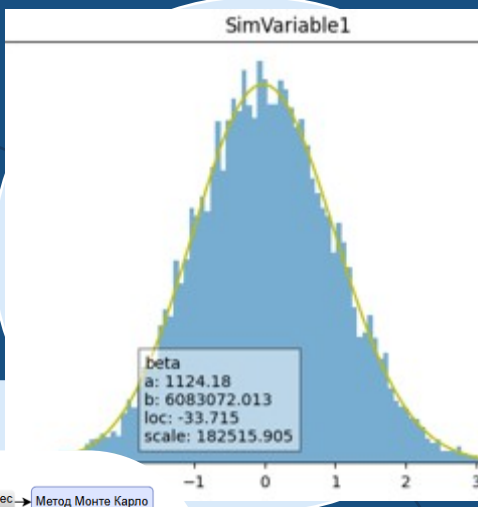
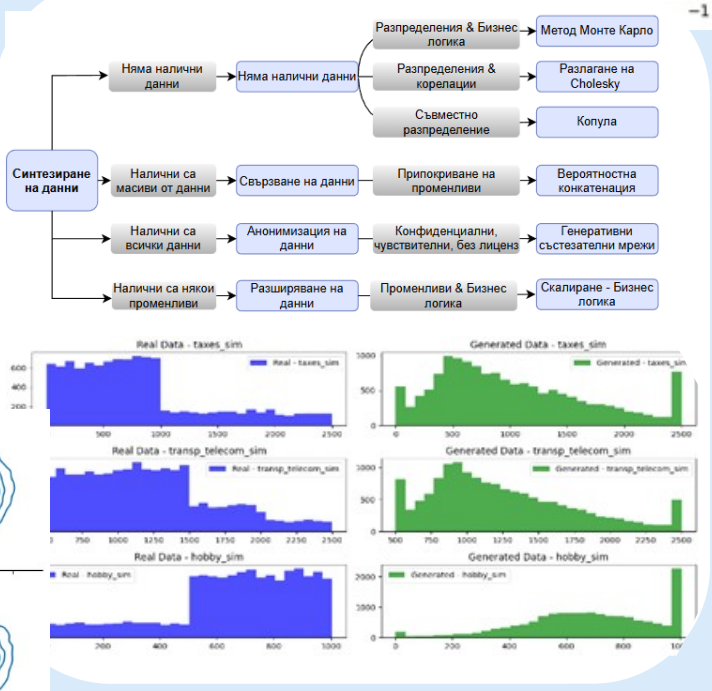
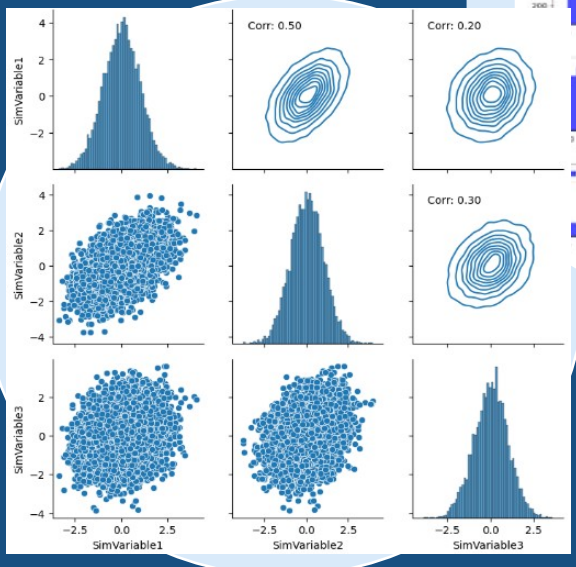
МОДЕЛИРАНЕ НА СИНТЕТИЧНИ ИНФОРМАЦИОННИ МАСИВИ. ПОДХОДИ ЗА СИНТЕЗ И ВАЛИДАЦИЯ







Колективна монография

2026

МОДЕЛИРАНЕ НА СИНТЕТИЧНИ ИНФОРМАЦИОННИ МАСИВИ. ПОДХОДИ ЗА СИНТЕЗ И ВАЛИДАЦИЯ

Колективна монография

София

2026

СЪДЪРЖАНИЕ

РЕДАКТОРСКО ПРЕДИСЛОВИЕ.....13
ИЗПОЛЗВАНИ ОЗНАЧЕНИЯ.....17
I. УВОД.....19
II. ОБЩИ НАСОКИ ПРИ РАБОТА СЪС СИНТЕТИЧНИ ДАННИ.....24
1 Големите данни в контекста на управленската наука.....25
1.1 Същност на проблематиката.....26
1.2 Големите данни като част от управленската наука.....26
1.2.1 Какво представляват големите данни.....26
1.2.1.1 Изучаване на начините за анализиране.....27
1.2.1.2 Извличане на информация от големи масиви.....28
1.2.2 Предизвикателства пред работата с големи масиви от данни.....28
1.2.2.1 Обезпечаване с данни.....28
1.2.2.2 Съхраняване на данни.....29
1.2.2.3 Анализ на данните.....29
1.2.2.4 Търсене, споделяне и трансфер на информация.....29
1.2.2.5 Визуализация на данните.....30
1.2.2.6 Обновяване на информацията.....30
1.2.2.7 Поверителност на данните.....30
1.2.3 Приложение на големите данни в управлението.....31
1.2.3.1 Прогнозни анализи.....31
1.2.3.2 Анализи на потребителско поведение.....31
1.2.4 Основна класификация на големите данни.....32
1.3 Законови рамки.....33
1.3.1 Общ регламент за защита на данните.....33
1.3.1.1 Принципи на регламента.....34
1.3.1.2 Права на субектите на данни.....35
1.3.1.3 Задължения на администраторите на лични данни.....36
1.3.1.4 Международни трансфери и прилагане.....37
1.3.2 Закон за кредитните институции.....38
1.4 Заключение.....41
2 Подходи за обезпечаване с информация.....45
2.1 Класически подходи.....45
2.1.1 Анкетни проучвания.....46
2.1.2 Интервюиране.....47
2.1.3 Системно наблюдение.....49

2.1.4 Дистанционни методи за набиране на информация....50
2.2 Симулация на данни....51
2.2.1 Постановка на задачата....51
2.2.2 Основни групи променливи....53
2.2.2.1 Демографски променливи....53
2.2.2.2 Социално-икономически променливи....54
2.2.2.3 Психологически и личностни променливи....54
2.2.2.4 Индивидуална склонност към риск....55
2.2.2.5 Финансови и банкови променливи....56
2.3 Заключение....57
3 Същност на синтетични данни. Приложение в управлението на бизнес структури....60
3.1 Въведение....60
3.2 Синтетични данни....61
3.2.1 Същност....61
3.2.2 Област на приложение....63
3.2.3 Предимства и предизвикателства....65
3.2.3.1 Предимства....65
3.2.3.2 Предизвикателства....65
3.3 Класификация на синтетичните данни....66
3.3.1 Според насоката за използване....67
3.3.1.1 Частично синтезиране на данни (обогатяване на данни)....67
3.3.1.2 Интерполация на данни....68
3.3.1.3 Изцяло синтезирани масиви от данни....68
3.3.2 Класификация според методите на генериране....69
3.3.2.1 Подходи базирани на вероятности....69
3.3.2.2 Подходи за симулация, които не са базирани на вероятности....70
3.4 Заключение....71
4 Валидиране на синтетични данни....74
4.1 Същност и дефиниция на процеса....74
4.1.1 Дефиниция....74
4.1.2 Същност....75
4.1.3 Цели на процеса....77
4.1.3.1 Осигуряване на статистическа прецизност на разпределенията..77
4.1.3.2 Максимизиране на аналитичната полезност....77
4.1.3.3 Осигуряване на структурна консистентност и логическа цялост 77

4.2 Необходимост и предпоставки....78
4.2.1 Контекстуална обусловеност....78
4.2.2 Регулаторни и етични норми....79
4.3 Класификация на методите за валидация....80
4.3.1 Количествени подходи....80
4.3.2 Качествени и функционални подходи....81
4.4 Тест на Колмогоров-Смирнов....82
4.4.1 Едноизвадков тест на Колмогоров-Смирнов....82
4.4.2 Двуизвадков тест на Колмогоров-Смирнов....83
4.4.3 Прилики и разлики между едноизвадков и двуизвадков тест на Колмогоров-Смирнов....84
4.5 Заключение....85
5 Бизнес логика....87
5.1 Същност на бизнес логиката в контекста на големите данни....87
5.1.1 Дефиниции....87
5.1.2 Функция....88
5.2 Стратегически приложения....90
5.2.1 Оптимизация на изработването на решения....90
5.2.2 Осигуряване на съответствие....90
5.2.3 Персонализация на крайния продукт....91
5.3 Ключови специфики на логическия слой....92
5.3.1 Унификация спрямо източника....92
5.3.2 Проследимост....92
5.3.3 Мащабируемост....93
5.4 Таксономия на видовете бизнес логика....93
5.4.1 Трансформационна бизнес логика....94
5.4.2 Интегрираща бизнес логика....94
5.4.3 Изчислителна логика....96
5.4.4 Филтрираща бизнес логика....96
5.5 Етапи на създаване на бизнес логика....97
5.6 Оценка на ефективността и качеството....98
5.6.1 Качество на бизнес логиката....98
5.6.2 Време за реакция....99
5.6.3 Адаптивност....99
5.7 Заключение....99

III. МЕТОДОЛОГИЯ ЗА ГЕНЕРИРАНЕ НА ДАННИ.....102
6 Методологични аспекти и методологично дърво.....103
6.1 Въведение.....103
6.2 Методологически насоки в процеса по създаване на данни.....103
6.2.1 Обогатяване на данни.....104
6.2.2 Интерполация на данни.....105
6.2.3 Синтез на данни.....106
6.2.3.1 При липса на данни.....106
6.2.3.2 При наличие на частични данни под формата на съществуващи характеристики.....106
6.2.3.3 Частични данни под формата на съществуващи вариращи набори от данни.....106
6.2.3.4 Съществуващ пълен нелицензиран набор от данни.....107
6.3 Работни фази.....108
6.3.1 Одит насочен към минимизиране на синтеза на данни.....108
6.3.2 Генериране на данни по (избран) метод.....109
6.3.3 Симулация на Монте Карло с оглед на характеристиките.....109
6.3.4 Разлагане на Холецки.....109
6.3.4.1 Обърната копула за генериране на извадки.....110
6.3.4.2 Разширяване на данните чрез бизнес правила.....110
6.3.4.3 Вероятностно конкатениране.....111
6.3.4.4 Генеративни състезателни мрежи.....112
6.3.5 Хоризонтална синхронизация.....112
6.3.6 Вертикално валидиране на характеристики.....112
6.4 Заключение.....113
7 Синтезиране на многомерни набори от лични данни.....116
7.1 Въведение.....116
7.2 Симулация базирана на общи разпределения и бизнес логика.....117
7.2.1 Методология.....117
7.2.1.1 Анализ на разпределения.....117
7.2.1.2 Бизнес логика.....118
7.2.1.3 Симулация.....118
7.2.2 Математически алгоритъм.....119
7.3 Симулация, базирана на общи разпределение и корелации.....120
7.3.1 Методология.....120
7.3.2 Математически алгоритъм.....121

7.4 Обърната копула за подбор на извадки....122
7.4.1 Методология....122
7.4.2 Математически алгоритъм....123
7.5 Симулация на данни, базирана на реални данни....123
7.5.1 Методология....123
7.5.2 Математически алгоритъм....125
7.6 Вероятностно конкатениране на бази данни....125
7.6.1 Методология....125
7.6.2 Математически алгоритъм....125
7.7 Генеративни състезателни мрежи (GAN)....126
7.7.1 Методология....126
7.7.2 Математически алгоритъм....127
7.8 Заключение....127
IV. ПОДХОДИ ЗА ГЕНЕРИРАНЕ НА ДАННИ....130
8 Автоматизиран алгоритъм за многовариантен синтез на данни с разлагане на Холецки....131
8.1 Методи за генериране на синтетични данни....131
8.1.1 Методи базирани на вероятности....131
8.1.1.1 Симулация Монте Карло....131
8.1.1.2 Генеративни състезателни мрежи (GAN)....132
8.1.2 Методи, които не се базират на вероятности....132
8.1.2.1 Обърната копула за генериране на извадки....132
8.1.2.2 Разлагане на Холецки....132
8.2 Синтез на данни въз основа на разпределения и корелации - дефиниране на проблема....133
8.2.1 Напасване на разпределенията....133
8.2.2 Формулиране на корелационна матрица....134
8.2.3 Разлагане на Холецки върху корелационната матрица....135
8.2.4 Симулация чрез метод Монте Карло....136
8.2.5 Матрично произведение на матрицата на Холецки....136
8.2.6 Валидиране на резултатите....137
8.3 Емпиричен пример....137
8.4 Заключение....140
9 Генериране на данни при наличие на частични информационни набори....143

9.1 Въведение.....143
9.2 Формулиране на задачата.....145
9.3 Генериране на данни.....146
9.4 Валидиране.....150
9.4.1 Бизнес логика.....150
9.4.2 Валидиране чрез бизнес логика.....152
9.4.3 Валидиране чрез псевдомоделиране.....154
9.5 Емпиричен пример.....156
9.6 Заключение.....159
10 Валидиране на синтетични данни за поведенческо моделиране.....162
10.1 Въведение.....162
10.2 Концептуален преглед.....163
10.2.1 Поведенчески данни.....163
10.2.2 Студен старт и синтетични данни.....165
10.2.3 Значение на валидацията на синтетични данни.....167
10.3 Предложени методи.....168
10.3.1 Одит на съгласуваност между експерти (HECA).....169
10.3.2 Програмен одит, базиран на Големи езикови модели (LLM-PA).....170
10.3.3 Автоматизирано калибриране с пропускане на една променлива.…171
10.4 Емпирични резултати.....173
10.4.1 Симулиран набор от данни за апробация.....174
10.4.2 Проучване на HECA.....175
10.4.3 LLM-PA пробен период.....177
10.4.4 LOFO-AC Trial.....178
10.5 Дискусия върху практически предложения и ограничения.....178
10.5.1 Генериране на извадки.....178
10.5.2 Правила.....178
10.5.3 LLM-PA заявка.....179
10.5.4 Настройване на LOFO-AC.....179
10.5.5 Мащабируемост и разход.....180
10.5.6 Поверителност и справедливост.....180
10.5.7 Ръководство за рамката за изработване на решения.....181
10.6 Заключение.....181
11 Генериране на данни при наличие на части от масиви (ProbCon)...186

11.1 Същност....186
11.2 Математически алгоритъм....187
11.3 Последователност и приложение на метода....188
11.3.1 Приложение на метода....188
11.3.2 Подробно описание на последователността на метода....189
11.3.2.1 Подготовка и първоначално вероятностно конкатениране....189
11.3.2.2 Нормализация посредством бинарна трансформация....190
11.3.2.3 Итеративна многослойна интеграция....190
11.3.2.4 Финализиране и почистване на масива....190
11.3.2.5 Анализ на процеса на редукция на данните....191
11.3.2.6 Резултати от проведената конкатенация....192
11.4 Предимства и потенциални ограничения на методологията....192
11.4.1 Предимства....192
11.4.2 Потенциални недостатъци....193
11.5 Заключение....194
12 Генериране на синтетични данни при наличие на пълни набори от данни. Анонимизация чрез генеративни състезателни мрежи (GAN).....196
12.1 Въведение....196
12.2 Теоретичен преглед....198
12.2.1 Основни моменти при генеративните състезателни мрежи....199
12.2.2 Математически алгоритъм....200
12.3 Внедряване на генеративни състезателни мрежи....201
12.3.1 Подготовка на данни за анонимизация с помощта на генеративни състезателни мрежи (GAN)....202
12.3.2 Конструиране на генератор за анонимизация на данни....203
12.3.3 Конструиране на дискриминатор за анонимизация на данни....205
12.3.4 Компилация на генеративната състезателна мрежа....206
12.3.5 Обучение на генеративната състезателна мрежа....208
12.3.6 Генериране на данни посредством обучена GAN....210
12.3.7 Последваща обработка на данни, генерирани с GAN....211
12.3.8 Оценка на косинусовото подобие на данни, генерирани от GAN....213
12.4 Заключение и бъдеща работа....215
13 Генериране на данни чрез съвместно разпределение....218
13.1 Въведение....218
13.2 Мотивация....220

13.3 Дефиниция на проблема.....221
13.4 Алгоритъм за синтез на данни - обърната копула за подбор на извадки 223
13.5 Емпиричен пример.....226
13.6 Сравнителен анализ.....230
13.7 Заключение.....232
14 Синтезиране на анонимизиран многомерен набор от данни.....236
14.1 Въведение.....236
14.2 Синтетични данни.....237
14.3 Симулиране на данни.....238
14.3.1 Речник на данните.....238
14.3.2 Променливи.....240
14.3.3 Анализ на разпределенията.....241
14.3.4 Разпределения.....242
14.3.4.1 Разпределения на демографски променливи.....243
14.3.4.2 Разпределения на индивидуални характеристики.....246
14.3.4.3 Доход.....247
14.3.4.4 Разходи.....248
14.3.4.5 Банкови и финансови характеристики.....248
14.3.4.6 Застрахователни услуги.....249
14.3.4.7 Допълнително здравно осигуряване.....250
14.3.4.8 Кредитни услуги.....250
14.3.5 Симулация на данни.....251
14.3.6 Бизнес логика.....254
14.3.7 Валидиране на данните.....255
14.4 Заключение.....256
15 Методологични подходи за създаване на многомерни лични данни 259
15.1 Въведение.....259
15.2 Основни фази на процеса.....260
15.2.1 Покритие.....260
15.2.1.1 Географско покритие.....260
15.2.1.2 Времево покритие.....261
15.2.1.3 Демографско покритие.....261
15.2.2 Източник на данни.....261
15.2.2.1 Първични източници на данни.....261
15.2.2.2 Доставчици на данни.....262

15.2.3 Методология....263
15.2.4 Обработка на данни....265
15.2.4.1 Метод на моментите....269
15.2.4.2 Обобщен метод на моментите (GMM)....269
15.2.5 Достъпност....269
15.2.6 Формат на данните....269
15.2.7 Осигуряване на качеството....270
15.2.8 Надеждност....272
15.2.9 Точност....272
15.2.10 Честота на актуализиране....273
15.3 Метаданни за статистически характеристики....274
15.4 Заключение....274
V. ЗАКЛЮЧЕНИЕ....277
VI. ОБОБЩЕНА БИБЛИОГРАФИЯ....285
Авторски публикации....299
VII. ПРИЛОЖЕНИЯ....301
Приложение 1 - Бизнес логиката на модела....301
Приложение 2 - Списък с описани променливи....306
Приложение 3 - Номинални променливи....328

РЕДАКТОРСКО ПРЕДИСЛОВИЕ

Настоящата монография е реализирана основно в рамките на научноизследователски проект НИД НИ 23/2023/В, финансиран от Университета за национално и световно стопанство. Авторите изразяват своята признателност към академичното ръководство за предоставеното финансиране и изследователска среда. В допълнение – и на Технически университет - София за предоставения достъп до MATLAB. Специални благодарности се отправят към изследователския колектив, чийто съвместен труд, експертиза и аналитичен принос са основополагащи за завършването на този труд.

В съвременната епоха на дигитална трансформация защитата на конфиденциалността и недостигът на качествени емпирични данни поставят синтеза на данни в центъра на научния интерес. Настоящото изследване адресира критичната нужда от методологически инструментариум за генериране на висококачествени синтетични масиви, които съхраняват статистическите свойства на оригиналите, без да компрометират личната неприкосновеност.

Интердисциплинарният подход в монографията създава предпоставки за обогатяване на теоретическата и приложна рамка. Изложението превежда читателя през целия процес – от изграждането на комплексно методологично дърво, през моделирането на съвместни разпределения и използването на набор от подходи за генериране на синтетични данни, до финалното валидиране на моделите. Крайният резултат представлява завършен модел за симулация на данни, приложим както в академичната сфера, така и в практическото бизнес прогнозиране.

Монографията е реализирана под редакторския прочит на гл. ас. д-р Васил Марчев.

Автори на отделните глави:

Големите данни в контекста на управленската наука (брой стр. 20)
гл. ас. д-р Васил Марчев
Катедра Управление, Университет за национално и световно стопанство, София

Подходи за обезпечаване с информация (брой стр. 15)
гл. ас. д-р Васил Марчев
Катедра Управление, Университет за национално и световно стопанство, София

Същност на синтетични данни и приложение в управлението на бизнес структури (брой стр. 14)
гл. ас. д-р Васил Марчев – 50%

Катедра Управление, Университет за национално и световно стопанство, София
доц. д-р Ангел Марчев – 50%
Катедра Управление, Университет за национално и световно стопанство, София

Валидиране на синтетични данни (брой стр. 13)
гл. ас. д-р Васил Марчев – 50%
Катедра Управление, Университет за национално и световно стопанство, София
доц. д-р Ангел Марчев – 50%
Катедра Управление, Университет за национално и световно стопанство, София

Бизнес логика (брой стр. 20)
гл. ас. д-р Васил Марчев – 50%
Катедра Управление, Университет за национално и световно стопанство, София
доц. д-р Ангел Марчев – 50%
Катедра Управление, Университет за национално и световно стопанство, София

Методологични аспекти и методологично дърво (брой стр. 13)
доц. д-р Ангел Марчев – 50%
Катедра Управление, Университет за национално и световно стопанство, София
гл. ас. д-р Васил Марчев – 50%
Катедра Управление, Университет за национално и световно стопанство, София

Синтезиране на многоизмерни набори от лични данни (брой стр. 14)
доц. д-р Ангел Марчев – 50%
Катедра Управление, Университет за национално и световно стопанство, София
гл. ас. д-р Васил Марчев – 50%
Катедра Управление, Университет за национално и световно стопанство, София

Автоматизиран алгоритъм за многовариантен синтез на данни с разлагане на Холецки (брой стр. 12)
доц. д-р Ангел Марчев – 50%
Катедра Управление, Университет за национално и световно стопанство, София
гл. ас. д-р Васил Марчев – 50%
Катедра Управление, Университет за национално и световно стопанство, София

Генериране на данни при наличие на частични информационни набори (брой стр. 19)
доц. д-р Ангел Марчев – 33.33%
Катедра Управление, Университет за национално и световно стопанство, София

гл. ас. д-р Васил Марчев – 33.33%
Катедра Управление, Университет за национално и световно стопанство, София
доц. д-р Александър Ефремов – 33.33%
Факултет Автоматика, Технически университет, София

Валидиране на синтетични данни за поведенческо моделиране (брой стр. 24)
доц. д-р Ангел Марчев – 33.33%
Катедра Управление, Университет за национално и световно стопанство, София
гл. ас. д-р Васил Марчев – 33.33%
Катедра Управление, Университет за национално и световно стопанство, София
доц. д-р Александър Ефремов – 33.33%
Факултет Автоматика, Технически университет, София

Генериране на данни при наличие на части от масиви (брой стр. 14)
гл. ас. д-р Васил Марчев – 50%
Катедра Управление, Университет за национално и световно стопанство, София
доц. д-р Ангел Марчев – 50%
Катедра Управление, Университет за национално и световно стопанство, София

Генериране на синтетични данни при наличие на пълни набори от данни. Анонимизация чрез генеративни състезателни мрежи (GAN) (брой стр. 22)
доц. д-р Ангел Марчев – 50%
Катедра Управление, Университет за национално и световно стопанство, София
гл. ас. д-р Васил Марчев – 50%
Катедра Управление, Университет за национално и световно стопанство, София

Генериране на данни чрез съвместно разпределение (брой стр. 17)
доц. д-р Ангел Марчев – 33.33%
Катедра Управление, Университет за национално и световно стопанство, София
докторант Димитър Любчев – 33.33%
Катедра Управление, Университет за национално и световно стопанство, София
гл. ас. д-р Васил Марчев – 33.33%
Катедра Управление, Университет за национално и световно стопанство, София

Синтезиране на анонимизиран многоизмерен набор от данни (брой стр. 23)
гл. ас. д-р Васил Марчев – 20%
Катедра Управление, Университет за национално и световно стопанство, София
доц. д-р Ангел Марчев – 20%

Катедра Управление, Университет за национално и световно стопанство, София
Докторант Милена Пирянкова – 20%
Катедра Управление, Университет за национално и световно стопанство, София
Даниел Масърлиев – 20%
Катедра Управление, Университет за национално и световно стопанство, София
Валентин Митков – 20%
Катедра Управление, Университет за национално и световно стопанство, София

Методологични подходи за създаване на многомерни лични данни (брой стр. 43)
гл. ас. д-р Васил Марчев – 10%
Катедра Управление, Университет за национално и световно стопанство, София
доц. д-р Ангел Марчев – 10%
Катедра Управление, Университет за национално и световно стопанство, София
доц. д-р Калоян Харалампиев – 10%
Катедра Социология, Софийски университет "Св. Климент Охридски"
доц. д-р Александър Ефремов – 10%
Факултет Автоматика, Технически университет, София
д-р Боян Марков – 10%
Катедра Управление, Университет за национално и световно стопанство, София
докторант Димитър Любчев – 10%
Катедра Управление, Университет за национално и световно стопанство, София
докторант Милена Пирянкова – 10%
Катедра Управление, Университет за национално и световно стопанство, София
докторант Богомил Филипов – 10%
Катедра Управление, Университет за национално и световно стопанство, София
Даниел Масърлиев – 10%
Катедра Управление, Университет за национално и световно стопанство, София
Валентин Митков – 10%
Катедра Управление, Университет за национално и световно стопанство, София

Настоящата разработка систематизира разгледаните в хода на проекта алгоритми и подходи, оформяйки интегрирана рамка за синтез на данни. Всяка глава притежава самостоятелна научна стойност, като същевременно е част от цялостния методологичен инструментариум, изграден в рамките на изследването.

ИЗПОЛЗВАНИ ОЗНАЧЕНИЯ

Номерация на разделите. Монографията е развита в увод, три основни раздела, заключение, обобщена библиография и приложения, обозначени с римски цифри. Всеки един от основните три раздела представлява съответната част от научното изследване: общи насоки при работа със синтетични данни, методология за генериране на данни и подходи за генериране на данни.

Номерация на главите. Главите в настоящия текст са номерирани във възходяща последователност, независимо от конкретния раздел, към който принадлежат. Отделните части могат да включват други съставни подчасти, номерирани последователно в рамките на главата. Примери: 8.2., 14.2.1. По този начин се избягва излишното добавяне на допълнителен индекс за номера на раздела в номерацията на главите.

Номерация на таблиците, фигурите и формулите. В номерацията се включва и номерът на главата. В рамките на една глава номерацията е в нарастваща последователност, като по отделно са означени таблиците, фигурите и схемите. Примери: Табл. 10.4.4-2, Фиг. 8.3-1. По този начин се заявява принадлежността на разглежданите елементи от изследването към съответните глави или подглави.

Номерация на информационните източници. Във възходяща последователност, приблизително по реда на употреба в текста, независимо от главата и подредбата по други признаци.

Номерата на цитираните източници в текста са оградени в квадратни скоби. Примери: [7], [2, 4, 14].

БНБ – Българска народна банка
БФБ – Българска фондова борса
ДЗО – Допълнително здравно осигуряване
ДК – дебитна карта
ДПО – Допълнително пенсионно осигуряване
ДЦК – Държавни Ценни Книжа
ИВ – инвестиционна възможност
ИИ – инвестиционен/ни инструмент/и
Инфостат - информационна система за разпространение на статистически данни (Infostat)
ИП – инвестиционен профил
КК – кредитна карта
НАП – Национална агенция по приходите
НСИ – Национален статистически институт
ТБ – търговска банка

ФГВБ – Фонд за гарантиране на влоговете в банките
ФЛ – физическо лице
GAN – генеративни състезателни мрежи (Generative adversarial network)
HECA - одит на съгласуваност между експерти (Human expert consistency audit)
LLM-PA – Програмен одит, базиран на големи езикови модели (LLM-based programmatic audit)
LOFO-AC – Автоматизирано калибриране с пропускане на една променлива (Leave-one-feature-out automated calibration)
MatLab – Матрична лаборатория (MATrix LABoratory - MATLAB, наименование на среда за моделиране и алгоритмичен език)
Microsoft Excel - наименование на аналитична и програмна среда
Python - наименование на среда за моделиране и алгоритмичен език

I. УВОД

В контекста на съвременната икономика на големите данни се наблюдава дефицит на достъпни масиви, съдържащи чувствителна по своето същество информация. Обект на строги регулаторни ограничения са данните, попадащи в обхвата на Общия регламент за защитата на данните (GDPR), Закона за кредитните институции (в частта му уреждаща банковата тайна), както и съпътстващото европейско законодателство за защита на личните данни. Настоящата монография изследва методологията за генериране на многомерни синтетични масиви, интегриращи личностни, демографски и поведенчески характеристики, приложими в банковия сектор, финансовите пазари и други социално-икономически сфери. Спецификата на разглежданите данни, тяхната структурна комплексност и правните рамки налагат необходимостта от алтернативни подходи за набавяне, симулиране и синтезиране на информационни масиви с високо качество и представителност.

В настоящото изследване се систематизира набор от разнообразни личностни характеристики, които позволяват изграждането на изчерпателни профили върху статистически значими съвкупности от физически лица. Фокусът е насочен към установяване на моделите на поведение на финансово активните потребители. Така синтезираният масив от данни предоставя надеждна основа за широка съвкупност от емпирични изследвания в социално-икономическата област, преодолявайки бариерите, свързани с поверителността и достъпа до първична информация.

Основната идея на монографията е да се изведе цялостен, методически описан и моделиран процес по генериране на набор от многомерни синтетични данни, интегриращ личностни, демографски и финансови характеристики.

Предметът на изследването се формулира като: процес на генериране на многомерен синтетичен информационен масив посредством разнообразна съвкупност от подходи, във вид подходящ за последващ анализ и интегриране в модели.

Обектът на изследване в настоящата разработка се дефинира като система от технологични ресурси и изчислителни архитектури за генериране на синтетични данни, изследвайки структурните компоненти на синтетичните информационни масиви и тяхната вътрешна логика. Това включва параметрите на цифровата среда и математическите закономерности в сложни социално-икономически системи.

Целта на научното изследване е да се установи набор от подходи за синтезиране/симулиране на масив от многомерна база от синтетични данни за финансови услуги, както и да бъде генериран многомерен масив от синтетични данни, съдържащ в себе си демографски, финансови и лични характеристики.

За реализиране на гореописаната цел спомагат следните задачи:

Теоретични:

1. Да се изяснят съвременните подходи за набиране на големи масиви от информация.
2. Да бъде осъществен преглед на научните подходи за генериране на масиви от информация. Поради спецификата на изследването е необходимо генерирането на базата данни, да се извърши посредством комбинация от методи за симулиране.

Методически:

3. Да се разработи методика за реалистична симулация на голям набор от лични данни чрез няколко алтернативни подхода.
4. Да се разработят алгоритми и съответното програмно осигуряване за установяване на компютъризирана симулационна среда, позволяваща автоматизиран процес за генериране на синтетичен многомерен масив от данни.

Емпирически:

5. Да се установят основните групи променливи, въз основа на които се генерира информационният масив.
6. Да се определят личностните преференции и индивидуалната склонност за риск на анкетираните физически лица. Провеждане на анкети с предварително установени въпросници.
7. Да се обединят генерираните масиви от данни. Обединяване на различните масиви от данни посредством интегриран авторски подход (вероятностна конкатенация).

Изследователска методика. Основният подход е свързан с изследването на различни възможности за изграждане на модел за генериране на многомерен масив от синтетични данни, приложим в банковото дело и икономико-финансови области. Методиката обхваща процес по изграждане на модел за генериране, състоящ се от няколко основни подфази. В ключовите подчасти се използват статистически подходи интегрирани в алгоритми. Като съществени елементи от изследователската методика могат да бъдат посочени:

- Използване на статистически разпределения
- Приложение на бизнес логика
- Приложение на авторски метод за вероятностно конкатениране на данни

Съществени ограничения

1. Дефинираният обект на изследване представлява алгоритмична система базирана на статистически принципи и парадигми, която следва стриктна последователност в процеса по генериране на синтетични данни.
2. Пълна липса на публично достъпни лични данни, които могат да се използват за валидиране на синтетичните масиви.
3. Данните за демографските характеристики са достъпни посредством последното Преброяване на населението в Република България 2021г, проведено в периода 07.09.2021 – 03.10.2021г.
4. Изкуствено е ограничен броят на характеристиките от отделните групи, до най-съществените, за да бъде съобразен с наличието на данни за съставяне на разпределения или възможността за съставяне на разпределение на база априорна информация

Основно допускане е, че симулираните и конкатенирани данни са достатъчно близки до реалните и като следствие, резултатите от проведената апробация са достоверни.

Представената система за генериране на многомерни синтетични информационни масиви е базирана върху неизчерпателен набор от подходи за симулация, сред които са подходи основани на разпределения, корелации, генеративни състезателни мрежи, копула и др. Авторите на монографията отчитат множеството необходими допускания за реализация на разглежданите подходи за синтезиране на данни. При процеса по изграждане на методологичната рамка и изчислителните процедури се вграждат множество имплицитни допускания относно ключови променливи. Тези структурни хипотези са критично важни за коректното функциониране на аналитичния алгоритъм и неговите компоненти. Процесът изисква прецизно дефиниране на вътрешните ограничения и параметрите на модела. Всяко допускане директно влияе върху финалната точност и адекватност на извежданите резултати. От съществена важност е допускането свързано с необходимостта за нормалност на разпределенията на разглежданите променливи.

Влиянието на описаните допускания е възможно да бъде опростено, в голяма степен, или ограничено посредством използването на подходящи разпределения, получени от реални данни или техни производни, спомагащи за намаляване на допусканията в първичната база от разпределения.

Към настоящия момент моделът за генериране на многомерни синтетични информационни масиви е концентриран върху генериране на данни с финансово-икономическа насоченост, като не са правени изследвания за приложението му в други научни области. Въпреки това, логиката заложена в

алгоритмичната обезпеченост на системата би позволила апробация с разнообразни характеристики, при налични разпределения и възможност за синтезиране на съответната бизнес логика. Резултатите от тези нови модели, разбира се, е необходимо да бъдат валидирани със съответните реални данни и да се установи дали качеството на новополучената информация е на необходимото ниво, както и дали данните са с подходяща структура за обучаване на модели в дадената научна област.

Работни хипотези

1. Възможно е да се конкатенира многомерен масив от данни от няколко различни източници, сведен до ниво индивид.
2. Възможно е да се използва набор от различни подходи в процеса на набиране на информационни масиви.
3. На база на извлечените индивидуални характеристики може да се генерира многомерна база от синтетични данни приложими във финансовите среди.
4. Установената бизнес логика ще може да отсее непостижимите или много слабо вероятни стойности, генерирани от модела.

Първичната емпирична база на изследването съдържа:

- Обобщени демографски данни от НСИ и Преброяване 2021, за населението на Република България по различни показатели;
- Данни за психологически профили. Предварително проведено международно проучване от Типатов, по отношение на темперамента и разпределение на основните групи.
- Данни от анкети за индивидуална склонност към риск и инвестиционни предпочитания;
- Правдоподобни допускания по отношение на характеристиките, за които не е налична първоначална база от данни за извличане на статистическо разпределение.

Обхват на изследването. Монографията обхваща теоретично и емпирическо изследване на набор от подходи за генериране на многомерен синтетичен набор от данни и приложение на подходите за получаване на необходимата база от данни. От гледна точка на теорията за синтезиране на масиви от данни, вниманието се фокусира главно върху използването на статистически парадигми, интегрирани в алгоритми за синтез на данни, генеративни състезателни мрежи и копула. Методически са включени широк набор от практики на машинно самообучение за симулация на данни.

Извън обхвата на разработката остават**:**

- Натрупването на априорна информация за осъществяване на бизнес логиката и статистическите разпределения, за които не са налични емпирични данни;
- Етапът на предварителен подбор на променливи, чрез които се реализира генерирането на данни;
- Всички останали методи, които не са включени в първоначалния неизчерпателен списък от подходи за симулиране на информационни масиви;
- Достигане на пълна автоматизация в процесите по синтезиране на характеристики и генериране на статистически разпределения в процеса на работа.

Структурата на монографията е в съответствие с целта и поставените задачи. Отчитайки основната идея, монографията е разработена в увод, три глави, заключение, към всяка от подглавите е добавен списък с информационни източници и приложения. Оформянето в три глави следва стандартната логическа последователност при подобни изследвания:

1. Първи раздел включва както основните насоки при работа със синтетични данни, така и съществени моменти свързани с регулаторните актове (първа глава), които създават необходимостта от приложението на подходи за генериране. Отделено е съществено внимание на концепциите свързани с развитието на Бизнес логиката (пета глава), като част от моделите за синтез на данни, валидирането на данни (четвърта глава), като основен етап в процеса по синтезиране на данни.
2. Втори раздел обхваща методологията в процеса по генериране на данни. Системно са описани методологическите аспекти на процеса, както и методологичното дърво със съответните основни разклонения при работа със синтетични данни (шеста глава). Освен това е направен подробен преглед на използваните подходи в процеса на симулиране на данни (седма глава).
3. Трети раздел включва подробен преглед и апробация на алгоритъма за генериране на данни, в зависимост от използваните различни подходи на работа. В последните две глави подробно е разписан процеса по генериране на многомерен синтетичен набор от данни (четиринадесета глава), както и пълно описание на метаданните. Готовият масив от синтетични данни е предоставен за последващо използване в Hugging face репозиториум – на адрес DOI https://doi.org/10.57967/hf/3701. Всички разработени софтуерни алгоритми в среда Python и MatLab са налични в GitHub репозиториум на адрес: https://github.com/Marchev-Science/synthetic-data.

II. ОБЩИ НАСОКИ ПРИ РАБОТА СЪС СИНТЕТИЧНИ ДАННИ

1 Големите данни в контекста на управленската наука

Автор: Васил Марчев

Резюме. Аспектите на големите информационни масиви интегрирани в управленската наука обхващат културни, технологични и научни аспекти, дефинирани чрез сложни взаимовръзки между хетерогенни и автономни източници. Ефективното боравене с тези масиви изисква преминаване през етапи на предварителна обработка, трансформация и моделиране, като основна роля играят парадигмите за паралелна обработка в разпределени клъстерни системи. Обезпечаването с данни се реализира посредством автоматизирано извличане, докато процесът по съхранение разчита на технологични инфраструктури за осигуряване на хоризонтална мащабируемост. Важна част от аналитичния цикъл е визуализацията, която предотвратява когнитивното претоварване и служи като мост между изчислителната мощ и човешкия интелект.

Приложенията на големите информационни масиви се простират от прогнозни анализи за оптимизация на риска до детайлен разрез на потребителското поведение, насочени към изработване на информирано управленско решение. Използването на невронни мрежи и клъстерен анализ позволява индивидуализиран маркетингов подход и преминаване към проактивно стратегическо планиране.

Ключово предизвикателство представлява Общият регламент за защита на данните (GDPR), който въвежда фундаментални принципи като законосъобразност, минимизиране на данните, точност и ограничение на съхранението. Системите са длъжни да осигурят запазване на неприкосновеността на правата на субектите, включително правото на достъп, коригиране, заличаване и преносимост. Специфика в процеса прибавя и Закона за кредитните институции, в който се урежда статута на банковата тайна и нейния обхват. Законът регламентира строги процедури за разкриване на информация през последователност от съдебни практики и решения. Потенциалното нарушаване на описаните регулации може да доведе до значителни административни санкции и мерки от страна на надзорните органи.

Ключови думи: големи данни, потребителско поведение, моделиране, банкова тайна, GDPR

1.1 Същност на проблематиката

В научната литература понятието за големи данни се разглежда през множество перспективи, включително като информационен актив, характеризиращ се с обем, скорост и разнообразие. Работната дефиниция в настоящия труд се фокусира върху многомерни информационни масиви, произтичащи от исторически събития или алгоритми, които носят допълнителна информация чрез сложни взаимовръзки между променливите. Фокусът върху управлението на структури с висока размерност изисква приложението на специализиран статистически софтуер и мащабируеми изчислителни архитектури с цел ефективно извличане на скрити зависимости.

Процесът на анализ на тези масиви надхвърля капацитета на традиционните системи за управление на бази данни и се основава на алгоритми за машинно и дълбоко обучение. Особено предизвикателство представлява недостигът на масиви, свързани с чувствителна информация, която попада под строги законови регулации като GDPR и Закона за кредитните институции. Разглеждането на тези аспекти е критично за осигуряването на поверителност и сигурност при трансформирането на суровите данни в стратегическа информация. Настоящата глава разглежда хипотезата, при която е налице необходимост от разработване на модел, базиран на голям обем от исторически данни, които попадат в гореизброените законодателни рамки.

1.2 Големите данни като част от управленската наука

1.2.1 Какво представляват големите данни

В научната литература съществува широк набор от дефиниции, свързани с понятието големи данни. Основните логически насоки се обуславят от следните смислови твърдения: „големи, хетерогенни, автономни източници с разпределен и децентрализиран контрол, стремящи се да изследват сложни и развиващи се взаимовръзки между данните“ [1], “големите данни са информационен актив, характеризиращ се с голям обем, скорост и разнообразие, който изисква специфични технологии и аналитични методи за превръщането му в стойност“ [2]. Други автори разглеждат описваното понятие от философска гледна точка: „големите данни са културно, технологично и научно явление, основаващо се на взаимодействието на технологии, анализ и митология” [3] или разглеждайки понятието от гледна точка на процесите, които го характеризират по следния начин- „големите данни описват ново поколение технологии и архитектури, предназначени за икономично извличане на стойност от много големи обеми от широкообхватни данни“ [4].

Като се вземат под внимание гореописаните дефиниции на понятието и се отчетат спецификите на настоящия труд, твърдението се използва със следната работна дефиниция: големите данни представляват многомерни информационни масиви, които от своя страна могат да бъдат както продукт на исторически събития, така и генерирани посредством алгоритми. Те се характеризират с големи мащаби, скорост, структурно разнообразие и взаимовръзки между променливите и носят допълнителна информация, чрез която подпомагат процесите по изработване на информирани управленски решения.

Разглеждайки последователността от действия се установява фокус върху анализа на съвременните информационни масиви и боравенето със структури, характеризиращи се с висока размерност на характеристиките и голям брой индивидуални записи. Това от своя страна създава предпоставки за използване на набор от специализирани софтуери, с цел да се подобри и ускори обработката на данните. Създаването на мащабируеми изчислителни архитектури, проектирани за паралелна обработка предоставя на изследователите възможност за идентифициране на съвкупност от скрити зависимости между променливите, които остават незабележими при анализ на по-малки и хомогенни извадки. Това от своя страна пречупва разбирането за големите масиви от информация през призмата на процесите по ефективно извличане на знания посредством набор от статистически алгоритми интегрирани в системите за управление.

1.2.1.1 Изучаване на начините за анализиране

Процесът по анализ на големи информационни масиви изисква приложението на набор от специализирани статистически и изчислителни методи и парадигми, които надхвърлят капацитета на традиционните системи за управление на бази данни. В основата на този процес залягат алгоритмите за машинно обучение и дълбоко обучение, които създават предпоставки за автоматизирано разпознаване на образи и откриване на скрити зависимости в неструктурирани данни [10]. Прилагането на методологии за паралелна обработка създават предпоставки за ефективното разпределение на изчислителния товар в разпределени клъстерни системи [8].

Съществува набор от различни усъвършенствани аналитични инструменти, които установяват възможности за обработка на масивите в реално време и посредством итеративен подход, което е от съществено значение за динамични сектори като управлението, финансите и здравеопазването. Процесът на извличане на знания преминава през етапи на предварителна обработка, трансформация и моделиране, осигурявайки

превръщането на необработения информационен поток в стратегическа информация за същността на системата [9]. Валидирането на тези модели посредством техники за кръстосана проверка, установява предпоставки за гарантиране на необходимата прецизност и приложимост при прогнозиране на бъдещи тенденции.

1.2.1.2 Извличане на информация от големи масиви

Извличането на знания от големи информационни масиви, по своята същност, представлява процес на комплексна трансформация на сурови записи в набор от интегрирани и приложими знания. За целите на така поставената задача е необходимо да бъдат приложени набор от специализирани методологии за извличане на знания от данни, чрез които се установява статистически значими модели в информационните масиви. Процесите по ефективно извличане на допълнителни знания от данните са свързани с етапи по предварителна обработка, обхващаща нормализация на данните и подготовка на структурите за последващи автоматизирани анализи. Като следствие от така описаните действия, големите масиви от информация се разглеждат през призмата на ценен управленски ресурс за извличане на знания и достигане до стратегически решения, а не само като абстрактно количество информация.

1.2.2 Предизвикателства пред работата с големи масиви от данни

1.2.2.1 Обезпечаване с данни

Обезпечаването на системите с големи масиви от информация представлява комплексен процес, който се базира на автоматизирано извличане и агрегиране на структурирана и неструктурирана информация посредством набор от алгоритмични подходи. В основата на този процес залягат технологиите за уеб скрапинг и приложните програмни интерфейси, които позволяват систематично събиране на набори от данни, намиращи се в цифровото пространство в реално време [5]. Съвременното управление на големи данни изисква използването на мащабируеми архитектури за съхранение и първоначална обработка, които да гарантират интегритета и валидността на постъпващите информационни потоци.

Разширяването на съвременната технологична екосистема въвежда нови парадигми в методологията за събиране и управление на информация, при която сензорните мрежи генерират масивни потоци от телеметрични данни [6]. Тези методи налагат прилагането на техники за филтрация и компресия още в точката на генериране на данните, с цел оптимизиране на изчислителните

ресурси. Интеграцията на облачните технологии допълнително улеснява централизираното управление на тези разнородни масиви, осигурявайки необходимата инфраструктура за последващ дълбочинен анализ и извличане на знания [7].

1.2.2.2 Съхраняване на данни

Използването на големи информационни масиви има пряка корелация с ефективното им съхраняване. За целта се използват както сървърни хардуерни конфигурации, така и облачни инфраструктури. Тези технологии осигуряват хоризонтална мащабируемост и висока наличност посредством репликация на данни върху множество клъстерни възли, което е критично за поддържането на производителността при нарастващи обеми [11]. Масивите от данни могат да бъдат съхранявани и достъпни в различни видове файлови формати.

За обезпечаването на управленските процеси е необходимо сигурността и дългосрочната устойчивост на масивите от данни да бъде гарантирана, като това може да се получи посредством набор от сложни механизми за криптиране и управление на жизнения цикъл на данните в хибридни облачни среди.

1.2.2.3 Анализ на данните

Процесът по анализ на големи информационни масиви се базира на имплементирането на набор от статистически парадигми и алгоритми с цел извличане на закономерности от високообемни и динамични структури [12]. Съществено място в последователността на анализа на информация заемат както процесите на описание на минали събития, така и прогнозирането на бъдещи тенденции с висока степен на достоверност.

Интегрирането на изкуствен интелект и методи за обработка на естествен език разширява обхвата на анализа към неструктурирани текстови масиви и мултимедийно съдържание [13]. Разглежданият цялостен подход към анализа на данни осигурява повишаване на прецизността при изработването на стратегически решения, базирани на доказателства. Ефективността на тези процеси се обуславя от способността на системите да скалират ресурсите в съответствие с нарастващата комплексност на информационните потоци.

1.2.2.4 Търсене, споделяне и трансфер на информация

Процесите по търсене, споделяне и трансфер на информация в мащабни системи се основават на протоколи за оперативна съвместимост и семантични технологии, които оптимизират откриваемостта на ресурсите [14]. Ефективното споделяне изисква използването на стандартизирани метаданни и

разпределени архитектури, които минимизират загубата на данни по време на пренос и гарантират сигурността чрез съвременни криптографски методи. Трансферът на големи масиви често се осъществява чрез паралелни потоци и оптимизирани мрежови пътища, за да се преодолеят ограниченията на пропускателната способност [15].

1.2.2.5 Визуализация на данните

Визуализацията на големи масиви от данни е съществен етап от аналитичния процес, който трансформира абстрактни цифрови структури в интуитивни графични репрезентации, улесняващи откриването на нови знания и тенденции [16]. Предизвикателствата при визуалното представяне на големи информационни масиви са свързани с високата плътност на информацията, което налага използването на усъвършенствани техники като клъстеризация, агрегиране и динамично филтриране за предотвратяване на когнитивното претоварване [17]. Визуалната интерпретация на информационните масиви се превръща в ключов управленски инструмент интегриращ изчислителната мощ на машините и творческия потенциал, с цел изработване на информирани решения.

1.2.2.6 Обновяване на информацията

Обновяването на големи масиви от данни изисква внедряването на усъвършенствани стратегии за инкрементална обработка (актуализация само на данните, които са били променени), които позволяват модифициране на информацията без необходимост от цялостно презаписване на масива [18]. Този процес се основава на механизми за улавяне на промените в реално време [19], които гарантират синхронизация между оперативните източници и аналитичните хранилища при минимално закъснение. Поддържането на актуалност е от критично значение за прецизността на моделите за машинно обучение, които разчитат на постоянен поток от свежи данни за избягване на деградация на резултатите [19]. Необходимо е поддържането на консистентност в процесите по обновяване на информационните масиви, като за целта се използват усъвършенствани протоколи за консенсус.

1.2.2.7 Поверителност на данните

Осигуряването на поверителност при обработката на големи масиви от данни представлява критично предизвикателство, произтичащо от риска за повторна идентификация на анонимизирани субекти чрез техники за кръстосани препратки [20]. В науката за големи данни са познати различни методологии за справяне с посочената проблематика. Такива подходи могат да

са свързани с добавянето на контролиран шум в данните, което спомага запазването на анонимността на потребителите, като същевременно не нарушава статистическата значимост на резултатите. Подобни мерки са от съществено значение за обезпечаване на данните от гледна точка на GDPR и Закона за кредитните институции.

Наблюдават се редица подходи за криптиране на информацията (като към това спада и блокчейн технологията), с цел да бъде осигурена максимална сигурност при минимална експозиция на чувствителна информация пред външни заплахи.

1.2.3 Приложение на големите данни в управлението

1.2.3.1 Прогнозни анализи

Големите информационни масиви намират своето приложение в процеса по анализ и прогнозиране в рамките на организационното управление, предоставяйки възможността да се предвиждат бъдещи събития посредством използването на прогнозиращи методологии, основани на исторически данни. В науката за големи набори от данни се разглеждат методи обхващащи регресионни анализи, невронни мрежи и други способи, които са основани на алгоритми за машинно самообучение. Тези прогнозни модели обработват исторически и текущи данни, за да идентифицират статистически значими закономерности, които служат като основа за оптимизация на веригите за доставки, управление на риска и персонализация на потребителското преживяване [21]. Чрез синергията между изчислителната мощ и статистическите принципи се постига преход от реактивно към проактивно стратегическо планиране и управление на процесите.

Способността за мащабиране на разглежданите аналитични рамки създава предпоставки, при наличие на експоненциален растеж на информационните потоци, да бъде осигурена точността на прогнозните модели. Посредством възможностите, които предоставят от една страна големите масиви от разнообразни данни, а от друга страна все по-динамично развиващата се изчислителна мощ, прогнозните модели се утвърждават като фундаментален инструмент за анализи, иновации и по-дълбоко разбиране както на бизнеса, така и на клиентското поведение, позволявайки на субектите да минимизират несигурността в една динамично променяща се среда.

1.2.3.2 Анализи на потребителско поведение

Съществен момент в приложението на големите масиви от данни представлява анализа и управлението на потребителското поведение. Създават

се предпоставки организациите да създават детайлни разрези, посредством които да подобрят разбирането на клиентското поведение. За тази цел се използва набор от подходи, базирани на алгоритми за клъстерен анализ и филтриране според предпочитанията на потребителите, които предоставят възможност за създаване на индивидуализиран подход към различните клиентски профили. Тези аналитични подходи трансформират маркетинговите стратегии от масов кампаниен принцип към индивидуално таргетирани взаимодействия, базирани на обективни данни за действията на потребителите.

Интегрирането на информация от социални мрежи и мобилни устройства добавя контекстуална дълбочина към поведенческия анализ, позволявайки идентифициране на емоционални състояния и социални влияния [22]. Прилагането и разглеждането на процеси като синергията между психологията на потребителя и статистическите алгоритми спомагат за създаване на динамични модели, които оптимизират жизнения цикъл на клиента и подобряват управлението на бизнес процесите.

Анализът на потребителското поведение чрез големи масиви от данни позволява детайлно картографиране на всяка точка на контакт между индивида и бранда в многоканална среда [23]. Оптимизацията на тези пътеки в реално време може да бъде постигната посредством интеграция на изкуствен интелект, за динамично адаптиране на съдържанието и навигацията спрямо текущото поведение на потребителя.

Използването на прогнозиращи модели създава предпоставки за предвиждане на следващото най-вероятно действие, което минимизира усилията на клиента и увеличава удовлетвореността [24]. Чрез тези действия организациите постигат не само подобряване на ефективността, но и изграждат дългосрочни взаимоотношения, основани на подобрено клиентско преживяване и взаимодействие.

1.2.4 Основна класификация на големите данни

Класификацията на големите данни се основава на степента на вътрешната им организация. Това представлява формално дефинирана схема, която определя методите за тяхното съхранение и последващ анализ [25]. В най-общ аспект големите масиви от информация могат да бъдат класифицирани както следва:

Структурирани данни. Структурираните данни представляват информация, организирана във фиксирани формати и таблични структури, обикновено управлявани чрез релационни бази данни. Осигуряването в такива формати позволява бързо търсене и по-прецизна статистическа обработка.

Структурираните данни се характеризират с висока степен на подреденост. Тези информационни масиви съставляват едва малка част от общия обем на генерираната в глобален мащаб цифрова информация

Полуструктурирани данни. Полуструктурираните данни съставляват значителен дял от съвременните информационни масиви. Този тип данни изискват по-гъвкави софтуерни архитектури, както и значително по-сложни алгоритми в процесите по анализ. Разглежданите масиви от информация могат да бъдат във формат XML, JSON и други. Те се характеризират с липсата на обособена таблична форма, но съдържат в себе си тагове и метаданни за семантично разделяне на елементите.

Неструктурирани данни. Неструктурираните данни по своята същност представляват информационни масиви обхващащи текстови документи, мултимедийно съдържание, публикации в социални мрежи и др. Този тип данни не съблюдава предварително дефиниран модел. Основните техники, които се използват при анализа им са свързани с филтрирането на естествен език, което предполага значителен времеви обем и висока цена за последваща обработка.

1.3 Законови рамки

Спецификата на данните, тяхната комплексност и законовите рамки, предполагат редица трудности, при обезпечаване на моделите с необходимото количество и качество на информацията. Данните, необходими за създаване на прогнозиращи модели свързани с финансови или процеси по управление на потребителското поведение често попадат под рамките на европейското законодателство: General Data Protection Regulation (GDPR) [26], Законът за кредитните институции (в частност глава Осма, чл. 62 – чл. 66, уреждащи принципите на банковата тайна) и др.

1.3.1 Общ регламент за защита на данните

Общият регламент за защитата на данните (GDPR) [26] влиза в сила от 27 април 2016 година, като посредством него се установява единна рамка за защита на физическите лица по отношение на обработването на лични данни, както и свободното движение на информация от такъв характер в рамките на Европейския съюз. Разглежданият регламент предоставя приоритет на защита на данните като основно право на физическите лица, като същевременно гарантира правна сигурност за икономическите оператори.

1.3.1.1 Принципи на регламента

Регламентът съдържа в себе си широк набор от основни принципи, които имат за цел да установят критична рамка за работа по отношение на въпросите свързани със защита на личните данни. Обработването и съхранението на информационни масиви, свързани с личността е необходимо да се придържа към няколко основни принципа, описани в следващите параграфи.

Законност, справедливост и прозрачност. Данните, които се предоставят на организациите, имащи статут като администратор на лични данни, трябва да се обработват законосъобразно и прозрачно. Администраторът на лични данни е задължен да предоставя навременна информация по отношение на това, как и за какви цели се използват личните данни на лицата.

Ограничение на целта. Данните, които се събират е необходимо да имат конкретни и ясно дефинирани цели, които са в съответствие с регламента. Описаните информационни масиви не подлежат на последваща обработка, целяща промяна на първоначално обособените намерения. В определени случаи последващата обработка и архивиране на събраните информационни масиви и тяхното използване в исторически изследвания е възможно да бъде потвърдено при наличие на специфични гаранции.

Минимизиране на данните. Натрупването на масиви, съдържащи лична информация, подлежи на принципа на адекватност, релевантност и ограниченост само до най-съществените променливи, необходими за постигане на посочените цели.

Точност. Събраните данни подлежат на щателна проверка по отношение на акуратността, като се предприемат редица стъпки за осигуряване на актуалност и точност на информацията, от страна на администраторите на лични данни. При установяване на данни, които са неточни или не подлежат на обработка във връзка с поставените цели е необходимо те да бъдат коригирани или премахнати от общите масиви.

Ограничение на съхранението. Събраните масиви от лична информация подлежат на съхранение във форма и обхват, позволяващи идентифициране на лицата само за целите на предварително конкретизирана обработка и времеви период. След приключване на първоначално зададените времеви аспекти и цели е необходимо данните да бъдат премахнати. Те могат да бъдат архивирани за по-дълги периоди, само в случай, че се обработват за научни или статистически цели, при условие на задълбочени предпазни мерки (анонимизация).

Почтеност и поверителност. Обработването на масивите от лични данни е необходимо да се осъществява по начин съответстващ на стандартите за

гаранции за сигурност и поверителност. Тук се обособява последователност от мерки съответстващи на процесите по защита от загуба на данни, унищожаване, повреда или неразрешена и незаконна употреба.

Обработването на данните се определя като законосъобразно единствено ако отговаря на поне едно от поставените условия, като тук могат да се обособят: изрично съгласие, договорна необходимост, спазване на правно задължение или законни интереси (при условие, че те не отменят правата на субекта на данните).

1.3.1.2 Права на субектите на данни

Освен гореописаните основни принципи, които са разписани в Регламента за защита на данни, в него съществено място намират и правата на субектите на лична информация. В най-общ вид правата могат да бъдат синтезирани както следва.

Право на достъп. Субектите на лични данни имат достъп до своята информация, като могат да получат копие от съхраняваната за тях информация, както и пълната и конкретна информация по отношение на целите и получателите. Също така е наличен достъп до потвърдителна информация във връзка с текущата обработка и използване на данните. От гореописаното може да се заключи, че правото на достъп създава възможност за проверка на законосъобразността и потенциалната обработка на данните в дискретни моменти от време.

Право на коригиране. Личните данни на физически лица подлежат на своевременна корекция при установяване на неточност. Същевременно, при установяване на неточност или при непълни записи, засегнатите индивиди могат да допълнят липсващите данни, чрез предоставяне на допълнителна информация.

Право на заличаване. Заличаването на лична информация е достъпно след изрично волеизявление на субекта на данни. При постъпване на желание за заличаване е необходимо администраторите на лични данни да се съобразят с него, когато данните не са необходими за първоначалната им цел и е оттеглено съгласието от страна на физическото лице, и не е налице правно основание, което да възпрепятства премахването. Премахването се осъществява без излишно забавяне. Правото за заличаване е приложимо, в случай че данните са обработени незаконно или е необходимо да бъдат заличени, като следствие от законово задължение.

Право на преносимост на данните. Субектът на данните има достъп до събраната за него информация, като администраторът на данни се задължава тя да му бъде предоставена в структуриран, машинно четим и широко използван

формат. Личните данни могат да бъдат прехвърлени към друг администратор на лични данни, в случай че обработването на информацията се осъществява автоматизирано и въз основа на предварително волеизявление (съгласие) или договор.

Право на възражение. Субектът на събраната лична информация има право на възражение, по отношение на конкретна ситуация, обработка или в съответствие със законови практики, изпълнявани в обществен интерес. Правото на възразяване е в сила в случаите, при които събраната информация се използва за целите на директен маркетинг, профилиране и сегментиране.

Автоматизирано вземане на решения. Според Регламента, решенията, изработвани посредством събраната лична информация, подлежат на преглеждане и от експерт, а не само посредством автоматизирана процедура. Разглежданото право подлежи на изключения, свързани със ситуации, при които изработваното решение е необходимо за сключване на договорни отношения или е като следствие от изрично волеизявление от страна на субекта на данните. Дори и при така описаните случаи, администраторът е длъжен да предостави необходимите гаранции, включващи правото за намеса от страна на експерт.

1.3.1.3 Задължения на администраторите на лични данни

Общият регламент за защита на данните очертава набор от насоки и специфични условия относно оперативните изисквания в процеса на работа както на администраторите на лични данни, така и на лицата действащи от тяхно име.

Водене на записи. Специфичните изисквания по отношение на събраната лична информация, задължават администраторите на лични данни да поддържат подробен регистър на дейностите по обработване на данните, който обхваща детайлна информация за целите на обработката и описание на различните категории данни. Разглежданото задължение обхваща компании по-големи от 250 души, освен в случаите, при които е налице риск за нарушаване на правата на субектите на данните, или в случаите, при които данните съдържат чувствителна информация.

Сигурност на обработването. Едно от основните задължения на администраторите на лични данни е свързано с прилагането на набор от подходящи похвати, от технически и организационен аспект, целящи да повишат сигурността в процеса на обработка. Мерките от такъв характер са свързани с анонимизация и криптиране на събраните масиви от лични и конфиденциални данни, в пропорционално съответствие с идентифицирани рискове. Необходимо е приложените мерки да гарантират поверителността,

целостта и наличността на системите на обработка, като същевременно е осигурена възможност за възстановяване на информацията при възникване на инцидент.

Уведомление при нарушаване на сигурността. При възникване на ситуация, нарушаваща сигурността на събраните лични информационни масиви, администраторът се задължава, в рамките на 72 часа от възникване (узнаване) на събитието, да уведоми компетентните надзорни органи. В случай на висок риск от нарушаване на правата и свободите на субектите на данни, администраторът е длъжен да ги уведоми без допълнително забавяне.

Оценка на въздействието. При възникване на високо рискова ситуация, свързана с обработване на информацията с методологии, използващи нови технологии, администраторът е задължен да извърши подробна оценка за очакваните потенциални въздействия върху процесите по защита на данните. Това задължение е особено необходимо в случаите на оценка на лични аспекти, основани на профилиране, мащабна обработка на специални категории данни и в случай на обширни и систематични наблюдения на обществени места.

Длъжностно лице по защита на данните. Администраторите на лични данни са задължени да определят лице, което е отговорно за процеса по защита на личните данни, в случаите, при които обработката се извършва от публичен орган или в случаите, при които администраторът извършва системни наблюдения и обработки на данни в голям мащаб. При обработка на информационни масиви, съдържащи класифицирана информация или данни подлежащи на администриране от други законови рамки, също е необходимо да бъде определено лице, което е отговорно за защитата на данните.

1.3.1.4 Международни трансфери и прилагане

Общият регламент за защита на данните прилага набор от строги правила, удостоверяващи суверенитета и отчетността на събраната информация, особено в случаите, при които данните подлежат на трансфер извън рамките на Европейския съюз. Установената рамка е приложима и в случаите, при които данните са обект на регулярен контрол.

Международен трансфер на данни. За да бъде осъществен трансфер на лични данни към трети държави или международни организации, необходимо е Европейската комисия изрично да издаде решение за съответстващото ниво на защита относно местоназначението. В случай на липсата на такова разрешение, администраторът на лични данни, е необходимо да предостави гаранции, разглеждащи съответните корпоративни политики по защита на личните данни, като същевременно се предоставят и гаранции за ефективни правни средства при възникване на инциденти.

Независим надзор. Всяка държава членка на ЕС е задължена да създаде независим публичен орган, който осъществява наблюдение на процесите по прилагане на регламента, с цел да бъдат защитени правата на субектите на данни, като същевременно се улесни свободното движение на данни. Съществено задължение, влизащо в прерогативите на описания орган е да действа с пълна независимост от всякакво външно влияние, като същевременно разполагат с необходимите технически, административни и финансови ресурси за изпълнение на поставените задачи.

Европейски комитет за защита на данните (EDPB). Комитетът представлява независим орган на ЕС, имащ за цел да се гарантира последователното прилагане на регламента във всички държави членки. Комитетът е съставен от ръководителите на националните надзорни органи и Европейския надзорен орган по защита на данните. Основната задача на комитета е да прецизира набор от насоки и препоръки за работа, като същевременно разрешава спорове, възникнали между националните органи.

В случаите, при които прилагането на GDPR не е налице са предвидени и съответните **административни санкции**. В случай, че нарушението е извършено от страна на администратор на личните данни, орган за мониторинг или сертифициращ орган, стойността на административната санкция може да достигне до 10 млн. евро. В случаите, при които нарушението е от страна на корпоративна организация, санкцията може да достигне горепосочения размер или 2% от стойността на оборота, като се прилага по-високата сума. При наличие на по-тежки нарушения, свързани както с нарушение на основни принципи, така и при потъпкване на правата на субектите на данни, административните санкции могат да достигнат и до 20 млн. евро, а ако става дума за корпоративни организации сумата може да достигне 4% от оборота, като се прилага принципа на по-високата стойност от двете.

1.3.2 Закон за кредитните институции

Според Закона за кредитните институции от 13 юли 2006 година, обнародван в „Държавен вестник “, бр. 59 от 21 юли 2006 г. (с последно изменение и допълнение - бр. 16 от 2026), „уреждащ условията и реда за лицензирането, осъществяването на дейността, надзора за спазването на пруденциалните изисквания и прекратяването на кредитните институции (банките)“ [27], **банковата тайна** обхваща всички факти и обстоятелства, засягащи наличностите и операциите по сметките и влоговете на клиентите на дадена търговска банка.

Запазването на банковата тайна представлява съществена част от ангажиментите на широк кръг от експерти в банковата сфера и институциите, упражняващи контрол върху банковия сектор. В това число влизат както представителите на БНБ, така и членовете на Фонда за гарантиране на влоговете в банките (ФГВБ) и синдиците.

Изискванията за конфиденциалност имат продължаващо действие и остават в сила дори и след прекратяване на трудовите правоотношения на експертите, които имат достъп до информация, която попада под действието на Закона за кредитните институции. Според цитираният закон е забранено разгласяването на информация или използването ѝ за лични цели. Посредством горепосочените рестрикции се цели да се предотврати всякакъв вид облагодетелстване на служителите или техните семейства. Законодателството поставя ясно откроваща се бариера пред евентуални злоупотреби с клиентски данни. Всички финансови институции следят стриктно за спазването на високи етични и професионални стандарти.

Според Закона за кредитните институции са предвидени случаите, при които банковата тайна може да бъде разкрита. Редът за разкриване на банковата тайна подлежи на строг регламент, обособен в действащото законодателство. Дадена финансова институция може да наруши банковата тайна при изрично волеизявление от страна на титуляра. Принудителното разкриване на банкова информация се осъществява посредством процедури влезли в сила като следствие от съдебно решение.

Съгласно Закона за кредитните институции, искане за разкриване на банкова тайна може да бъде направено от органи на съдебната система и висши държавни служители, в това число: прокурор, главен прокурор, органи на съдебната власт, министърът на финансите и др. Процедурата по разкриване на банковата тайна преминава през съдебно решение, издавано на закрито заседание в рамките на 24 часа от постъпване на искането.

В банковата система и системите за финансов надзор е предвиден автоматичен обмен на финансово данни. Банковият сектор е задължен да предоставя пълен набор от информация, необходима за изпълнение на функциите на КФН и БНБ. Данните, постъпващи в тези две институции, са критични за ефективността на функциониране на централния кредитен регистър, а описаният процес осигурява стабилността на цялата банкова система.

В Закона за кредитните институции е предвидена и възможността за достъп до информация за наличностите по банкови сметки, по искане постъпило от специализирани държавни органи. Прокуратурата, органите на МВР и ДАНС разполагат с широк набор от законови правомощия при

разследване на престъпления, като процедурата отново преминава през съдебно решение. Механизмът съчетава нуждите на правосъдието с високата степен на защита на личните финансови данни.

В Закона за кредитните институции се урежда и допускането за нарушаване на банковата тайна, без да е необходимо съдебно решение. Това е възможно във връзка с осъществяване на дейността на две големи централизирани системи, поддържани от БНБ, които позволяват оторизиран достъп до определени данни.

Централен кредитен регистър (ЦКР). ЦКР представлява специализирана, централизирана информационна система, поддържана от БНБ. Тя съхранява данни за всички парични задължения на потребителите към кредитни институции. Достъп до информацията в тази система е осигурен на редица държавни институции, в това число и на: МВР, ДАНС, Прокуратура, НАП и митници, които също могат да извършват проверки без да е необходимо изрично съдебно решение.

Електронен обмен на информация за сметки и сейфовете. Системата предоставя бърза и актуална информация за текущото местоположение на активите. В нея се води подробен регистър с номерата на сметките и действителните собственици на средствата. Системата не съдържа информация за наличностите по дадените сметки.

В Закона за кредитните институции се урежда процедурата на действия за защита на личните данни при изпадане на дадена банкова институция в несъстоятелност. При възникване на банкова несъстоятелност, правният режим на банковата тайна претърпява редица изменения. Синдикът получава разширен достъп до цялата налична документация за активите и пасивите на управляването от него дружество. Също така, част от информацията, в най-честия случай това са големи вложители, може да бъде публично оповестена на сайта на Фонда за гарантиране на влоговете.

Нарушаването на правилата за конфиденциалност и защита на банковата тайна води до тежки санкции от страна на регулатора. При доказани нередности, БНБ има правомощия за отстраняване на администратори. На санкции подлежат също така физическите и юридически лица, които са нарушили Закона за кредитните институции, в частта му уреждаща защита на банковата тайна. Осъществяваният стриктен контрол цели да бъдат възпрепятствани всякакви опити за злоупотреби с клиентски данни.

1.4 Заключение

В заключение на настоящата глава може да се обобщи, че теоретичната и практическата интеграция на големите информационни масиви в управленската наука маркира фундаментална промяна в начина на извличане на знания. Големите данни са не само статичен информационен актив, а представляват динамично технологично и научно явление, дефинирано посредством сложни взаимовръзки между хетерогенни и автономни източници. Работната дефиниция в настоящия труд акцентира върху многомерни масиви, произтичащи от исторически събития или алгоритми, които носят допълнителна информация за подпомагане на процеса по изработване на информирани управленски решения.

Установява се, че ефективното боравене с големи информационни структури изисква преход от традиционни системи за управление на бази данни към мащабируеми архитектури и парадигми за паралелна обработка в разпределени клъстерни системи. Аналитичният цикъл, основан на алгоритми за машинно и дълбоко обучение, превръща суровия поток от структурирани, полуструктурирани и неструктурирани данни в основа за автоматизирано разпознаване на скрити зависимости и знания в информационните масиви.

Стойността на тези технологии се наблюдава в прогнозните анализи и детайлното анализиране на потребителското поведение. Посредством използването на технологични подходи се създават предпоставки за преминаване от реактивно към проактивно стратегическо планиране, позволяващо индивидуализиран маркетингов подход и оптимизация на риска. Интеграцията на информация в реално време и последователната обработка гарантират актуалността на моделите и предотвратяват деградацията на резултатите в динамична среда.

Паралелно с технологичния напредък, работата с големи данни е неразривно свързана със законовото съответствие и етичните норми. Рамките, очертани от Общия регламент за защита на данните (GDPR), въвеждат императивни принципи като законосъобразност, минимизиране на данните и прозрачност, които директно влияят върху архитектурата на аналитичните и прогнозиращи системи. Защитата на правата на субектите, включително правото на достъп, заличаване и преносимост, изисква висока степен на автоматизация на процесите по отчетност.

Необходимостта от защита на личната информация в процесите на работа с големи информационни масиви се наблюдава и при операциите в банковия сектор. Регламентирането на основните насоки на работа се осъществява посредством Закона за кредитните институции. Разглежданият закон поставя допълнителни изисквания относно конфиденциалността и банковата тайна,

като очертава строги процедури за разкриване на информация, преминаващо през съдебни решения или специализирани системи като Централния кредитен регистър. Нарушаването на тези регулации носи значителни административни санкции, което превръща сигурността и анонимизацията в стратегически приоритет.

Съвременната управленска наука, основаваща се на информирани решения базирани на големи информационни масиви, се основава на симбиозата между изчислителна прецизност и стриктно спазване на нормативната уредба. Големите данни се утвърждават като фундаментален инструмент за намаляване на несигурността, изграждане на дългосрочни взаимоотношения с клиентите и достигане до стратегически решения, базирани на обективни доказателства.

Използвана литература

[1] X. Wu, X. Zhu, G. -Q. Wu and W. Ding, "Data mining with big data," in IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97-107, Jan. 2014, doi: 10.1109/TKDE.2013.109.

[2] De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features.

[3] Boyd, D., & Crawford, K. (2012). Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication & Society, 15(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

[4] Gantz, J., & Reinsel, D. (2011). Extracting Value from Chaos.

[5] Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: A review of classification and combining techniques. *Artificial Intelligence Review, 26*(3), 159–190. doi.org

[6] Manyika, J., et al. (2011) Big Data: The Next Frontier for Innovation, Competition, and Productivity. San Francisco, McKinsey Global Institute, CA, USA.

[7] Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Ullah Khan, S. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, *47*, 98–115. https://doi.org/10.1016/j.is.2014.07.006

[8] Dean, J. and Ghemawat, S. (2008) MapReduce: Simplified Data Processing on Large Clusters. Communications of the ACM, 51, 107-113. https://doi.org/10.1145/1327452.1327492

[9] Chen, H., Chiang, R. H. and Storey, V. C. (2012) Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly, 36, 1165-1188

[10] Jordan, M. I. and Mitchell, T. M. (2015) Machine Learning: Trends, Perspectives, and Prospects. Science (New York, N. Y.), 349, 255-260. https://doi.org/10.1126/science.aaa8415

[11] Grolinger, K., Higashino, W. A., Tiwari, A., & Capretz, M. A. M. (2013). Data management in cloud environments: NoSQL and NewSQL data stores. *Journal of Cloud Computing: Advances, Systems and Applications, 2*(1), 22. https://doi.org/10.1186/2192-113X-2-22

[12] Gandomi, A. and Haider, M. (2015) Beyond the Hype: Big Data Concepts, Methods, and Analytics. International Journal of Information Management, 35, 137-144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

[13] Russom, P. (2011) Big Data Analytics. TDWI Best Practices Report, Fourth Quarter, 19, 1-34

[14] Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. A New Form of Web Content That Is Meaningful to Computers Will Unleash a Revolution of New Possibilities. Scientific American, 284, 1-5.

[15] F. Berman, G. Fox, T. Hey. (2003) Grid Computing: Making the Global Infrastructure a Reality. John Wiley and Sons. P747-771.

[16] Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J. and Melançon, G. (2008) Visual Analytics: Definition, Process, and Challenges. In: Kerren, A., Stasko, J. T., Fekete, J.-D. and North, C., Eds., Information Visualization, Springer, Berlin, 154-175. https://doi.org/10.1007/978-3-540-70956-5_7

[17] Shneiderman, B. (2003), The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations, The Craft of Information Visualization Readings and Reflections Interactive Technologies 2003, Pages 364-371, https://doi.org/10.1016/B978-155860915-0/50046-9

[18] Golfarelli, M., & Rizzi, S. (2009). Data Warehouse Design: Modern Principles and Methodologies. McGraw-Hill Education

[19] Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys*, *46*(4). https://doi.org/10.1145/2523813

[20] Dwork, C. (2008). Differential privacy: A survey of results. In M. Agrawal, D. Du, Z. Duan, & A. Li (Eds.), *Theory and Applications of Models of Computation* (pp. 1–19). Springer. https://doi.org/10.1007/978-3-540-79228-4_1

[21] Waller, M. A., & Fawcett, S. E. (2013). Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management. *Journal of Business Logistics*, *34*(2), 77-84. https://doi.org/10.1111/jbl.12010

[22] Erevelles, S., Fukawa, N., & Swayne, L. E. (2016). Big Data consumer analytics and the transformation of marketing. *Journal of Business Research*, *69*(2), 897–904. https://doi.org/10.1016/j.jbusres.2015.07.001

[23] Lemon, K. N. and Verhoef, P. C. (2016) Understanding Customer Experience throughout the Customer Journey. Journal of Marketing, 80, 69-96. https://doi.org/10.1509/jm.15.0420

[24] Hoyer, W. D., Chandy, R., Dorotic, M., Krafft, M., & Singh, S. S. (2010). Consumer Cocreation in New Product Development. *Journal of Service Research*, *13*(3), 283-296. https://doi.org/10.1177/1094670510375604

[25] Russom, P. (2011) Big Data Analytics. TDWI Best Practices Report, Fourth Quarter, 19, 1-34

[26] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

[27] Закон за кредитните институции, В сила от 01.01.2007 г. Обн. ДВ. бр. 59 от 21 Юли 2006г., изм. ДВ. бр. 105 от 22 Декември 2006г., изм. ДВ. бр. 52 от 29 Юни 2007г., изм. ДВ. бр.59 от 20 Юли 2007г., изм. ДВ. бр. 109 от 20 Декември 2007г., изм. ДВ. бр. 69 от 5 Август 2008г., изм. ДВ. бр. 23 от 27 Март 2009г., изм. ДВ. бр. 24 от 31 Март 2009г

2 Подходи за обезпечаване с информация

Автор: Васил Марчев

Резюме: Настоящата глава систематизира методологическите подходи за информационно обезпечаване на управленските системи, анализирайки класическите методи като: анкети, интервюта и системно наблюдение. Тези инструменти позволяват събиране на детайлни количествени и качествени данни, но се характеризират със значителна времева и финансова интензивност. Критичен аспект в процеса представлява фазата на пречупване на информацията през призмата на регулаторните рамки като GDPR и Закона за кредитните институции в частта му, уреждаща банковата тайна. Въпреки прецизността им, класическите подходи често са недостатъчни за генериране на мащабни масиви, необходими за обучение на съвременни модели, приложими в управлението.

Анализирайки недостатъците на класическите методи е предложен алтернативен подход за набавяне на данни с необходимото количество и качество, като за целта се използва набор от статистически подходи за симулация на многомерни синтетични масиви. Моделите интегрират йерархична структура от променливи, включваща демографски, социално-икономически, психологически и тясно финансови показатели. Чрез прилагане на бизнес логика и корелационни зависимости се постига висока автентичност на потребителските профили. Описаните методи минимизират правните рискове и елиминират субективността, присъща на традиционните методи. Получените данни служат като надеждна база за прогнозен анализ и изследване на пазарни тенденции и приложението им в процесите по изработване на управленски решения.

Ключови думи: подходи за информационно обезпечаване, симулация на данни, синтетични данни,

2.1 Класически подходи

Главата разглежда основните класически подходи за подбор на информация, като обособява както позитивните аспекти на разглеждания подход, така и негативните от гледна точка на процеса по обезпечаване на управленските системи с необходимото количество и качество на информационните масиви.

2.1.1 Анкетни проучвания

Анкетните проучвания представляват съществен инструмент в процесите по количествено измерване на нагласите и събиране на информация относно съществени аспекти от живота на физическите лица. Този подход позволява събиране на структурирани данни, посредством предварително дефинирани скали и затворени въпроси. Структурирана по този начин, информацията обезпечава системите за генериране на управленски въздействия. В контекста на европейското законодателство за защита на личните данни, въпросниците изискват прецизно дефинирано информирано съгласие от страна на респондента. Правилното структуриране на инструментите гарантира надеждност на резултатите и улеснява последващата статистическа обработка на данните.

При прилагането на разглежданата методология е необходимо строго съблюдаване на принципите за минимизиране на обработваната информация. За последваща работа, събраните масиви от данни е необходимо да бъдат подложени на допълнителна анонимизация с цел сигурност. Академичните стандарти изискват детайлно описание на извадката и методологията на разпространение на анкетните карти. Така се осигурява прозрачност на научните изследвания в съответствие с етичните норми за работа със субекти на данни.

Съществена фаза в процеса по набиране на информация посредством анкетни проучвания е необходимостта от фокус върху процеса по операционализация на променливите. Този етап се фокусира върху трансформацията на теоретичните постановки в конкретни, измерими въпроси, които гарантират вътрешната валидност на научния труд [1]. Необходимо е да се установи баланс между различните типове въпроси, с цел да се оптимизира качеството на събраната първична информация, като същевременно се цели избягването на пристрастност по отношение на въпросниците и очакваните отговори [2]. Друга съществена част от процеса е правилното дефиниране на измервателните скали, което се разглежда през призмата на по-сложни статистически методи за анализ на зависимостите.

Осигуряването на представителност на извадката представлява един от най-критичните моменти в при работа с анкетни методи. Приложим вариант за справяне с тази трудност може да бъде използването на вероятностни методи за подбор, с цел постигане на валидност и обобщаване на изводите. Необходимо е данните да бъдат подлагани на системен анализ и контрол, с цел установяване на наличие на системна липса на отговор от определени групи физически лица.

В контекста на GDPR и обработката на данни на физически лица, сигурността на платформите за онлайн анкетиране е от критично значение.

Трансграничният пренос на информация изисква сървърите на доставчика да бъдат локализирани в рамките на Европейското икономическо пространство. Изследователите са длъжни да внедрят технически мерки като криптиране преди и по време на пренос. Тези гаранции защитават правата и свободите на субектите и повишават доверието в самия научен процес.

Като съществен недостатък в разглеждания подход може да бъде обособен рискът от ниска степен на валидност, породен от субективността в процесите на самоотчитане. Наблюдава се специфика, според която респондентите често дават социално желателни отговори или интерпретират въпросите по начин, различен от заложения в съответното изследване. На лице е и възможността за установяване на нисък процент на отговорилите, което също допринася за компрометиране на представителността на цялото изследване. Поставените въпроси са задълбочено разгледани от Couper [3].

Друго ограничение е свързано с твърдата структура, която не позволява гъвкавост в хода на самото проучване. Ако в дизайна на въпросника е допусната системна грешка, тя рефлектира върху целия масив от данни. Методическите предизвикателства изискват внимателна предварителна проверка и пилотно тестване на всеки инструмент.

Също така, като съществен недостатък от управленска гледна точка, могат да бъдат обособени сложността и цената на разглежданата методология. Събирането на информация посредством анкетни проучвания представлява времеемък процес, съпътстван от допълнителни разходи свързани както с времевия аспект, така и с необходимата последваща анонимизация на събраните информационни масиви.

2.1.2 Интервюиране

Интервюирането по същество представлява качествен метод, осигуряващ дълбочина и контекст на изследваните социални и поведенчески феномени. Чрез директна комуникация може да се извлече специфична информация, която често остава скрита при количествените методи [4]. Този процес изисква внимателно планиране на въпросите и висока степен на емпатия. Събраните данни се характеризират с висока степен на детайлност по отношение на личностните аспекти на изследваното лице.

Процесът на провеждане на интервюта изисква детайлно планиране на взаимодействието между изследователя и субекта на данни. Изборът между полуструктурирано и задълбочено интервю определя рамката, в която физическите лица разкриват своя субективен опит и възприятия [5]. Ключово предимство тук е възможността за адаптиране на въпросите в реално време въз

основа на получените отговори. Това позволява идентифициране на скрити причинно-следствени връзки, които количествените методи често пропускат.

Валидността на данните от интервютата зависи в голяма степен от уменията на водещия да избягва насочващи въпроси. Анализът на полученото съдържание изисква систематично кодиране, за да се превърнат изложенията в структурирано научно знание. Използват се разнообразни софтуери, подпомагащи процеса за управление на големите обеми от текстова информация. Правилната интерпретация на контекста е критична за научната стойност на получените изводи.

От гледна точка на GDPR интервюирането поставя специфични предизвикателства пред защитата на неприкосновеността. От гледна точка на разглежданата методология е наложително прилагането на строга политика за анонимизация при транскрибирането. Съгласието на участника трябва да обхваща не само участието, но и специфичния начин на съхранение на цифровите записи. Изследователят е длъжен да осигури прозрачност относно това кой ще има достъп до суровия материал.

Защитата на личните данни изисква внедряването на стриктни протоколи за криптирано съхранение и ограничен достъп до първичните материали. Транскрибирането на разговорите трябва да бъде придружено от премахване на всички идентифициращи белези на субекта. Тези стъпки са ключови по отношение на легитимността на изследването в рамките на регулаторната рамка на ЕС.

Разглежданата методология изисква значителен ресурс както във времеви, така и във финансов аспект, свързан най-вече с последващата обработка на информационните масиви. Също така, налице е висок риск от предубеденост в процеса на събиране на първичната информация, което поставя под въпрос бъдещите резултати, базирани на събраната информация. Съществен риск крие и степента на анонимизация на получените данни.

Разглежданата методология за набиране на информационни масиви е подходяща при малки изследвания, базирани на ограничен брой частни случаи. Този процес не е приложим в случаите, при които е необходимо голямо количество данни, което да послужи в процеса по обучаване на управленски или финансови модели, поради невъзможността да бъде събрано необходимото количество информация в разумни времеви граници.

2.1.3 Системно наблюдение

Системното наблюдение е класически подход за събиране на данни чрез непосредствено регистриране на действията на физическите лица. То позволява обективно документиране на поведението на потребителя в неговата

естествена среда. Методът ограничава риска от субективни изкривявания при самоотчитането от страна на респондентите. Данните се записват хронологично и се класифицират според предварително изготвени кодиращи схеми.

Подходът осигурява достъп до реалното поведение на физическите лица, което често се различава от декларираното в анкети или интервюта. Процесът изисква предварително разработване на единна категоризираща система, с цел да се гарантира обективност и повторяемост на резултатите. Високата степен на прецизност при документирането превръща суровите наблюдения в количествено измерими данни, подходящи за използване при обучение на модели за управление на потребителското поведение.

Използването на технологични средства за наблюдение, като видеокамери или сензори, изисква ясна правна база и прозрачност спрямо субектите. Резултатите от наблюдението трябва да се съхраняват по начин, който изключва случайно разкриване на самоличността на участниците [6]. Тези процедури гарантират, че научният прогрес не компрометира основните права на личността.

В контекста на европейското законодателство (GDPR), системното наблюдение поставя сериозни въпроси относно правото на личен живот и неприкосновеност. Съществен момент при провеждане на подобни изследвания е да се определи дали събраната информация позволява идентифициране на конкретни физически лица чрез визуални белези или метаданни [7]. Внедряването на принципи за анонимизация при проектиране на системите за на записване изиграва ключова роля за потвърждаване на легитимността на изследването.

При провеждане на наблюдение е наложително да се балансира научният или корпоративният интерес с правото на неприкосновеност на личния живот. Субектите трябва да бъдат информирани за присъствието на наблюдател, освен в специфични случаи на публично поведение. Визуалните данни от наблюдението подлежат на засилен контрол и не могат да бъдат споделяни без правно основание [8]. Регламентацията изисква ясно документиране на времето, мястото и обхвата на извършваното научно проследяване.

Прилагането на анализираната методология е свързано със значителни времеви и финансови ресурси, обусловени най-вече от комплексната последваща обработка на информационните масиви. Допълнително предизвикателство представлява гарантирането на адекватна степен на анонимизация на генерираните данни в съответствие със съвременните изисквания.

Този методологичен подход е приложим предимно при изследвания в ограничен мащаб, концентрирани върху специфичен брой казуси. Методът се оказва практически неприложим в ситуации, изискващи мащабни масиви от данни за обучение на модели, поради системната трудност да се акумулира необходимият обем информация в приемливи срокове.

2.1.4 Дистанционни методи за набиране на информация

Дистанционните методи включват набор от подходи, използващи технологични средства за събиране на данни от разстояние без пряк контакт. Те обхващат анализ на цифрови следи, телеметрия и други пасивни форми на регистриране на активността [11]. Въпреки че са технологично напреднали, те често се причисляват към класическите заради своя емпиричен характер. Този подход позволява обхващането на огромни масиви от информация за кратък период от време.

Дистанционните методи създават предпоставки за събиране на информация за физически лица посредством използването на цифрови технологии и телеметрични системи. Този подход позволява регистрирането на обективни „цифрови следи“, като метаданни от комуникация, лог-файлове или сензорна активност в реално време. За научните изследвания това е ценен източник, тъй като данните се генерират пасивно, без пряка намеса в ежедневието на субекта. Тези методи са способни да обхванат огромни масиви от информация, които надхвърлят възможностите на традиционните методи за събиране на данни [9].

От гледна точка на GDPR, дистанционното събиране на данни е зона с висок риск поради възможността за автоматизирано профилиране [10]. Дори когато данните изглеждат технически анонимизирани, съчетаването на различни цифрови следи често води до де-анонимизация на конкретно физическо лице. Необходимо е изследователят да гарантира, че обработката е прозрачна, както и че субектите са информирани за обхвата на дистанционното проследяване. Прилагането на принципа на ограничаване на съхранението е императивно за минимизиране на потенциалните вреди.

Дистанционните подходи изискват специфична техническа експертиза за филтриране на „шума“ в големите масиви от данни. Анализът често се базира на алгоритми за машинно обучение, което налага допълнителна проверка за етични пристрастия в самите модели. Въпреки че тези методи предлагат мащабируемост, те не винаги могат да обяснят мотивите зад дадено поведение. Ето защо в научните трудове и в света на бизнеса те често се комбинират с класически подходи за постигане на по-голяма дълбочина.

Предизвикателствата при тези методи са свързани предимно с автоматизираното профилиране на физическите лица. Научната етика изисква пълна прозрачност относно алгоритмите, използвани за извличане и филтриране на получената информация, което в пряка корелация с принципите застъпени в GDPR. Необходимо е да бъдат гарантирани правата на субектите при установяване на дистанционен достъп, с цел съблюдаване на европейското законодателство и Закона за кредитните институции, в частта му уреждаща принципите на банковата тайна.

Описаната методология осигурява възможност за анализ на големи информационни масиви, като акцентира върху мащабируемостта на използваните подходи. Също така се откриват и перспективи при анализа на данни в реално време и елиминиране на субективизма в процеса на работа. Разглежданите подходи осигурява висока продължителност на наблюдение на субектите, с цел постигане на мониторинг на тенденции за дълги периоди. Посредством това се създават предпоставки за осигуряване на големи информационни масиви за обучаване на модели базирани на данни, които обезпечават изработване на управленски решения, основани на емпирични изследвания.

При използването на описаната методология се регистрират и набор от отрицателни въздействия. Типичните негативни явления са свързани с липсата на контекст в извличаните данни. Установяват се действията на субекта, но липсата на допълнителен контекст не предоставя информация относно причините за действията на изследваното лице [12]. Други негативни аспекти в процеса са свързани с изискваната висока изчислителна мощ и специфични софтуерни умения, което допълнително оскъпява процеса по обезпечаване с необходимите данни. Също така е необходимо да се подчертае и необходимостта от последващо анонимизиране на събраните масиви от данни, с цел удовлетворяване на принципите на GDPR.

2.2 Симулация на данни

2.2.1 Постановка на задачата

В информационните източници се наблюдава недостиг на набори от данни, съдържащи чувствителна информация, например лична и индивидуална информация, банкови данни и др. Информационните масиви, обобщени в тази група, попадат в рамките на европейското законодателство (GDPR), Закона за кредитните институции в частта му уреждаща банковата тайна и други. Това от своя страна затруднява изследователите да осигурят необходимите набори от информация за моделите основани на данни.

Научните изследвания, базирани на големи масиви от синтезирани бази данни, предполагат набор от предварителни действия, насочени към получаване на необходимата информация както в количествено, така и в качествено отношение. Спецификата на данните, тяхната комплексност и правните рамки предполагат редица различия в подходите за генериране на необходимите набори от информация.

При описаните по този начин обстоятелства се наблюдават два основни подхода за справяне с наблюдаваната ситуация: стандартни/класически методи и алтернативни подходи. При стандартните /класически/ методи (описани в предходната част на настоящата глава) изследователите получават достъп до необходимата им информация чрез стандартните начини за анкетиране и попълване на съответната информация. В самата си същност, това представлява сложен и бавен процес. Спецификата на необходимата информация, при работа с чувствителни лични данни, предполага негативен отклик от страна на потенциалните респондентите. Продължителността във времеви аспект се увеличава и от необходимото последващо анонимизиране на вече събраните набори от данни. Като се вземат предвид описаните по-горе трудности, се установява необходимост от структуриране на алтернативен подход за получаване на потребните масиви от данни. Този подход се основава на набор от статистически методи, чрез който се симулира информационен масив въз основа на предварително избрани фактори, които включват демографски, личностни, индивидуални и банкови променливи и показатели.

За целите на поставената задача се използва набор от разнообразни статистически техники, включващ общи разпределения на данните за всеки един от избраните фактори, бизнес логика, корелации, обърната копула за генериране на извадки и др. Разпределенията, като съществен елемент от процеса, са получени като следствие от изследователска дейност, априорни допускания, анализ на банковата среда и др.

Главата предоставя задълбочен преглед на възможностите за генериране на многомерен масив от данни, интегриращ персонални и демографски характеристики, личностни преференции и др., приложим в управлението на бизнес структури, банковото дело, финансовите пазари и други икономико-финансови области.

В настоящата глава се разглеждат множество разнообразни личностни характеристики, обособени в отделни групи. Всяка една от разгледаните променливи спомага за изграждане на изчерпателен профил на статистически значимо множество от физически лица. Установяват се финансово активните индивиди и тяхното поведение във финансово-икономически аспект. Това

способства събраният масив от данни да намери приложение в широк набор от икономически и социални изследвания.

2.2.2 Основни групи променливи

Обособени са няколко основни групи променливи, които спомагат за разработване на многомерен набор от данни. В зависимост от проведените наблюдения и анализи, както и априорни знания относно динамиката на финансовите пазари се очертават следните групи променливи:

2.2.2.1 Демографски променливи

Демографията е научна дисциплина, която систематично изследва структурата, динамиката и пространственото разпределение на човешките популации. В рамките на настоящото изследване тя се разглежда като критичен аспект при формирането на синтетични данни насочени към обезпечаване на модели в управление на финансови структури. В съвременния изследователски контекст демографските данни служат като фундамент в процеса на установяване на социално-икономическите процеси и за формиране на дългосрочни стратегически прогнози. Посредством прилагане на статистически парадигми и методи, демографията спомага за идентификацията на тенденции, които са от съществено значение както за публичното планиране, така и за частните корпорации при създаването на прогнозиращи модели.

При разработването на съвременни модели за генериране на синтетични данни, демографските променливи играят ролята на основни параметри, гарантиращи репрезентативност и точност на симулациите. Моделите приоритетно интегрират водещи показатели като пол, възраст, семейно положение и образование, които се използват като основна рамка за социалните структури и установяват основните насоки за анализ на потребителското поведение. Наред с тях, за постигане на по-висока детайлност на анализа се включват и голям набор от по-второстепенни характеристики, сред които етническа и религиозна принадлежност, националност и др. Тези допълнителни характеристики позволяват на моделите да уловят специфичните културни и социални нюанси, които влияят върху динамиката на ниво микро група. Интеграцията на такъв широк спектър от променливи осигурява висока степен на представителност на генерираните информационни набори.

2.2.2.2 Социално-икономически променливи

Към тази подкатегория спадат основни фактори като: социално-икономически статус, приходи, персонални разходи, разходи за данъци,

разходи за жилище и по-второстепенни променливи като: застрахователни разходи, разходи за хоби и забавление, и др.

Социално-икономическите променливи представляват комплексна група от характеристики, която разглежда в детайли позицията на физическите лица в социално-икономическата йерархия, посредством съчетаване на финансовото благосъстояние, както и социалния и обществен статус на лицата. От гледна точка на настоящото изследване тази група от променливи се разглежда като фактор за по-лесно обособяване на хомогенни групи и подобряване на взаимодействията между променливите. При изграждането на аналитични модели социално-икономическият статус се разглежда като група от променливи с висока прогнозираща мощ. Тази група от променливи спомага за установяване на корелации между жизнения стандарт и дългосрочните поведенчески модели на населението. По този начин се осигурява надеждна основа за подобряване на моделите за синтезиране на информационни масиви, съответстващи на динамичното съвременно общество.

При проектирането на модели за генериране на данни, социално-икономическите променливи представляват набор от характеристики, които повишават точността на използваните методологии, по отношение на подобието на новогенерираните данни спрямо реалните масиви. Към основните фактори се причисляват общите приходи, персоналните разходи, данъчните задължения и разходите за жилище, които формират базисната икономическа структура на индивида. Тези водещи параметри са критични за определяне на разполагаемия доход и финансовата стабилност в рамките на генерирания масив от данни. За постигане на по-висока детайлност в модела се интегрират и вторични променливи, свързани със застрахователните навици на популацията, както и разходи за хоби и забавления. Добавянето на по широка гама от специфични характеристики позволява на модела да пресъздаде реалистичен потребителски профил и да отрази начина на живот отвъд задоволяването на първичните потребности. Тази многопластова структура на данните е от съществено значение за провеждане на задълбочени анализи в сферата на управленските науки.

2.2.2.3 Психологически и личностни променливи

Психологическите и личностни променливи са група от индикатори, които могат да бъдат разгледани като фундаментален слой при осъществяване на анализ на човешкото поведение. Наборът от променливи, съставляващи тази група, изследва корелацията на различните записи от гледна точка на влиянието им върху социално-икономическите и финансови променливи. Крайъгълен камък в обособената група заема изследването на темперамента.

Темпераментът се дефинира като съвкупност от вродени характеристики от психологическа гледна точка, които определят интензивността и скоростта на реакциите, по отношение на емоциите и поведението, на индивида. Той се разглежда като фундамент, върху който се изгражда характера, посредством взаимодействие с околната среда и социалния опит. Анализът на разглежданата група от променливи дава възможност да се проучи както цялостния фокус на индивида, така и специфичната динамика в действията при различни стимули. Като следствие от казаното до тук психологическият профил, може да бъде описан като ключов инструмент за прогнозиране на индивидуалните различия в социална и професионална среда.

В процеса по генериране на многомерни масиви от синтетични данни, психологическите и личностни характеристики имат за задача да осъществят в значителна степен подобряването на автентичността на новоизградените профили (записи). Посредством разглежданите променливи се цели да се симулира последователност от реалистичен реакция спрямо събития от околната среда, които са установени на база на личностните черти на индивида и особеностите на темперамента му. Наблюдаваната група от характеристики представлява значим елемент в процесите по симулация и анализ на отклика по отношение на пазарните промени и последващите от това взаимодействия в социален аспект. Внедряването на личностни характеристики повишава прецизността на генерираните данни, като осигурява не само статистическата вероятност на генерираните масиви, а също така и психологическият елемент, който присъства при всички реални информационни масиви. Широкият спектър от променливи, които се използват за генериране, създава предпоставки за повишаване на надеждността прогнозиращите модели използвани в управленските системи.

2.2.2.4 Индивидуална склонност към риск

В динамичната среда, съпътстваща процесите по инвестиране на финансови активи, като съществен фактор се разглежда индивидуалната склонност към риск. Този индикатор е в пряка корелация с нивата на неопределеност, свързани с инвестирането на финансови активи. От гледна точка на настоящото изследване, индикаторът се дискутира като строго индивидуална променлива. Това се дължи както на разликите в личностните характеристики на инвеститорите, така и на опита, който те притежават в търговията на организирани пазари.

В процесите по установяване на цялостен синтезиран набор от данни се търси корелацията, която характеризира индивидуалната склонност към риск и психологическите и личностни характеристики. Разграничаването на

реакцията по отношение на негативните събития, които може да съпътстват процесите по инвестиране се намира в основата на коректната оценка за индивидуална склонност към риск. Настоящата работа цели да дефинира изследваната характеристика като компонент от цялостен масив от синтетични данни, приложими в дейността на икономическите субекти, академичните среди и анализаторите на големи информационни масиви.

2.2.2.5 Финансови и банкови променливи

Съществен аспект в процеса по генериране на синтетични масиви от данни, за целите на настоящата разработка, представлява коректното изграждане на задълбочен профил на активния потребител на банкови и финансови услуги. Разглежданата съвкупност от характеристики се фокусира върху установяването на специфичните необходимости и предпочитания на физическите лица, по отношение на достъпните финансови инструменти и услуги.

Посредством наблюдение на поведението на потребителите на банкови услуги и процесите по управление на финансовите наличности, аналитичните системи създават предпоставките за установяване на набор от устойчиви поведенчески модели. Наблюдаваните характеристики целят да се идентифицира мотивацията, стояща зад последователността от решения, съпътстваща управлението на финансови активи. По този начин се създава възможност за създаване на клиентски профили и подобрено сегментиране на потребителите. Използването на подобен набор от променливи осигурява детайлен разрез на клиентските потребности, като създава възможности за подобряване на клиентската удовлетвореност, разкривайки механизмите на финансовата динамика.

Обособяването на съвкупност от личностни преференции на физически лица създава предпоставки за установяване на подбудите, които тласкат различните потребители към избора на набор от конкретни банкови и финансови продукти. Моделите за синтезиране на данни разглеждат този аспект на променливите като интегрират в алгоритмите си критични променливи като: лични инвестиционни цели, планиран срок на инвестицията и индивидуална склонност към риск. Внедряването на такива характеристики позволява детайлен разрез на индивидуалните стратегии на физическите лица по отношение на управлението на финансови активи. Наблюдаването на финансовите и банкови променливи има съществена роля в последователността на генериране на реалистични симулации на пазарното поведение. Интегрирането на вторични фактори, като канали за банкиране и

честота на използвани кредитни инструменти създава една цялостна картина на съвременния активен потребител на финансови услуги.

С помощта на така представената структура на данните се създава възможност за по-пълен и адекватен анализ на поведението на пазарите на финансови услуги, като се откриват възможности за подобряване на прогнозната точност на моделите за установяване на бъдещите тенденции.

2.3 Заключение

Настоящата глава разглежда методологическото обезпечаване на съвременните управленски системи. Това е процес, който е изправен пред необходимостта от съчетаване на нуждата от мащабни информационни масиви и строгите регулаторни рамки за защита на личните данни. Анализът на класическите инструменти като анкетни проучвания, интервюта и системно наблюдение показва, че тяхната специфика да осигуряват висока степен на детайлност и контекстуална дълбочина се сблъсква с приложимостта им в сферата на големите данни. Разгледаните подходи са силно ограничени от значителния времеви и финансов аспект. В условията на настоящата високо технологична икономика тези методи се оказват недостатъчни за генериране на необходимите количества информация, приложими в процесите по обучение на модели, предназначени за управление на потребителското поведение и банковите процеси.

Дистанционните методи за набиране на информация създават предпоставки за набавяне на необходимото количество от данни. Те от своя страна поставят набор от въпроси относно автоматизираното профилиране и риска от де-анонимизация. Предизвикателствата, произтичащи от GDPR и Закона за кредитните институции, в частта му уреждаща принципите на банковата тайна, налагат търсенето на алтернативни подходи, които да гарантират както аналитичната стойност на масивите, така и правната сигурност на изследователския процес. В разглеждания контекст предложеният модел за симулация на многомерни синтетични масиви се очертава като най-ефективното иновативно решение. Чрез интегрирането на набор от променливи, обхващащи демографски, социално-икономически, психологически и финансови показатели, се постига високо ниво на автентичност на профилите без директно използване на чувствителни лични данни.

Елиминирането на субективността, характерна за традиционните методи, и прилагането на строга бизнес логика и корелационни зависимости при генерирането на синтетични данни, създава надеждна основа за прогнозно

моделиране и стратегическо планиране. Синтетичните масиви не просто запълват информационния дефицит, но и трансформират начина, по който управленските системи взаимодействат с емпиричните доказателства.

Процесите по информационното обезпечаване в управленските науки, от гледна точка на съвременната технологична реалност, преминават през прецизното комбиниране на статистическите симулации с дълбокото разбиране на човешкото поведение, което гарантира едновременно технологичен прогрес и спазване на високите стандарти за неприкосновеност в глобалната информационна среда.

Използвана литература

[1] Saris, W. E., & Gallhofer, I. N. (2014). Design, Evaluation, and Analysis of Questionnaires for Survey Research. John Wiley & Sons.

[2] Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. Annual Review of Psychology.

[3] Couper, M. P. (2000). Web surveys: A review of issues and approaches. Public Opinion Quarterly.

[4] Boyce, C., & Neale, P. (2006). Conducting in-depth interviews: A guide for designing and conducting in-depth interviews for evaluation input. Pathfinder International.

[5] Legard, R., Keegan, J., & Ward, K. (2003). In-depth interviews. In J. Ritchie & J. Lewis (Eds.), *Qualitative research practice: A guide for social science students and researchers* (pp. 138–169). Sage Publications.

[6] Spiekermann, S. (2012). The challenges of privacy by design. Communications of the ACM, 55(7), 33–35. https://cacm.acm.org/opinion/the-challenges-of-privacy-by-design/

[7] European Data Protection Board. (2020, January 30). *Guidelines 3/2019 on processing of personal data through video devices* (Version 2.0). https://www.edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-32019-processing-personal-data-through-video_en

[8] McKechnie, L. E. F. (2008). Observational Research. In L. M. Given (Ed.), The SAGE Encyclopedia of Qualitative Research Methods (pp. 573-577). Sage.

[9] Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures & their consequences*. SAGE Publications Ltd, https://doi.org/10.4135/9781473909472

[10] Zarsky, Tal, (2017), Incompatible: The GDPR in the Age of Big Data (August 8, 2017). Seton Hall Law Review, Vol. 47, No. 4(2), 2017, Available at SSRN: https://ssrn.com/abstract=3022646

[11] Salganik, M. J. (2017). Bit by Bit: Social Research in the Digital Age. Princeton University Press.

[12] Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878.

3 Същност на синтетични данни. Приложение в управлението на бизнес структури

Автори: Васил Марчев, Ангел Марчев, мл.

Резюме: Настоящата глава предоставя подробен преглед на класификацията и методологията за генериране на синтетични данни в ерата на големите данни. Направени са съпоставки между частично и напълно синтезираните масиви, както и диференциацията между вероятностните подходи и тези базирани на техники за машинно обучение, невронни мрежи и правила, пречупени през призмата на симулацията на многомерни набори от синтетични данни. Главата подчертава стратегическото значение на описаните технологии за преодоляване на информационния дефицит и оптимизиране на бизнес процесите.

Акцентът е върху стратегическата роля на синтетичните данни като ефикасно средство за преодоляване на правните и регулаторни ограничения, произтичащи от GDPR и Закона за кредитните институции. Прилагането на генеративни методи предоставя възможност за създаване на мащабни информационни масиви, които съхраняват статистическите зависимости на реалния свят, без да компрометират поверителността на индивидуалните субекти. Съществен момент в приложението на синтетични данни е свързан с феномена „студен старт“ при внедряването на нови аналитични системи, нови продукти или проникване на непознати пазари, позволявайки по-бърза реализация и оптимизация на бизнес стратегиите още в начална фаза. Налице са редица предимства по отношение на разходите и скоростта по обезпечаване на моделите с данни, но настоящата глава подчертава критичното значение на валидирането на данните, за да се избегнат рискове от несъответствие на синтезираните масиви с реалните данни.

Ключови думи: синтетични данни, генериране на данни, вероятностни модели, машинно обучение

3.1 Въведение

Процесът на анализ и моделиране на данни обхваща няколко основни момента. Значително място заема фазата на установяване на информация с необходимото количество и качество от гледна точка на модела. Спецификата на разглежданите информационни набори, тяхната комплексност и юридическото обезпечаване на процеса, свързан с приложение на европейските

регламенти по защита на данните, предполагат редица трудности при осигуряване на необходимите масиви. Данните, фокус на настоящото изследване, попадат в рамките на GDPR и Закона за кредитните институции в частта му уреждаща въпросите за банковата тайна.

Фазата на събирането на данни посредством познатите класически методи за анкетиране и интервюиране, разгледани в предходната глава, представлява сложен и бавен процес. Като се вземат като отправна точка гореописаните трудности в последователността по обезпечаване с данни, авторите идентифицират алтернативен подход за получаване на необходимото количество информация. За целта е използван набор от подход за симулиране на синтетична база данни, основана на предварително синтезирана съвкупност от различни демографски, лични, индивидуални и банкови променливи и показатели.

По своята същност синтетичните данни представляват масив от записи, които се генерират посредством статистически зависимости, интегрирани в софтуерни алгоритми. От своя страна тези информационни масиви намират приложение в обучението на валидирането на модели за управление, могат да бъдат използвани като тестови данни и др. Предимствата на този вид информация са свързани с липсата на ограничения, както и с възможността, която те предоставят за запазване на взаимовръзките между различните променливи.

Основните недостатъци на разглежданите подходи са свързани с трудностите, които възникват в процеса по генериране на данни. Качеството на новосъздадените масиви е в пряка корелация с уместността и релевантността на използвания метод. Друг наблюдаван недостатък е свързан с необходимостта от фаза, в която новите данни подлежат на процес по валидиране, посредством реален масив от информация. В този случай е необходимо да се осигури реален набор от данни - с цел валидация.

Поради спецификата на различните емпирични изследвания, се установяват редица случаи, при които е необходимо, за целите на успешното моделиране или аналитичен процес, да се използва комбинация от различни симулационни методи, за да се постигне необходимото количество и качество на данните.

3.2 Синтетични данни

3.2.1 Същност

Технологичното настояще и все по-нарастващите необходимости от детайлни прогнозиращи модели установяват критичната необходимост от

масиви от данни, съдържащи информация, посредством която може да бъде подобрена аналитичната и прогнозираща мощ, по отношение на клиентското поведение и необходимости. Натрупването на все по-големи масиви от информация предполага задълбочаване в сферата на защитата на личните данни в кореспонденция с европейското законодателство (GDPR) и Закона за кредитните институции, в частта уреждаща въпроса с банковата тайна. В някои случаи спецификата на необходимите данни, сложността и законосъобразността предполагат редица трудности при осигуряването на модела с необходимото количество и качество на информация [1].

Методите за генериране на синтетични данни предлагат потенциално решение на този проблем чрез създаване на многомерни набори от данни, които интегрират различни променливи, включително демографски, поведенчески, финансови и лични характеристики. Генерираните данни не са обект на посочените ограничения, тъй като не идентифицират реални лица, като същевременно запазват разпределенията и връзките между променливите. Фундаменталната необходимост от подробни набори от данни, от една страна и такива, които са с достатъчно голяма степен на сигурност, от друга страна, създават предпоставки за възникване и развитие на синтетични информационни набори.

Генерирането на информационни масиви е ефективна техника за обработка на небалансирани данни, инженерство на характеристики, обработка на липсващи данни, отговаряне на изискванията за поверителност на данните, справяне с реалния недостиг на данни и постигане на икономии на разходи, а също така и ефективна техника за последващ анализ на информацията. Налице е широк набор от разнообразни подходи за генериране на многомерни набори от данни и множество автори и водещи организации в областта на изкуствения интелект са разглеждат засегнатата проблематика [2].

Събирането и съхраняването на реални данни представлява скъп и бавен процес. Съществен момент в работата с големи информационни масиви представлява необходимостта от намаляване на разходите по набиране и съхранение на данните. От тази гледна точка синтетичните данни се разглеждат като устойчива алтернатива на реалните масиви и тяхната сложност, времеемкост и капиталоемкост по набавяне, съхранение и използване. Разходите за производството на синтетична информация представляват незначителна част в сравнение с необходимите за използване на реални данни ресурси.

3.2.2 Област на приложение

Синтетичните данни са подходящи и най-често се използват за тестване, валидиране и обучение на модели за машинно обучение. „Едно от многото предимства на синтетичните данни е, че те могат да се генерират в голям мащаб, осигурявайки изобилие от данни за обучение и тестване на модели с изкуствен интелект. Това е особено ценно в области, където данните от реалния свят са оскъдни или трудни за получаване“ [3]. Друго предимство, което произтича от използването на синтетични данни, е, че новополучените данни „могат да бъдат пригодени към специфични изисквания, като например осигуряване на балансирано представяне на различните класове чрез въвеждане на контролирани вариации“ [3].

Математически този процес може да се обозначи като:

$$D_{synthetic} = f\left(D_{real}\right)$$

Където с $D_{synthetic}$ са означени синтетичните данни. Това са данни които са получени като функция на данните от реалния свят, като се съпоставят реалните данни със синтетичния им еквивалент [4].

Във финансовата индустрия разпоредби като Закона за кредитните институции и по-конкретно, в частта му уреждаща банковата тайна, налагат защитата на чувствителна информация, включително данни за трансакции, банкови сметки, заеми и други. Чрез използването на синтетични набори от данни, банките и регулаторните органи могат да осъществяват анализи в процеса за борба с прането на пари, прогнози за развитието на финансовата система и др., без да се прави компромис с поверителността на реалните клиенти.

Синтетичните данни насърчават иновациите, като същевременно гарантират поверителност и спазване на разпоредбите. Чрез използването на усъвършенствани статистически техники и такива за машинно обучаване се създават реалистични набори от данни, които се използват от организациите и анализаторите с цел да експериментират и да изграждат модели, дори когато са изправени пред строги регулаторни ограничения. Освен това, синтетичните данни могат да помогнат за смекчаване на отклоненията или пропуските в събраните набори от данни чрез свръхсемплиране на недостатъчно представени групи или сценарии, като по този начин се повишава справедливостта и надеждността на моделите.

Синтетичната информация представлява критичен инструмент за балансиране на аналитичната полезност и защитата на поверителността в рамките на процесите по управление на бизнес структури. Тя представлява

стратегическо решение за оптимизиране на бизнес процесите и преодоляване на информационния дефицит. Приложението на усъвършенствани техники за синтез, създава предпоставки организациите да подобрят своята аналитична и прогнозираща мощ по отношение на клиентското поведение и необходимости. Това е особено ценно в области, в които реалните данни са оскъдни или трудни за получаване.

Моделите за генериране на симулирани данни са специфични, което позволява широк спектър от приложения. Генерираните информационни набори могат да се използват в области като медицина, финанси и автономни превозни средства. Разглежданите масиви намират своето приложение и в процесите на моделиране, тестване и отстраняване на грешки за аналитични и прогнозни задачи. Тази възможност позволява интегрирането им както в тестови среди, така и в продуктивна среда. Съществен момент на приложение в реални условия е използването им по време на студен старт (внедряване на нови продукти или проникване на непознат пазар). Използването на синтетични данни може да доведе до по-бързи иновационни процеси и подобрения в услугите.

Освен описаните до момента приложения, синтетичните данни се утвърждават като стратегическо решение за елиминиране на феномена „студен старт“ при внедряването на нови аналитични системи, погледнато през призмата на бизнес контекста. Чрез симулиране на обширни масиви от потребителско поведение и трансакционни исторически записи, организациите могат да обучат своите прогнозиращи модели без да изчакват натрупването на реална историческа информация [20].

Този подход позволява бързо и адекватно обособяване на персонализирани препоръчителни алгоритми, което може да бъде разгледано като критична фаза, съпътстваща внедряването на нови продукти и/или услуги. Прилагането на усъвършенствани методи за синтез гарантира, че генерираните масиви притежават необходимата статистическа правдоподобност от гледна точка на управлението на бизнес процеси, като същевременно с това се осигурява и високо ниво на защита на конфиденциалността на информацията. Синтетичните данни, в контекста на управлението на процесите повлияни от явлението студен старт, създават предпоставки за подобряване на прогнозиращите модели и създаване на един по-пълен разрез на планираните иновативни продукти и/или услуги, като същевременно се намалява времето за обучение на моделите и се намалява възможността за изработване на неподходящи бизнес стратегии в начална фаза.

3.2.3 Предимства и предизвикателства

3.2.3.1 Предимства

Синтетичните данни предлагат редица предимства, които ги правят привлекателен вариант за изследователи и практици. Едно от основните предимства на синтетичните данни е, че те се създават или генерират, а не се събират от реалната среда. Едно от съществените предимства е липсата на ограничения, с които се свързват данните от реалния свят, като например опасения за поверителност и регулаторни ограничения, като същевременно се запазват взаимовръзките между променливите. Този аспект е особено важен в чувствителни области като здравеопазването или финансите, в които поверителността на данните е от първостепенно значение. Освен това, синтетичните данни са икономически ефективни и спестяват време, тъй като елиминират необходимостта от прилагане на усилия за събиране и обработка на данни. Тази ефективност е особено полезна при трениране на модели за машинно обучение, провеждане на симулации и изпълнение на различни аналитични задачи. Синтетичните данни могат да бъдат генерирани с помощта на различни техники, като например симулации Монте Карло, генеративни състезателни мрежи (GAN) или други статистически методи основани на предварително установени разпределения, корелации и бизнес логика.

Регламентите за поверителност на данните често ограничават достъпността на анализаторите до информационни масиви за анализ. За да се преодолее това, могат да се създадат синтетични набори от данни, които запазват свойствата на оригиналните данни, но не съдържат никаква лична информация.

В ситуации, в които реалните данни са оскъдни или трудни за получаване, синтетичните данни могат да послужат като жизнеспособна алтернатива.

Събирането и съхраняването на информационни масиви от реалния свят представлява бавен, труден и скъп похват. Синтетичните данни често могат да бъдат генерирани на много по-ниска цена, което води до значителни икономии на разходи. Генерирането на синтетични данни позволява също така създаването на големи количества данни, способстващи за обучение на модели, прилагани за машинно обучение, без свързаните с това разходи по събиране и съхраняване на данни от реалния свят.

3.2.3.2 Предизвикателства

Генерирането на синтетични данни, от своя страна, представлява последователност от действия, които имат своите предизвикателства. Самият процес на синтезиране и адаптиране на новополучените данни в използваем

набор от информация представлява комплексна задача, изискваща сложни алгоритми и задълбочено разбиране на основните статистически взаимовръзки.

Съществен момент в процеса по генериране на данни представлява фазата на осигуряване на фактология и прецизност на синтетичните данни, тъй като моделите, обучени върху неверни, халюцинирани или пристрастни синтетични данни, създават предпоставки за неуспех в последователността за обобщаване в реални сценарии [5, 6, 7, 8].

В процеса на генериране на данни значително място заема проблемът със запазването на установените модели на зависимости и взаимовръзки, открити в данните от реалния свят. Друго предизвикателство е потенциалът на синтетичните данни да усилват отклонения или да въвеждат нови отклонения, в случай че не са внимателно проектирани и валидирани [9, 10].

Предвид това, резултатите, получени от синтетични данни, трябва да се използват с повишено внимание:

$$R_{real} \neq R_{synthetic}$$

Където R_{real} представляват резултатите от реалния свят, докато с $R_{synthetic}$ се обозначават тези, които са получени като следствие от генеративни процеси.

Осигуряването на качеството на новосъздадения набор от данни е друго съществено предизвикателство. От съществено значение е синтетичните данни да бъдат валидирани, за да се потвърди, тяхната представителност по отношение на явленията от реалния свят, с необходимата прецизност и достоверност. Този процес е от съществено значение, за да се гарантира надеждността и полезността на синтетичните данни и тяхното практическо приложения. Валидирането е стъпка, при която, посредством набор от реални данни, се гарантира надеждността на констатациите, осъществени чрез синтетични данни [4]. В литературата съществува широк набор от разнообразни подходи за валидация. Като такива могат да бъдат посочени тестът на Колмогоров-Смирнов, тестът за максимално сходство, тестът за обучение върху синтетични данни, изследването за тест върху реални данни (TSTR) и др. [11].

3.3 Класификация на синтетичните данни

Класификацията на синтетичните данни представлява многопластов процес, който се определя, в най-общ смисъл, от архитектурата на генеративните процеси и степента на запазване на оригиналната информация. Процесът по систематизиране според методологията на създаване и интензитета на синтез, създава предпоставки за оптимизиране на анализа в

среди с висока степен на неопределеност. Това стратегическо разделение е крайъгълен камък за развитието на надеждни модели на машинното обучение в ерата на големите данни.

3.3.1 Според насоката за използване

3.3.1.1 Частично синтезиране на данни (обогатяване на данни)

Частично синтезираните данни са информационни масиви, които запазват определени оригинални атрибути, докато в процеса по синтез се заменят само чувствителните променливи със синтетични стойности с цел анонимизация или надграждане на съществуващите масиви от информация. Разглежданият подход създава предпоставки за поддържане на висока аналитична полезност при специфични изследвания, в които връзките между реалните и синтетичните компоненти са критични [12]. Посредством осигуряване на комбинация между автентични записи и симулирани данни, организациите балансират между информативната стойност и защитата на личните данни, което обезпечава възможностите за изработване на информирано управленско решение. Същевременно се създават предпоставки за разширяване на съществуващите анализи посредством подобряване на наличната база от данни със синтетични масиви, които осигуряват възможност за създаване на по-детайлни разрези на анализираните клъстерни образувания.

Съществува набор от разнообразни методи за генериране на частично синтезирани данни. Възможни подходи са свързани с както с методологията за многократното входиране на синтетични записи, така и с тази за разширяване на данните чрез мащабиране на бизнес логиката. При подхода за многократно входиране се моделират и заменят само стойностите на предварително дефинирани чувствителни променливи [13], докато при методологията за мащабиране на бизнес логика се стъпва върху последователност за интегриране на взаимовръзки, извлечени от бизнес правилата. Споменатите концепция за частично синтезиране създава предпоставки за поддържане на висока степен на прецизност от страна на статистическите модели.

Съществен аспект в процеса по синтезиране на мащабни информационни масиви е свързан със способността за съхраняване на аналитичната полезност на специфичните подмножества от данни. Частичният синтез показва някои преимущества пред пълната симулация в разглежданата специфика [14]. Посредством частичните синтезирани масиви могат да се извличат валидни статистически изводи за генералната съвкупност, тъй като допълнителната вариация, въведена от процеса на синтез, се отчита в изчисленията на стандартните грешки. Успешното прилагане на този метод изисква прецизно

дефиниране на вероятностните модели, за да се предотвратят потенциални атаки чрез съпоставяне на реалните части от записи (част от общия масив) с външни публични бази данни.

3.3.1.2 Интерполация на данни

Процесът по интерполация на синтетичните масиви от данни представлява усъвършенстван подход за попълване на информационни празнини. За целта се използва математическо генериране на нови стойности в рамките на съществуващия диапазон. Този процес е критичен за осигуряване на структурна цялост при аналитични модели, които изискват непрекъснатост на числовите редове. Чрез прилагане на алгоритми за сближаване, системата създава предпоставки за успешно преодоляване на дефицитите в извадките, като същевременно не се нарушава общата логика на разпределението [15]. Правилното конфигуриране на тези изкуствени точки гарантира, че статистическите изводи остават надеждни и устойчиви на външни аномалии [16].

Интегрирането на обособените нови информационни сегменти позволява на алгоритмите за машинно обучение да функционират в среда с висока плътност на признаците. Специализираните техники за апроксимация минимизират риска от изкривяване на сигналите при обработка на мащабни масиви от информация [17]. Така създадената синтетична среда служи като здрава основа за тестване на хипотези и оптимизиране на прогнозните параметри. Разглежданата методология подобрява адаптивността на архитектурата спрямо специфичните динамики на изследваната област.

3.3.1.3 Изцяло синтезирани масиви от данни

От своя страна изцяло синтезираните данни представляват информационни масиви, които биват напълно генерирани посредством набор от статистически парадигми, интегрирани в алгоритми. Тези масиви не съдържат директни елементи от оригиналния източник, което подпомага процеса на анонимизация и елиминира риска от попадане под ударите на законодателството за защита на личните данни и Закона за кредитните институции. Това създава възможности за широко приложение на подходи за изцяло синтезиране на информационни масиви в управлението на финансови структури и бизнеси съдържащи чувствителна информация като анализи с медицинска насоченост. Тези набори се изграждат въз основа на оценени параметри на разпределението на генералната съвкупност [18]. Синтезираните масиви от информация осигуряват максимална защита на поверителността,

същевременно точността и прогнозиращата им мощ зависи изцяло от способността на модела да улови сложните многовариантни зависимости в оригиналната структура.

Посредством изцяло синтезираните масиви може да се постигне пълна анонимизация. При тези данни се конструира изцяло нов информационен обект въз основа на извлечени статистически параметри. Този методологичен подход създава предпоставки за практически невъзможно установяване на еднозначно съответствие между генерираните записи и индивидуалните субекти от оригиналния източник, което елиминира риска от деанонимизация [19]. Съществен аспект в този подход е използването на многомерни съвместни разпределения за подсигуряване на прецизна симулация, като същевременно новосинтезираните масиви от данни биват подложени на филтрация посредством бизнес логика. Друг подходящ подход е използването на корелационни матрици в среди с голяма размерност. Така синтезираните данни функционират като статистически еквивалент на реалната популация, позволявайки мащабни експерименти без правни ограничения.

3.3.2 Класификация според методите на генериране

3.3.2.1 Подходи базирани на вероятности

Подходите базирани на вероятности се основават на експлицитното моделиране на теоретичните разпределения на оригиналната извадка. За целта се използва набор от статистически функции за плътност на вероятността, които позволяват генерирането на нови записи със същите стохастични характеристики [12]. Тази методология се фокусира върху запазването на математическата коректност на модела спрямо генералната съвкупност от реални данни.

Основните фази в този тип подходи включват дефиниране на съвместни разпределения, както и изчисление на корелациите между всяка двойка променливи. Подходът се основава на парадигмата за Бейсови мрежи, както и на концепциите на различните функции на копула, които структурират йерархичните зависимости в големи информационни масиви. Коректното напасване на параметрите е от съществено значение за аналитичната валидност на синтетичния информационен набор.

Подходите, базирани на вероятности са широко разпространени в статистическите изследвания и медицинските проучвания. Тези подходи създават предпоставки за сигурно споделяне на чувствителна информация. Използват за създаване на публични версии на национални преброявания или данни свързани с управлението на здравни досиета. Тези масиви позволяват на

външни изследователи да тестват хипотези, без да компрометират личната поверителност.

Основното предимство е високата степен на интерпретируемост и статистическа строгост на генерираните данни. Като фундаментален недостатък се отчита трудното мащабиране при голяма размерност на данните [14]. Сложните нелинейни връзки в модерните системи, основани на големи информационни масиви, често остават извън обхвата на стандартните вероятностни модели.

3.3.2.2 Подходи за симулация, които не са базирани на вероятности

От своя страна подходите, които не се базират на вероятности, разчитат на алтернативни техники като алгоритми, базирани на правила, или модели за машинно обучение, които не разчитат изрично на вероятностните разпределения. При тези методи аналитичното описание на разпределенията е заменено от разучаване на структурата на данните посредством итеративни изчислителни процеси. В съвременната наука те често се идентифицират като генеративни AI модели [21].

Съществен момент в разглежданите подходи заема архитектурното проектиране на невронни мрежи и дефинирането на специфични за средата логически правила. При генеративните състезателни мрежи (GAN) се осъществява процес на съревнование между две невронни мрежи за постигане на максимален реализъм. Важен аспект е дефинирането на функциите на загуба, които направляват модела към точно структурно репликиране.

Методите, които не се базират на вероятности доминират при работа с неструктурирани масиви от големи данни, включително изображения, аудио и комплексни финансови трансакции. Те имат съществена роля в процеса по обучение на автономни системи и за изкуствено балансиране на редки събития. Подходът е много подходящ за случаи, при които традиционните статистически предположения не са приложими.

Най-голямото предимство е способността за моделиране на изключително сложни и високоразмерни зависимости. При използването на методи, които се базират на машинно обучение се наблюдава значителен риск от "пренапасване", при което моделът започва да възпроизвежда реални записи вместо нови синтетични копия. За да се предотврати това е необходимо да се вземат изрични мерки за добавяне стохастичен шум към градиентите, посредством който алгоритъма се принуждава да обобщава структурите, ограничавайки влиянието на всяка индивидуална трансакция или субект [22].

3.4 Заключение

В епохата на големите информационни масиви управлението, базирано модели, основани на масиви от синтетичните данни, се утвърждава като критичен инструмент за балансиране на аналитичната полезност и защитата на поверителността. Тяхното основно преимущество се състои във възможността за генериране на мащабни информационни масиви, които репликират прецизно статистическата структура, като същевременно запазват взаимовръзките между променливите на оригиналните данни. Описаният подход не само създава предпоставки за удовлетворяване на правилата на регулаторни рамки като GDPR и Закона за кредитните институции, в частта уреждаща банковата тайна, но и осигурява достъп до висококачествена информация за провеждане обучение на сложни управленски модели.

Предимствата на синтетичните масиви създават възможности за елиминиране на проблема със „студения старт“ и ефективно балансиране на редки събития, като по този начин катализира внедряването на иновации и намалява времето за пазарна реализация.

Генерирането на синтетични данни представлява комплексен и многоетапен процес, който изисква прилагането на усъвършенствани алгоритми и задълбочено разбиране на фундаменталните статистически закономерности. Критично предизвикателство е гарантирането на фактологична прецизност и обективност. Това е необходимо за да се предотврати обучението върху неверни, халюцинирани или пристрастни масиви, следствие на което да се възпрепятства ефективното обобщаване на моделите в реални. Успешното създаване и адаптиране на синтетичните данни зависи от способността за съхраняване на автентичните взаимовръзки между променливите от реалните данни. Същевременно се минимизира рискът от усилване на съществуващи или въвеждане на нови системни отклонения.

Използвана литература

[1] Marchev, V., Marchev, A., Jr. (2024). Anonymizing Personal Information Using Distribution-based Data Synthesis, XXII International Scientific Conference “Management and Engineering’24”, Sozopol, Bulgaria (in publishing)

[2] Nikhil. 2024. Google AI Introduces CodecLM: A Machine Learning Framework for Generating High-Quality Synthetic Data for LLM Alignment], [How to Use Synthetic and Simulated Data Effectively; https://towardsdatascience.com/how-to-use-synthetic-and-simulated-data-effectively-04d8582b6f88

[3] Liu, R., Wei, J., Liu, F., Si, C., Zhang, Y., Rao, J., Zheng, S., Peng, D., Yang, D., Zhou, D., and Dai., A., M., (2024). Best Practices and Lessons Learned on Synthetic Data for Language Models

[4] Marchev, A., Jr, Marchev, V., (2023). Automated Algorithm for Multi-variate Data Synthesis with Cholesky Decomposition, ICACS '23: Proceedings of the 7th International Conference on Algorithms, Computing and Systems, https://doi.org/10.1145/3631908.3631909

[5] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S., (2017) Gans trained by a two-time-scale update rule converge to a local Nash equilibrium. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 6626–6637, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

[6] Wood, E., Baltrusaitis, T., Hewitt, C., Dziadzio, S., Cashman, T. J., and Shotton, J., (2021). Fake it till you make it: face analysis in the wild using synthetic data alone. In 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, pages 3661–3671. IEEE, 2021. DOI: 10.1109/ICCV48922.2021.00366. URL https://doi.org/10.1109/ICCV48922.2021.00366.

[7] Guarnera, L., Giudice, O., and Battiato, O., (2020). Deepfake detection by analyzing convolutional traces. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pages 666–667.

[8] Van Breugel, B., Qian, Z. and Van Der Schaar, M., (2023). Synthetic data, real errors: how (not) to publish and use synthetic data. In International Conference on Machine Learning, pages 34793–34808. PMLR.

[9] Barbierato, E., Vedova, M. L. D., Tessera, D., Toti, D., and Vanoli, N., (2022). A methodology for controlling bias and fairness in synthetic data generation. Applied Sciences, 12(9):4619

[10] Gupta, A., Bhatt, D., and Pandey, A. (2021). Transitioning from real to synthetic data: Quantifying the bias in model. ArXiv preprint, abs/2105.04144, 2021. URL https://arxiv.org/abs/2105.04144.

[11] Skabar, A., 2024. Evaluating Synthetic Data — The Million Dollar Question, https://towardsdatascience.com/

[12] Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. Survey Methodology, 29(2), 181–188.

[13] Little, R. J. A.: Statistical Analysis of Masked Data. Journal of Official Statistics 9, 407–426 (1993)

[14] Drechsler, J. (2011). Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation.

[15] He, H., and Garcia, E. A. (2009) Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, 21, 1263-1284. http://dx.doi.org/10.1109/TKDE.2008.239

[16] Little, R. J. and Rubin, D. B. (2019) Statistical Analysis with Missing Data. Vol. 793, John Wiley & Sons, Hoboken. https://doi.org/10.1002/9781119482260

[17] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357

[18] Rubin, D.B. (1993) Discussion: Statistical Disclosure Limitation. Journal of Official Statistics, 9, 461-468.

[19] Raghunathan, T. E., Reiter, J. P., & Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics, 19*(1), 1-16.

[20] Patki, N., Wedge, R., & Veeramachaneni, K. (2016). *The Synthetic Data Vault.* In 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA) (pp. 399-410). IEEE. https://doi.org/10.1109/DSAA.2016.49

[21] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014) Generative Adversarial Nets. Proceedings of the 27th International Conference on Neural Information Processing Systems, 2, 2672-2680.

[22] Xie, L., Lin, K., Wang, S., Wang, F., & Zhou, J. (2018). Differentially Private Generative Adversarial Networks for Time Series, Continuous, and Discrete Data. arXiv:1802.07831.

4 Валидиране на синтетични данни

Автори: Васил Марчев, Ангел Марчев, мл.

Резюме: Настоящата глава разглежда процеса на валидиране на синтетични данни. Най-общо последователността по валидиране на информация се дефинира като последователност от статистически способи, които установяват съответствието на генерираните масиви спрямо реалните данни. Основните изследователски фази включват осигуряване на статистическа прецизност на разпределенията, максимизиране на аналитичната полезност на сценариите и поддържане на структурна консистентност. Методологичната класификация най-общо може да се раздели на количествени метрики (pMSE, TSTR, K-S тест, косинусово подобие) и качествени подходи(PCA, t-SNE, логически ограничения).

Централно място в анализа заема непараметричният тест на Колмогоров-Смирнов, използван за оценка на съответствието - едноизвадков или за сравнение между независими извадки - двуизвадков. Двуизвадковият K-S тест намира своето приложение при валидиране на алгоритми за машинно обучение поради своята чувствителност към формата и мащаба на разпределенията. Валидирането осигурява спазването на регулаторните и етични стандарти за анонимност, които са от решаващо значение за интегрирането на синтетични масиви в съвременните научни и корпоративни реалности.

Ключови думи: валидиране на синтетични данни, статистическа полезност, тест на Колмогоров-Смирнов

4.1 Същност и дефиниция на процеса

4.1.1 Дефиниция

Процесът по валидиране на синтетични данни представлява критичен компонент от жизнения цикъл на процеса по генериране на данни, насочен към гарантиране на статистическата автентичност и аналитичната стойност на масивите. В научната литература съществуват множество различни дефиниции. Част от най-съществените са обособени следващите абзаци.

Процесът по валидиране на данните може да бъде обособен като последователност от действия, имаща за цел да оцени статистическата полезност, разделена на обща (сравнение на разпределенията) и специфична (съответствие при конкретни статистически модели). Процесът изисква

синтетичните данни да запазват съвместното разпределение на променливите и техните взаимозависимости, така че резултатите от анализите да съвпадат с тези от оригиналния масив [1].

Други автори разглеждат процеса като „конгениалност“ (близост) между модела за въвеждане/генериране и анализа на данните. Последователността по валидиране гарантира, че синтетичните масиви създават предпоставки за извършване на валидни статистически изводи, посредством фокусиране върху съвпадението на средните стойности, дисперсиите и корелационните структури на оригиналните данни [2].

В научната литература процесът по валидиране се разглежда и като систематичен преглед, включващ проследяване на съответствието на новополучените данни с реалните разпределения. Последователността обхваща подфази на детайлен анализ и сравнение на честотите, като се цели избягване на резки колебания при прехода между съседни стойности, за да се потвърди, че първоначалните разпределения са запазени [3].

В научните среди се разглежда и подхода за валидиране от гледна точка на машинното обучение, представляващо количествено измерване на прецизността на синтетичните данни. Фокусът е върху способността на синтетичния модел да възпроизведе съвместните разпределения на променливите, без да нарушава поверителността на индивидуалните записи, като се извършва проверка на разпределенията [4].

Според други автори процеса се дефинира като балансиране между запазването на аналитичната структура на данните и минимизирането на риска от разкриване. Валидирането включва верификация на всяка синтетична променлива спрямо нейния оригинален еквивалент, за да се гарантира, че статистическите взаимовръзки са автентични и използваеми за публични изследвания [5].

Обобщавайки гореописаните формулировки в настоящата разработка се използва следната работна дефиниция: процесът по валидация на синтетичните данни представлява последователност, при която посредством последователност от статистически способи и математически похвати, интегрирани в алгоритми, се установява съответствието на генерираните масиви по отношение на разпределенията на реалните данни.

4.1.2 Същност

Процесът по валидиране на данни представлява фундаментален етап в цикъла на генериране на синтетични масиви. Посредством алгоритмична последователност се осигурява аналитичната стойност на новогенерираните

данни [2]. Последователността функционира като критичен механизъм за контрол на качеството, целящ потвърждаването на статистическата адекватност на изкуствено създадената информация. Евентуалната липса на стриктна верификация поставя под въпрос полезността на данните за научноизследователски цели. Процесът се утвърждава като задължителна фаза в последователността по запазване на доказателствената сила на моделите.

Основно предизвикателство в процеса на работа по валидация на данните се явява точното възпроизвеждане на съвместното разпределение на променливите, както и сложни взаимозависимости [1]. В процеса по генериране на данните се установява специфика в репликирането на корелациите между променливите, които са присъщи на оригиналния набор данни. Установява се възможност за осъществяване на неточно отразяване на корелациите и взаимодействията между променливите в синтезираните масиви. Описаните специфики предполагат анализ на резултатите на всяка отделна променлива и тяхната верификация спрямо първоначалното разпределение. Тази съпоставка гарантира структурната цялост и логическата консистентност на генерирания продукт.

Процесът обхваща редица подфази, сред които детайлен анализ, в чиито рамки се осъществява подробен преглед и сравнение на информацията от различни източници. Проследяването на съответствието между новополучените и реалните разпределения се разглежда като последователност от действия, която е от съществено значение за установяване на статистическата валидност. Съществен момент в процеса по валидация на данни представлява анализа на близките честоти, като се цели да се установи плавен преход без резки колебания в стойностите. Чрез разглеждания метод се създават предпоставки за избягване на аномалии, които биха могли да изкривят представителността на новосинтезираните масиви.

Финалният етап във верификацията на новите разпределения има за цел да установи до каква степен са запазени първоначалните параметри. Посредством описания систематичен подход се създават предпоставки за околичествяване на степента на сходство между синтетичните данни и реалните масиви. Посредством използването на подобни тестове се цели минимизиране на риска от систематични грешки и алгоритмично отклонение. Описаните действия имат решаващо значение за успешното прилагане на синтетичните данни в академичните и корпоративните среди.

4.1.3 Цели на процеса

Валидирането на синтетични масиви представлява не само техническа проверка, а многопластов анализ, целящ установяване на достоверността на синтетичните масиви от информация. За установяването на коректността на процеса е формулиран списък от фази.

4.1.3.1 Осигуряване на статистическа прецизност на разпределенията

Съществен аспект в процеса по валидиране на данните е установяване на прецизността им по отношение на реалните разпределения. Статистическата точност на синтетичните масиви предоставя възможност за коректно възпроизвеждане на едномерните и многомерни разпределения на характеристиките, съставляващи оригиналния информационен масив. Процесът на валидиране подробно проучва дали моментите на разпределението и корелационните структури запазват своята стабилност след процеса по синтезиране [1]. Чрез тази фаза се цели избягване на алгоритмични изкривявания, които биха компрометирали научната стойност на информацията. Високото ниво на прецизност създава предпоставки изследователите да използват синтетичните масиви като надежден източник за обучаване на модели.

4.1.3.2 Максимизиране на аналитичната полезност

Посредством процесите на валидация се цели максимизиране на аналитичната полезност, което е насочено към практическото приложение на данните в конкретни управленски сценарии. Чрез валидирането се създават предпоставки за установяване на това, дали резултатите от комплексните анализи на синтетични масиви съвпадат с тези, извлечени от емпиричните наблюдения [6]. Тази цел изисква прецизен баланс между сложността на синтетичния модел и неговата интерпретируемост за крайните потребители. Постигането на висока полезност спомага за генериране на масиви с подходяща структура и качество за интегрирането им в управленски модели в корпоративните и научни среди. По този начин валидираните данни могат да служат като стабилна основа за провеждане на експерименти без риск от погрешни заключения.

4.1.3.3 Осигуряване на структурна консистентност и логическа цялост

Следващата основна цел на процеса по валидация е осигуряването на структурна консистентност и логическа цялост. Тази фаза на удостоверяване на

данните създава предпоставки за установяване на нивото на запазване на логическите връзки и йерархичните зависимости в данните. Валидирането в този контекст се фокусира върху идентифицирането на логически аномалии или несъвместими стойности, които не съществуват в реалните набори от данни [3]. Предотвратяването на нелогични колебания в честотите е ключова фаза при поддържане на реалистичен профил на данните. Това подпомага изграждането на стабилни прогнозни модели, които са устойчиви на статистически шум. Консистентността се явява фундаментален стълб за дългосрочния интегритет на синтетичната информационна система.

4.2 Необходимост и предпоставки

Използването на синтетични масиви в съвременната научна практика изисква прецизна теоретична и методологична съпоставка на техните качества спрямо оригиналните набори от данни. Валидирането се явява като фундамент, трансформиращ генерираната информация в надежден аналитичен инструмент. разглежданият процес е от решаващо значение за преодоляване на скептицизма относно достоверността на синтетичните резултати [2]. Липсата на рамка установяваща процеса по валидация създава предпоставки за риск при използване на синтетични данни. Посредством стъпките в процеса на валидация се формира първичната предпоставка за интегриране на синтетичните масиви в конвенционалните научни и корпоративни работни процеси.

Генерираните информационни масиви представляват обект на анализ, установяващ деликатния баланс между статистическата значимост на резултатите и защитата на конфиденциалността. Предпоставките за успешно приложение включват доказване на прецизността на модела и липсата на рискове за разкриване на лична информация [5]. Систематичната оценка позволява на изследователите да идентифицират границите на приложимост на всеки синтетичен масив. Това осигурява необходимата прозрачност и методологична строгост, които представляват задължителен елемент в процеса по валидиране. Процесът по валидиране не представлява опционална стъпка, а структуроопределящ елемент в процеса по генериране на синтетични информационни масиви.

4.2.1 Контекстуална обусловеност

Използването на синтетични масиви без предварителна критична оценка крие сериозни рискове както за прецизността на анализите и последващите от

това управленски решения, така и за научната прецизност на изследванията. Използването на генеративни модели създава предпоставки за генериране на неточности или за пропускане на зависимости, характерни за първоначалния набор от данни [2]. Липсата на процес по верификация на новополучения масив създава предпоставки за трудно разграничаване между реални тенденции и погрешни заключения, генерирани като следствие от работата на модела. Наличието на подобна зависимост може да доведе до компрометиране на аналитичния процес. От казаното дотук може да се направи извод, че валидирането служи като задължителен филтър, осигуряващ съответствие между изкуствената среда и емпиричната реалност.

Статистическата значимост на синтезираната информация зависи от способността на алгоритмите да съхранят структурните връзки между променливите. Без задълбочена проверка на разпределенията, изследователите рискуват да публикуват резултати, които не могат да бъдат репликирани в реална среда [1]. Валидирането предоставя количествени метрики, които потвърждават достоверността на синтетичния аналог, разгледан през призмата на заместител на оригинала. Това обуславя невъзможността за директно внедряване на генерираните данни в критични области като медицината или управлението на финансови процеси. Оценката на прецизността на генерираните масиви се явява като функционална защитна рамка срещу информационен шум и алгоритмична неточност.

4.2.2 Регулаторни и етични норми

Регулаторните рамки изискват строги доказателства за анонимност при обработката на информация, дори когато тя е синтезирана алгоритмично. Процесът на валидиране е необходимо да демонстрира устойчивост срещу атаки за повторна идентификация, за да се гарантира пълната защита на личните данни [7]. Съответствието със съвременните стандарти представлява не само правно задължение, но и етична норма в модерната наука за данни. Липсата на ясна методология за оценка на поверителността, поставя организациите на значителни правни и репутационни рискове. Ефективната валидация създава необходимата прозрачност и сигурност за използване на данните в публичната сфера.

Етичният аспект на синтетичните данни се фокусира върху отговорното управление на рисковете от разкриване на чувствителна информация. Валидирането на анонимността посредством стандарти като диференциалната поверителност гарантира, че индивидуалните записи не могат да бъдат изолирани [8]. Последователността от подобни действия създава доверие

между субектите на данни и анализаторите, улеснявайки споделянето на ценни ресурси. Поддържането на високи етични стандарти изисква непрекъснат мониторинг на съответствието с актуалните стандарти за безопасност. Процесът на валидиране може да бъде разгледан и като мост между технологичните иновации и защитата на персонална информация.

4.3 Класификация на методите за валидация

Класификацията на методите за валидация най-общо може да бъде диференцирана на две основни групи: количествени подходи, базирани на математически метрики, и качествени/функционални подходи, фокусирани върху визуалната интуиция и логическата структура.

4.3.1 Количествени подходи

Количествените методи осигуряват обективна и възпроизводима оценка на сходството между оригиналните и синтетичните данни. Тези подходи са фундаментални за автоматизираните системи за оценка на алгоритми за синтез. Разглежданите способи за валидация формират обективната основа за оценка на статистическата значимост. Тяхната основна цел е да определят, посредством количествени и статистически методи, загубата на информация и статистическото сходство между реалния и синтетичния набор. Те се разделят на метрики за локално сходство (на ниво променлива) и глобални метрики за структурно съответствие. Някой от най-често използваните количествени методи са разгледани в следващите абзаци.

Propensity Score Mean Squared Error (pMSE). Разглежданият метод се използва за измерване на общата полезност на данните. Той обхваща дискриминативен модел, който се опитва да различи реалните от синтетичните записи. Ниските стойности на pMSE индикират, че моделът не може да направи разлика между двата набора, което е доказателство за високо качество на синтеза [1].

Ефикасност на машинното обучение. Този метод, известен още като "Train on Synthetic, Test on Real" (TSTR), оценява до каква степен моделите, обучени върху синтетични данни, запазват прогнозиращата си мощ по отношение на реални тестови набори. Високото съответствие в точните метрики (като F1-score или RMSE) показва, че синтетичните данни са запазили сложните зависимости между променливите [4].

Тест на Колмогоров-Смирнов (KS test). По своята същност това представлява непараметричен статистически тест, подходящ за оценка на съответствието между две разпределения. За целта се осъществява измерване

на максималното разстояние между техните кумулативни функции на разпределение. В контекста на валидацията на синтетични данни, KS тестът се прилага за всяка една от променливите поотделно, като по този начин се удостоверява дали синтезираният набор успешно репликира статистическата структура на реалните данни. Ниската стойност на тест-статистиката и високото p-value са индикатор, че двете извадки произхождат от едно и също разпределение, което потвърждава високата статистическа значимост на синтеза [9].

Косинусово подобие. Косинусовото подобие измерва косинуса на ъгъла между два вектора в многомерното пространство. Посредством тази метрика се оценява тяхната ориентация независимо от разглежданите мащаби. При валидиране на данни този метод се използва основно за сравняване на структурните зависимости. Подходящ е за съпоставяне на корелационни матрици или вграждане на вектори на реалния и генерирания набор. Резултатите, близки до единица, показват запазване на многовариантността на връзките и насочеността на данните, което е от съществено значение за моделите на машинното обучение [10].

4.3.2 Качествени и функционални подходи

Качествените и функционалните подходи поставят в своя фокус процеса по анализ на семантичната цялост и структурната логика на данните, които често остават скрити за чисто статистическите метрики. Докато количествените методи осигуряват числови доказателства, качествените и функционалните подходи се фокусират върху смисловата взаимосвързаност и практическата приложимост на данните по отношение на управленските системи. Разглежданата група от подходи обезпечава, че генерираната информация не само изглежда статистически коректна, но и функционира логично в реални сценарии. Някой от най-често използваните подходи са разгледани в следващите абзаци.

Визуализация чрез намаляване на размерността (t-SNE и PCA). Разглежданият подход създава предпоставки за визуално инспектиране на припокриването на клъстерите от реални и синтетични точки в двуизмерно пространство. В случай че синтетичните данни формират идентични структури с оригиналните, това е значим индикатор за запазване на глобалната структура на данните [11].

Валидация чрез логически ограничения: Функционалната валидация удостоверява дали синтезираните записи спазват първоначално дефинирания набор от бизнес правила и логически връзки. Разглежданият метод е от

съществено значение за оперативната съвместимост на данните в реални информационни системи [12].

Анализ на екстремните стойности и редки събития. Качественият преглед се фокусира върху способността на генеративния модел да произвежда правдоподобни редки случаи, без да се нарушава принципа на анонимизацията. Това изисква експертна оценка в съответната бизнес област, за да се потвърди, че синтетичните "аномалии" имат смисъл в контекста на реалния свят [8].

4.4 Тест на Колмогоров-Смирнов

В рамките на настоящата работа съществен момент в процеса по валидиране на данни заема тестът на Колмогоров-Смирнов (KS). Разглежданият метод е непараметричен тест, който се използва за две основни цели: първо, той оценява съвместимостта между извадково разпределение и теоретично разпределение в едноизвадковия KS тест. Втората цел, за която се използва, е да се изследва сходството между две извадкови разпределения в двуизвадковия KS тест. KS тестът разчита на емпиричната кумулативна функция на разпределение (ECDF), за да определи дали наборът от данни съответства на предварително дефинирано разпределение.

4.4.1 Едноизвадков тест на Колмогоров-Смирнов

Едноизвадковият тест на Колмогоров-Смирнов представлява основен непараметричен метод за оценка на съответствието между емпиричното разпределение на дадена извадка и предварително дефинирано теоретично непрекъснато разпределение [18]. Основната сила на разглеждания подход се крие в изчисляването на максималната абсолютна разлика между кумулативната функция на разпределение на извадката и тази на целевата популация [9]. Методът намира своето приложение по отношение на строга проверка за нормалността, осигурявайки статистическа акуратност при анализ на непрекъснати променливи [13]. Неговата инвариантност спрямо монотонни трансформации на скалата го прави предпочитан избор за прецизно моделиране в сферата на науката за данни [14]. Чрез използването на този тест изследователите могат обективно да потвърдят дали анализираните данни следват очаквания теоретичен модел, без да се опират на рестриктивни параметрични предположения [15]. Разглежданият аналитичен инструмент представлява крайъгълен камък в процесите за валидиране на едномерни структури в сложни масиви от информация.

Едноизвадковият тест на Колмогоров-Смирнов (KS) е алгоритъм, използван за сравняване на емпиричната кумулативна функция на разпределение (ECDF) на извадка и кумулативната функция (CDF) на теоретичното разпределение. ECDF представлява делът на наблюденията в извадката, които са по-малки или равни на определена стойност, докато CDF на теоретично разпределение предоставя теоретичната вероятност случайно наблюдение, извлечено от разпределението, да бъде по-малко или равно на определена стойност. Използва се едноизвадковия KS, в случай че е получено базисното разпределение чрез предположение или чрез други теоретични априорни данни.

Тестовата статистика D за едноизвадковия KS тест представлява максималната абсолютна разлика между ECDF на извадката и CDF на теоретичното разпределение. Математически еднопробната KS тестова статистика може да бъде изразена като:

$$D = \max_i \left| F_0(x_i) - F_n(x_i) \right|$$

Където:
$F_0(x_i)$ = теоретична кумулативна функция на разпределение (CDF)
$F_n(x_i)$ = емпиричната кумулативна функция на разпределение (ECDF).

4.4.2 Двуизвадков тест на Колмогоров-Смирнов

Двуизвадковият тест на Колмогоров-Смирнов е инструмент за установяване на това дали две независими извадки произхождат от едно и също непрекъснато разпределение. Разглежданият непараметричен метод се основава на изчисляването на супремума на множеството от разлики между емпиричните кумулативни функции на двете групи [16, 19]. В контекста на валидирането на синтетични данни, този подход служи като строг показател за това дали генерираният набор успешно репликира статистическата структура на реалния източник [9]. Методът се отличава със своята чувствителност не само към централната тенденция, но и към вариациите във формата, мащаба и асиметрията на разпределенията [17]. Поради своята изчислителна ефективност, двуизвадковият K-S тест представлява предпочитан избор при оценка на полезността на синтетични масиви от данни, използвани в алгоритмите за машинно обучение. Чрез него се постига необходимата аналитична прецизност за доказване на еквивалентност между синтетични и реални информационни масиви.

Двуизвадковият KS тест сравнява ECDF на две независими извадки, за да определи дали те са извлечени от една и съща популация (т.е. дали извадките имат едно и също разпределение). Тестовата статистика D, за двуизвадковия KS тест представлява максималната абсолютна разлика между двете ECDF.

Статистиката на двуизвадковия KS тест може да се изрази като:

$$D = \max_i \left| F_{1,n}(x_i) - F_{2,m}(x_i) \right|$$

Където:

$F_{1,n}(x_i)$ - емпирична кумулативна функция на разпределение (ECDF) на първата извадка

$F_{2,m}(x_i)$ - емпирична кумулативна функция на разпределение (ECDF) на втората извадка.

4.4.3 Прилики и разлики между едноизвадков и двуизвадков тест на Колмогоров-Смирнов

Както при едноизвадкови, така и при двуизвадковия KS тест, нулевата хипотеза (H0) гласи, че извадката(ите) следва(т) зададеното разпределение (еднопробен) или че двете извадки са резултат от едно и също разпределение (двупробен). Алтернативната хипотеза (H1) е, че извадката(ите) не следва(т) зададеното разпределение (еднопробен), или че двете извадки произлизат от различни разпределения (двупробен). Ако изчислената D статистика е по-малка от критичната стойност, нулевата хипотеза не се отхвърля.

Критичната стойност зависи от нивото на значимост (обикновено обозначено като) D_α и размера на извадката (за еднопробния KS тест) или размерите на извадките (за двупробния KS тест).

За еднопробния KS тест, критичната стойност може да бъде приблизително определена с помощта на формулата:

$$D_\alpha = c(\alpha) / \sqrt{n}$$

Където с $c(\alpha)$ е означена обратната стойност на KS разпределението при $1-\alpha$, а n е размерът на извадката.

За двупробния KS тест, критичната стойност се изчислява по формулата:

$$D_\alpha = c(\alpha) * \sqrt{(n_1 + n_2) / (n_1 * n_2)}$$

4.5 Заключение

Процесът по валидиране на синтетични данни представлява критичен компонент от жизнения цикъл на генерирането на масиви, целящ гарантиране на статистическата значимост на данните и аналитичната им стойност. В сферата на науката за големи данни това представлява фундаментален етап, който осигурява необходимата статистическа адекватност на изкуствено създадената информация. Липсата на фаза, в която се осъществява стриктна верификация би компрометирала полезността на данните, използвани както за обучаване на управленски модели, така и за научноизследователски цели, и би поставила под въпрос доказателствената сила на аналитичните модели. Валидирането служи като задължителен филтър за съответствие между алгоритмичната среда и емпиричната реалност, предоставяйки количествени метрики за потвърждаване на достоверността.

От изследователска перспектива основното предизвикателство е точното възпроизвеждане на съвместните разпределения и сложните взаимозависимости между променливите. Процесът изисква прецизен баланс между статистическата полезност и защитата на конфиденциалността. Оценката на прецизността служи като функционална защитна рамка срещу информационен шум и алгоритмична неточност. Това обуславя невъзможността за директно внедряване на генерирани данни в критични области. За целта е необходимо да се осъществи предварителна оценка на сходството на разпределенията на синтетичните данни с тези на реалните информационни масиви.

Прилагането на системен подход позволява околичествяване на степента на сходство между синтетичните и реалните масиви. Методи като теста на Колмогоров-Смирнов осигуряват статистическа акуратност при анализа на непрекъснати променливи, като същевременно са устойчиви спрямо монотонни трансформации. Процесът по валидиране представлява структуроопределящ елемент, който изгражда доверие между анализаторите и субектите на данни, гарантирайки дългосрочния интегритет на управленските системи.

Използвана литература

[1] Snoke, J., Raab, G. M., Nowok, B., Dibben, C., & Slavkovic, A. (2018). General and specific utility measures for synthetic data. Journal of the Royal Statistical Society: Series A (Statistics in Society), 181(3), 663-688.

[2] Reiter, J. P. (2003). Inference for partially synthetic public use microdata sets. *Survey Methodology*, 29(2), 181-188.

[3] Raab, G. M., Nowok, B., & Dibben, C. (2016). synthpop: Bespoke Creation of Synthetic Data in R. *Journal of Statistical Software*, 74(11), 1-26.
[4] Jordon, J., Yoon, J., & van der Schaar, M. (2018). PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees. *International Conference on Learning Representations (ICLR)*.
[5] Drechsler, J. (2011). Synthetic Datasets for Statistical Disclosure Control. Springer Series in Statistics
[6] Reiter, J. P. (2005). Significance tests for multi-component estimands from multiply imputed, synthetic microdata. Journal of Statistical Planning and Inference, Volume 131, Issue 2, 1 May 2005, Pages 365-377
[7] Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science, 9(3–4), 211–407. https://doi.org/10.1561/0400000042
[8] El Emam, K., Mosquera, L., & Hoptroff, R. (2020). Practical Synthetic Data Generation: Balancing Privacy and Utility in Data Sharing. O'Reilly Media.
[9] Massey, F. J. (1951). The Kolmogorov-Smirnov Test for Goodness of Fit. Journal of the American Statistical Association, 46(253), 68-78
[10] Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. Morgan Kaufmann.
[11] Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9(11), 2579-2605.
[12] Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The synthetic data vault. IEEE International Conference on Data Science and Advanced Analytics (DSAA), 399-410.
[13] Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. Journal of the American Statistical Association, 62(318), 399–402.
[14] Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. Journal of the American Statistical Association, 69(347), 730–737
[15] Conover, W. J. (1999). Practical nonparametric statistics (3rd ed.). John Wiley & Sons.
[16] Smirnov, N. V. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. Bulletin Mathématique de l'Université de Moscou, 2(2), 3–14.
[17] Darling, D. A. (1957). The Kolmogorov-Smirnov, Anderson-Darling tests of goodness of fit. The Annals of Mathematical Statistics, 28(4), 823–838
[18] Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione [On the empirical determination of a distribution law]. *Giornale dell'Istituto Italiano degli Attuari*, *4*, 83–91
[19] Smirnov, N. (1948) Table for Estimating the Goodness of Fit of Empirical Distributions. *The Annals of Mathematical Statistics*, 19, 279-281. https://doi.org/10.1214/aoms/1177730256
[20]

5 Бизнес логика

Автори: Васил Марчев, Ангел Марчев, мл.

Резюме. Настоящата глава разглежда в детайли същността на бизнес логиката като набор от правила и ограничения, съпътстващи системите за генериране на данни. Разглежданата логическа рамка изпълнява ключови функции като осигуряване на семантична цялост, дефиниране на специфични бизнес зависимости и подпомагане на процеса по извличане на нови знания. Текстът класифицира видовете логика в четири основни категории: трансформационна, интегрираща, изчислителна и филтрираща бизнес логика. Всяка от разглежданите класификации открива своята роля в ефективния синтез на данни. Разгледани са критичните характеристики на логическия слой, включително: унификация спрямо източника, технологична проследимост и мащабируемост чрез паралелна обработка. Методологическият цикъл на изграждане обхваща фазите на установяване на правила, формализация, валидация, приложение и непрекъсната обратна връзка. Внедряването на бизнес логика в процеса по генериране на данни може да се разглежда както през призмата на персонализацията на потребителското преживяване, така и през последователността по гарантиране на регулаторното съответствие чрез автоматизирани проверки. Качеството на системата се определя от нейната точност, оптимизация на времето за реакция и способност за динамична адаптация към промените в средата.

Ключови думи: бизнес логика, синтез на данни, мащабируемост, бизнес правила и ограничения

5.1 Същност на бизнес логиката в контекста на големите данни

5.1.1 Дефиниции

В литературата се срещат множество автори, които обособяват широк набор от дефиниции. Предложените формулировки определят същността, съдържанието и границите на понятието „бизнес логика“. Някои автори разглеждат бизнес логиката през призмата на домейн логика. Това е онази част от приложната система, която е отговорна за кодирането на реалните бизнес правила, определящи как данните могат да бъдат създавани, съхранявани и променяни [1].

Според други автори бизнес логиката може да бъде дефинирана като формално представяне на бизнес процесите и политиките на организацията, които са интегрирани в софтуерни компоненти, с цел да се осигурят изпълнението на конкретни бизнес задачи и цели [2].

Последователи намира и дефиницията на понятието свързана с пречупването на бизнес логиката през призмата на оперативните правила и зависимости, които управляват поведението на информационната система в отговор на външни или вътрешни бизнес събития [3].

От системна гледна точка разглежданото понятие включва алгоритмичните ограничения и функционалните изисквания, които осигуряват семантичната съгласуваност на данните при тяхното синтезиране от набор от хетерогенни източници [4].

В литературата съществуват и други смислови насоки по отношение на бизнес логиката. Някои автори разглеждат понятието през призмата на фундаментална система от формални логически изображения, които осигуряват семантично съответствие между хетерогенни структури [5]. Според други, разглежданото понятие може да бъде дефинирано и като критичен елемент за стратегическо съгласуване, свързващ организационните цели с архитектурата на информационните системи [6]. Други смислови насоки разглеждат бизнес логиката като набор от декларативни правила, които управляват интелигентния синтез и търсене в уеб базирани масиви от информация [7].

Като се вземат под внимание разгледаните по-горе дефиниции на понятието и се вземат под внимание спецификите на настоящия труд, твърдението се използва със следната работна дефиниция: бизнес логиката представлява набор от правила, интегриращи потенциално взаимозависимите фактори, техните ограничения, възможни комбинации и невъзможни такива.

5.1.2 Функция

Бизнес логиката може да бъде разгледана като основен архитектурен слой, който превръща изолираните информационни масиви в съгласувани стратегически ресурси. Процесът на синтезиране на данни не се ограничава само до техническото обединяване на таблици, той изисква обособена рамка, която да интерпретира взаимовръзките между променливите с оглед на конкретните управленски цели. В тази връзка, бизнес логиката може да бъде разглеждана като семантичен филтър, който гарантира, че синтезираният резултат е не само математически точен, но и контекстуално валиден.

Една от основните функции на бизнес логиката е осигуряването на семантична цялост. В хетерогенните набори едно и също понятие може да има различни измерения. При наличие на разнородни значения е необходима намесата на теоретичната рамка на интеграцията, при която глобалната схема трябва да бъде изградена чрез формални логически съответствия, които преодоляват структурните различия между източниците [5]. Липсата на изградена рамка от правила и нейното приложение може да доведе до генериране на информационен шум, лишен от управленска и бизнес стойност.

Друга ключова функция на бизнес правилата е дефинирането на специфични ограничения. Бизнес логиката обособява експертните познания за разглежданата бизнес среда. В процесите по синтезиране на информационни масиви е необходимо изградените алгоритмите да следват реални зависимости, които да подсигурят надеждността на изводите базирани на новогенерираните набори. Подобни зависимости могат да бъдат свързани с логиката на кредитния риск или управление на веригата на доставките и др. В общия случай генерираните правила представляват набор от обусловености, които, на пръв поглед, може да останат скрити от анализаторите.

Бизнес логиката може да бъде разглеждана като движещата сила в процеса по извличане на нови знания чрез деривация. Това е последователност, при която се създават нови променливи или показатели, които не съществуват в суров вид. Казаното до тук може да бъде разгледано през призмата на анализа на процесните логове и прилагане на бизнес правила, чрез които се постига дълбоко разбиране на оперативната реалност [8]. Генерирането на синтетични масиви, базирани на бизнес логика, създава предпоставки за автоматизирано изработване на информационни масиви, отговарящи на предварително синтезиран набор от правила, които съответстват на реалните условия на средата.

Бизнес логиката може да бъде разглеждана като инструмент, създаващ предпоставки за устойчивост и проследимост на процесите по генериране на данни. Ясно дефинираните правила за обработка и тяхната имплементация в техническата инфраструктура, благоприятстват за бързото адаптиране на управленските модели към променящата се пазарна среда. Правилното структуриране на домейн логиката позволява успешното управление на сложни бизнес сценарии, като минимизира риска от грешки при интеграцията на мащабни масиви от данни [1].

5.2 Стратегически приложения

Синтезирането на данни не е просто техническа операция по комбиниране на хетерогенни източници, това е стратегически процес, който изисква дълбока интеграция на бизнес логика. В контекста на архитектурата за генериране на информационни масиви, бизнес логиката служи като семантичен мост между суровите структури от данни и крайните стратегически цели. Ефективният синтез трансформира отделните информационни набори в съгласувани масиви, подходящи за анализ и прогнозиране на потребителското поведение.

5.2.1 Оптимизация на изработването на решения

Интегрирането на бизнес логика в алгоритмите за синтез на данни създава предпоставки за филтриране на шума и фокус върху метрики, които имат реална икономическа и прогнозна стойност. Прилагането на бизнес логика трансформира моделите от строго статистически, фокусирани върху корелациите между променливите, в подходи, интегриращи както статистически парадигми, така и причинно-следствени ограничения посредством предварително установените правила. Според някои автори, разглеждащи интегрирането на бизнес логика в моделите, бизнес аналитиката е необходимо да бъде дълбоко вкоренена в организационния и управленски контекст, за да превърне „големите данни“ в „голямо въздействие“ [9]. Генерирането на масиви с добавена логика спомага за автоматизирането на процесите, като по този начин се намалява когнитивното натоварване, като се елиминират субективните пристрастия. Този подход осигурява методологическа рамка за превръщането на аналитичните резултати в изпълними стратегии [10].

5.2.2 Осигуряване на съответствие

В ера на строги регулации, бизнес логиката може да приеме ролята на регулаторен филтър в процесите по генериране на данни. Алгоритъмът на синтезиране е необходимо да включва серия от автоматизирани проверки за съответствие, които да гарантират, че крайните набори от данни не нарушават поверителността или етичните норми. Синтетичните данни, генерирани посредством специфична бизнес логика, позволяват тестване на системи без излагане на чувствителна информация. Някои автори считат, че етичната рамка представлява неразделна част от управлението на данни, като бизнес логиката трансформира абстрактните принципи в набор от технически спецификации

[11]. Посредством интегриране на правила за осигуряване на съответствие директно в алгоритмите за генериране на данни се минимизира рискът от възникване на правни санкции или несъответствие на данните спрямо реалните информационни масиви.

В процеса на синтез на данни бизнес логиката създава предпоставки за технологична проследимост, която е критична за одиторските процеси. Посредством дефиниране на експлицитни правила за трансформация, изследователите могат да докажат произхода и легитимността на всяка точка от данни в крайния синтезиран продукт. Така обособеният механизъм на „управление чрез дизайн“ минимизира вероятността от неволно включване на дискриминационни променливи или непредвидени изкривявания в аналитичните модели. Според някой автори липсата на прозрачност в алгоритмичния синтез може да генерира системни рискове, докато интегрираната бизнес логика функционира като регулаторна рамка, гарантираща алгоритмична отчетност и прозрачност в съответствие със съвременните стандарти за управление на информацията [12].

5.2.3 Персонализация на крайния продукт

Приложението на бизнес логиката в синтеза на данни улеснява процесите по персонализация и управление на потребителското търсене. Посредством комбинация на поведенчески данни с профилни характеристики и пазарни тенденции, алгоритмите могат да бъдат програмирани да създават уникални профили, които да се превърнат в съществен обучителен инструмент на препоръчителните системи. Персонализацията изисква динамичен синтез, в който бизнес логиката е незаменим помощник в последователността по адаптиране на моделите към съответстващия контекст на потребителя, като по този начин се осигурява релевантност и ангажираност. Стойността на този процес се създава съвместно, посредством внедряване на институционални логики, които в дигиталния свят се реализират с помощта на интелигентен синтез на информация [13].

Персонализацията изисква синтез на данни, който надхвърля статичните потребителски сегменти и се адаптира динамично към индивидуалните микро-контексти. Бизнес логиката в този процес дефинира критичните параметри на релевантност, като осигурява отразяване на актуалните нужди и предпочитания на конкретния субект. Този подход създава възможност за преминаване от реактивно обслужване към проактивно предлагане на стойност, където моделите предвиждат бъдещи потребности въз основа на сложни исторически образци и поведенчески сигнали. Посредством интегриране на

организационните и управленски цели в логиката на алгоритмите за синтез, персонализацията се трансформира от техническа функция в устойчиво стратегическо предимство.

5.3 Ключови специфики на логическия слой

В настоящите технологични реалности, синтезирането на данни представлява процес, който не се ограничава до простото генериране на информационни масиви и тяхното интегриране в модели. Това представлява сложна последователност, включваща интегриране на семантични правила по отношение на разнородни информационни слоеве. Бизнес логиката служи като мост между суровите технологични ресурси и процесът по изработване на стратегическо решения. В литературата се е наложило мнение, че предизвикателствата пред анализа на големи данни често произтичат не от липсата на алгоритми, а от трудността при трансформирането на неструктурираната маса в смислен бизнес контекст [14].

5.3.1 Унификация спрямо източника

Един от основните критични моменти в синтезирането на данни е способността на бизнес правилата да функционират независимо от формата на входящия поток. Традиционно структурираните релационни бази данни и неструктурираните потоци (като JSON обекти от сензори или текстови логове) се обработват в изолирани информационни хранилища. Бизнес логиката изисква абстракция, която унифицира тези източници. Предпоставките за интеграция на данните в мащаб се нуждае от системи, които могат да разпознават единици и връзки независимо от тяхното първоначално представяне [15]. Като се вземат предвид гореизложените постулати, се установява, че ако дадено бизнес правилото дефинира "активен клиент", критериите за това трябва да се прилагат еднакво към всички съществуващи бази данни, независимо от техния формат или ниво на структурираност. Тази унификация може да бъде постигната посредством изграждане на семантичен слой, който отделя дефиницията на логиката от физическото хранилище.

5.3.2 Проследимост

В контекста на множеството регулаторни рамки и нуждата от висока степен на доверие в новополучените резултати, проследимостта на данните се превръща в императив. Според някои автори тази фаза има ключова роля в процеса по отстраняване на грешки и осигуряване на качеството в сложни

аналитични вериги [16]. Всяко едно от установените бизнес правила представлява хипотеза, която променя състоянието на данните. В случай че крайният резултат представлява невъзможна комбинация, алгоритъмът може да се върне назад и да осъществи одит на системата. По този начин се идентифицира дали проблемната ситуация е в суровия източник или в конкретната логическа стъпка. В системите за синтез това се реализира посредством метаданни, които придружават всеки агрегиран обект, описвайки конкретните стъпка от процеса на генериране.

Концепцията за проследимост на процесите в бизнес логиката включва и управлението на версиите на поредицата от интегрирани правила в дискретен времеви период. Според някои автори ефективната система за проследяване на версиите е необходимо да съхранява не само трансформираните данни, но и точното състояние на алгоритмите към момента на тяхното изпълнение [17]. Това позволява на изследователите да възпроизвеждат идентични резултати от минал период, дори след като логическите зависимости са били актуализирани. Разглежданата форма на проследимост осигурява възможности за математическа валидация на синтезираните масиви, предотвратявайки появата на непрозрачни алгоритмични процеси.

5.3.3 Мащабируемост

Съществен момент при синтеза на големи масиви от информация се свързва с предпоставките за мащабируемост на процесите. Предизвикателство в разглежданата система представлява поддържането на сложни зависимости (например рекурсивни изчисления или многослойни присъединявания) при минимална латентност. Според някои автори [18] унифицираните двигатели за обработка представляват единственият начин за справяне със скоростта и обема на съвременните данни. Мащабируемостта изисква бизнес логиката да бъде декомпозирана на паралелни задачи, които да се изпълняват върху разпределени клъстери. Ключов момент представлява "изтласкването" на логиката възможно най-близо до източника, за да се минимизира трансферът на ненужна информация през мрежата.

5.4 Таксономия на видовете бизнес логика

Бизнес логиката в процеса на генериране на данни действа като семантичен слой, който трансформира суровите измервания в използваеми активи. Съществен момент в процеса по генериране на данните представлява структурирането на метаданните и правилата, така че те да отразяват реалните

бизнес цели [19]. От гледна точка на таксономията и възможностите за класификация на видовете бизнес логика може да се определят няколко основни насоки, включващи: трансформационна бизнес логика, логика на интегритета, деривационна бизнес логика и филтрираща.

5.4.1 Трансформационна бизнес логика

В класификацията на бизнес логиката, трансформационната логика разглежда процесите през призмата на последователността от действия, чрез които данните променят своята форма, структура и мерни единици, за да съответстват на унифициран стандарт. Посочените действия включват както технически конверсии, като промяна на типове данни, така и бизнес операции като валутна конверсия и нормализиране на времеви интервали. За да се постигне ефективност на процеси като извличане, трансформиране, зареждане е необходимо да се разчита на строго дефинирани трансформационни правила, които гарантират, че синтезираният резултат е консистентен в рамките на разглеждания масив [20].

В случай че трансформационната логическа рамка бъде разгледана в контекста на времевите редове се установява, че тя играе ключова роля в процеса по агрегиране на синтетични информационни масиви. Като съществени моменти подкрепящи тази теза могат да бъдат разгледани данните получени от сензори, генерирани през милисекунди, които е необходимо да бъдат трансформирани в минутни или часови обобщения, посредством прилагане на набор от специфични статистически функции (средна стойност, медиана и др.). Този процес изисква последователност от действия, удостоверяващи, че е проектиран да функционира по аналогичен начин спрямо съответните източници, като това предполага, че логиката за конверсия работи идентично върху набор от различни входящи потоци. Управлението на системи от такъв тип и техните трансформации представлява критична фаза, чрез която се цели предотвратяване на информационни загуби по време на интеграцията [21].

5.4.2 Интегрираща бизнес логика

Основен момент в класификацията на видовете бизнес логика заема интегриращата. Този тип логическа структура може да бъде сравнена със защитния механизъм в системите за генериране на информационни масиви. Логическата рамка се състои от набор от декларативни правила, които посредством интеграция на потенциално взаимозависимите фактори

установяват техните ограничения, възможни комбинации и невъзможни такива. Най-съществения момент при интегриращата бизнес логика е стриктното изпълнение на хронологичната последователност, съдържаща се в променливите. Пример, който може да бъде изведен, с цел илюстрация на подхода, представлява притежаването на банкова сметка от физическо лице, което няма навършени определен брой години. Евентуалната липса на слой, съдържащ интегрираща бизнес логика в рамките на моделите за генериране на информация, създава предпоставки за загуба на надеждността на синтезираните масиви, което, от своя страна, би довело до неверни аналитични изводи и управленски решения. От казаното до тук може да се направи извод, че почистването на данните и проверките на интегритета представляват неделима част от процеса на синтез, тъй като грешките в началните етапи се мултиплицират експоненциално [22].

Освен гореспоменатите хронологични проверки, тази логика включва и широк набор от верификации на релационни зависимости. Към този тип проверки спадат първоначално зададени зависимости между нива на доход и генерирани разходи, степен и сфера на образование, така и доходи на лицето. Друг пример в тази посока може да бъде свързан с банковия сектор и предотвратяване на синтезирането на трансакции, които превишават разполагаемия кредитен лимит, или приписването на активи на несъществуващи сметки. Интегриращите бизнес правила удостоверяват и верифицира, че всеки запис в общия информационен масив отговаря на предварително зададените бизнес аксиоми.

Съществен аспект в работата на бизнес логика от този тип е представлява последователността по приложение на правилата. Тази специфика в работата на системата, прилагаща бизнес логиката, спомага за подобряване на проследимостта. Разглежданата последователност от действия създава възможност за установяване на причината за отхвърляне или коригиране на определен запис по време на процеса на валидация.

За целите на настоящата разработка се използва интегрираща бизнес логика, като установения пълен набор от логически принципи наброява повече от сто взаимозависими правила. Те са базирани на априорни познания на авторите за икономическата и управленска среда, наблюдения и анализи на социалните и банкови системи.

Установяването на този набор от правила осигурява структурна стабилност и предвидимост при управлението на сложни финансови операции в динамична пазарна обстановка. Експертните познания позволяват прецизното моделиране на взаимовръзките между макроикономическите индикатори и специфичните изисквания на регулаторната рамка. Чрез задълбочен анализ на

социалните потоци се идентифицират потенциални рискове и възможности за оптимизация на симулираните информационни масиви. Посоченият пример за бизнес логиката създава предпоставки за ефективно установяване на потенциалните взаимозависими фактори в процеса на генериране на синтетични информационни масиви, техните възможни и невъзможни комбинации. Крайният вариант на бизнес логиката може да бъде разгледан в Приложение 1.

5.4.3 Изчислителна логика

В процесите по генериране на синтетични масиви изчислителната логика играе ролята на двигател за създаване на нова стойност и знания. За разлика от трансформационната логика, настоящата рамка използва математически модели и формули за изчисляване на нови метрики, които не съществуват директно в изходните данни. Такъв тип метрика може да бъде изчисляването на индекс на клиентска лоялност, базирано върху множество променливи като честота на покупките, средна стойност и взаимодействие с поддръжката. В литературата съществува мнението, че [23] аналитичните възможности на големите данни се реализират напълно, само когато сложната изчислителна логика се прилага автоматизирано върху големи информационни набори.

Разглежданата логическа рамка изисква значителни изчислителни ресурси. Тя може да включва статистически изводи, вероятностни модели и дори резултати от машинно обучение, интегрирани директно в процеса на синтез. Ключовият момент при разработване на логика от този характер е свързан с мащабируемостта – способността алгоритъмът да се изпълнява паралелно върху разпределени клъстери.

5.4.4 Филтрираща бизнес логика

От гледна точка на класификацията, филтриращата бизнес логика спада към логическите рамки за редукция на параметри. Най-съществените аспект от работата на подобен тип структурирани модули е за намаляване на шума в данните и оптимизация на фокуса. Филтриращите правила се използват за премахване излишни записи, дубликати или информация, която е извън текущия контекст на анализа. Този процес представлява не просто изтриване, а последователен подбор, извършен въз основа на предварително установени бизнес приоритети. Според някои автори [24], в системите с интензивно използване на данни, способността да се филтрира ненужната информация на ранен етап е от решаващо значение за ефективността на цялата архитектура.

Филтриращата бизнес логика се използва както за техническо пречистване на наличната информационна база, така и за приложение в процеса за сегментиране на пазарите и изготвяне на целеви анализи. Разглежданият тип логическа рамка, както и предходната класификация, е тясно свързан с концепцията за мащабируемост, тъй като намалява обема на данните, които преминават към последващите, по-тежки изчислителни слоеве. Посредством последователността от подобни действия се постига търсения баланс между пълнота на информацията и производителност на управленските системи.

5.5 Етапи на създаване на бизнес логика

Методологическата рамка за изграждане на бизнес логика представлява структуриран цикъл, целящ превръщането на абстрактни представи в дискретни алгоритмични функции. Този процес е ключов за осигуряване на семантичната коректност на генерираната информация и съдържа в себе си следната последователност от фази:

Установяване на набор от правила

Първата фаза от процеса се фокусира върху извличането и структурирането на априорната информация, на чиято основа се изгражда цялостната рамка на бизнес логиката. Последователността обхваща подфаза, в която се извършва задълбочено наблюдение и изучаване както на бизнес средата, така и на управленските процеси в организацията, като целта е да се идентифицират скритите зависимости между различните променливи. В резултат на тези действия се осъществява строго дефиниран набор от правила, които са придружени от съответните логически и физически ограничения, дефиниращи границите на допустимите стойности на синтетичните масиви.

Формализация

Втората фаза от процеса е тази на формализацията. На този етап се осъществява трансформация на дескриптивните бизнес изисквания в алгоритмичен вид. В разглежданата фаза качествените правила се трансформират в математически зависимости и логически вериги. Имплементирането им в конкретни алгоритми позволява на системата да интерпретира бизнес контекста без двусмислие, превръщайки стратегическите насоки в изпълнима последователност от алгоритмични действия, който се имплементират в модела за генериране на синтетични данни.

Валидация на бизнес логиката

Валидацията на бизнес логиката е ключов момент за проверка (в случай на налични реални масиви от информация), при който новосъздадената логика

се подлага серия от тестове върху исторически масиви от реални данни. Целта е да се направи проверка на коректността на алгоритмите посредством съпоставка на синтезираните изходи с вече известни резултати. Одитът спомага за идентифициране на логически грешки или прекомерни отклонения преди системата да бъде имплементирана в продуктивна среда, като по този начин се създават предпоставки за подобряване на статистическата устойчивост на модела.

Приложение

Етапът на приложение обхваща стъпките по имплементирането на установената бизнес логика в процеса по синтезиране на информационни масиви, Основният фокус е върху установяването на невъзможните или слабо вероятни записи и тяхното премахване от финалния набор от синтетични данни. Настоящата фаза осигурява запазване на висока степен на достоверност на генерираните стойности, както и създаване на данни, които са подходящи както за обучение на модели, така и за сложни симулационни процеси, интегрирани в управлението.

Обратна връзка

Процесът завършва с фазата на обратна връзка. Въз основа на установени отклонения от първоначално зададената концепция по време на експлоатацията или настъпване на промяна във външните фактори, е необходима корекция на правилата и ограниченията. Този етап превръща бизнес логиката в адаптивна система, която има установен цикъл на прецизиране, за да съответства на променящата се динамика на данните и околната среда.

5.6 Оценка на ефективността и качеството

5.6.1 Качество на бизнес логиката

Точността на бизнес логиката се измерва посредством прецизността на синтезираните данни спрямо реалните информационни масиви. Тази метрика оценява доколко алгоритмите успяват да запазят както количествените характеристики, така и семантичното качество на първоначалните данни. Високото ниво на съвпадение гарантира, че синтезираният изход запазва всички статистически зависимости. Наличието на такова съответствие създава предпоставки за пълен информационен трансфер между реалните масиви и симулираната среда.

5.6.2 Време за реакция

Времето за реакция представлява съществен показател за ефективност на системата при обработка на големи обеми информация. То отразява скоростта, с която бизнес логиката се прилага върху масивни структури от данни. Малките времеви диапазони са индикатор за висококачествена оптимизация на алгоритмите и ефективно използване на хардуерните ресурси, което създава предпоставки за получаване на изчистени синтезирани информационни масиви, без съществено забавяне. В контекста на мащабируемостта, този параметър дефинира способността на архитектурата да поддържа висока пропускателна способност при нарастваща сложност на правилата.

5.6.3 Адаптивност

Адаптивността измерва динамичния потенциал на системата да интегрира промени в кратки срокове. Този показател се фокусира върху времето, необходимо за техническо внедряване и валидиране на ново бизнес правило в отговор на корекции в пазарната стратегия или регулаторната рамка. Високо адаптивните системи създават възможности за модифициране на логическите зависимости, като по този начин се осигурява съответствие между синтезираните данни и актуалните цели на бизнеса. Тази гъвкавост е критична за поддържане на конкурентно предимство в бързо променяща се среда.

5.7 Заключение

Бизнес логиката представлява ключов архитектурен слой в моделите за генериране на големи масиви от данни. Тази рамка представлява не само техническа спецификация, а критичен семантичен мост между суровите информационни структури и крайните стратегически и управленски цели, осъществени посредством генерираните информационни масиви. Настоящата глава подчертава, че интегрирането на бизнес правила в алгоритмите за синтез гарантира контекстуална валидност и семантична цялост, предотвратявайки генерирането на информационен шум. Съществена е ролята на бизнес логиката като регулаторен филтър, създаващ предпоставки за удовлетворяване на строгите регулации за поверителност. Посредством приложената в настоящото изследване таксономия, обхващаща трансформационна, интегрираща, изчислителна и филтрираща логика, се открива възможността за автоматизиран синтез на знания. Посредством стъпките на формализация и валидация се постига трансформация на абстрактните бизнес изисквания в изпълними алгоритмични последователности. Ефективността на бизнес логиката,

измервана чрез точност, време за реакция и адаптивност, определя способността на моделите да осигурят адекватност на синтезираните масиви в динамичната пазарна среда. Бизнес логиката способства за трансформиране на аналитичните модели от чисто статистически инструменти в подходи, интегриращи причинно-следствени ограничения, което е от съществено значение за оптимизацията на процесите по изработване на управленски решения. Така структурираният логически слой изиграва съществена роля в процеса по установяване на цялостен и адекватен информационен набор, посредством минимизиране рисковете от грешки при генериране на данни.

Използвана литература

[1] Fowler, M. (2002). Patterns of enterprise application architecture. Addison-Wesley Professional

[2] Papazoglou, M. P. (2003). Service-oriented computing: Concepts, characteristics, and directions. In Fourth International Conference on Web Information Systems Engineering pp. 3-12. IEEE

[3] Van Der Aalst, W. M. P. (2013). Business Process Management: A Comprehensive Survey. ISRN Software Engineering, 2013, 1-37

[4] Erl, T. (2005) Service-Oriented Architecture: Concepts, Technology, and Design. Prentice Hall, Upper Saddle River.

[5] Lenzerini, M. (2002). Data integration: A theoretical framework. Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, 233–246.

[6] Rosemann, M., vom Brocke, J. (2010). The Six Core Elements of Business Process Management. In: Brocke, J. v., Rosemann, M. (eds) Handbook on Business Process Management 1. International Handbooks on Information Systems. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-00416-2_5

[7] Ceri, S., Bozzon, A., Brambilla, M., Della Valle, E., Fraternali, P., & Quarteroni, S. (2013). Web Information Retrieval. Springer Science & Business Media

[8] Van der Aalst, W. M., Adriansyah, A., de Medeiros, A. K. A., Arcieri, F., Baier, T., Blickle, T., & Wynn, M. T. (2011). Process mining manifesto. In Business Process Management Workshops (pp. 169-194). Springer.

[9] Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. MIS Quarterly, 1165-1188.

[10] Davenport, T. H., & Harris, J. (2017). Competing on Analytics: Updated, with a New Introduction: The New Science of Winning. Harvard Business Review Press.

[11] Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. Harvard Data Science Review, 1(1). https://doi.org/10.1162/99608f92.8cd550d1

[12] O'Neil, C., (2016) Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York: Crown Publishers, 2016. 272p. Hardcover, ISBN 978-0553418811

[13] Vargo, S. L., Lusch, R. F. (2016). Institutions and axioms: an extension and update of service-dominant logic. J. of the Acad. Mark. Sci. 44, 5–23. https://doi.org/10.1007/s11747-015-0456-3

[14] Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. Journal of Business Research, 70, 263-286. https://doi.org/10.1016/j.jbusres.2016.08.001

[15] Dong, X. L., & Srivastava, D. (2013). Big data integration. In Proceedings of the 2013 IEEE 29th International Conference on Data Engineering (ICDE), 1245-1248. DOI: 10.1109/ICDE.2013.6544914

[16] Herschel, M., Diestelkämper, R. & Ben Lahmar, H. A survey on provenance: What for? What form? What from? The VLDB Journal 26, 881–906 (2017). https://doi.org/10.1007/s00778-017-0486-1

[17] Simmhan, Y. L., Plale, B., & Gannon, D. (2005). A survey of data provenance in e-science. ACM SIGMOD Record, 34(3), 31-36. https://doi.org/10.1145/1084805.1084812

[18] Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., Franklin, M. J., Ghodsi, A., Gonzalez, J., Shenker, S., & Stoica, I. (2016). Apache Spark: A unified engine for big data processing. Communications of the ACM, 59(11), 56–65. https://doi.org/10.1145/2934664

[19] Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2014, July). Big Data and Its Technical Challenges. Communications of the ACM, 57. https://doi.org/10.1145/2611567

[20] Vassiliadis, P. (2009) A survey of Extract-transform-Load technology. International Journal of Data Warehousing and Mining, 5, 1-27. https://doi.org/10.4018/jdwm.2009070101

[21] Bernstein, P. A., & Melnik, S. (2007). Model management 2.0: Manipulating richer objects. In Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data (pp. 1–12). Association for Computing Machinery. https://doi.org/10.1145/1247480.1247482

[22] Rahm, E. and Do, H. H. (2003) Data Cleaning Problems and Current Approaches. IEEE Bulletin on Data Engineering, 7, 182-185.

[23] Wang, Y., Kung, L., and Byrd, T. A. (2018) Big Data Analytics: Understanding Its Capabilities and Potential Benefits for Healthcare Organizations. *Technological Forecasting and Social Change*, 126, 3-13. https://doi.org/10.1016/j.techfore.2015.12.019

[24] Wu, X., Zhu, X., Wu, G.-Q. and Ding, W. (2014) Data Mining with Big Data. IEEE Transactions on Knowledge and Data Engineering, 26, 97-107. https://doi.org/10.1109/TKDE.2013.109

III. МЕТОДОЛОГИЯ ЗА ГЕНЕРИРАНЕ НА ДАННИ

6 Методологични аспекти и методологично дърво

Автори: Ангел Марчев, мл., Васил Марчев

Резюме: Настоящата глава представя преглед на ключови подходи за създаване на многомерни синтетични набори от данни, включително вероятностно конкатениране, мащабиране с бизнес правила, обратна копула за семплиране и генеративни състезателни мрежи. Тези методи са насочени към справяне с предизвикателства като недостиг на данни, запазване на поверителността и поддържане на точност на представяне. В процеса на работа се имплементира методологично дърво, с цел да се систематизират случаите, при които генерирането на синтетични данни е от съществено значение, от обработката на непълни набори от данни до обогатяването на съществуващи. Акцентът е поставен върху осъществяването на предварителни одити за минимизиране на данните, осигуряване на изчислителна ефективност и подобряване на резултатите в процеса на валидиране. Въпреки че настоящият текст не е емпиричен, в него са синтезирани теоретични постановки, като същевременно е предложен синтезиран набор от данни за справка. Констатациите допринасят за установяването на набор от най-добри практики в генерирането на синтетични данни, позволявайки на изследователите ефективно да отговорят на изискванията на съвременните изследвания, основани на данни.

Ключови думи: синтез на данни, невронна мрежа, синтетични данни.

6.1 Въведение

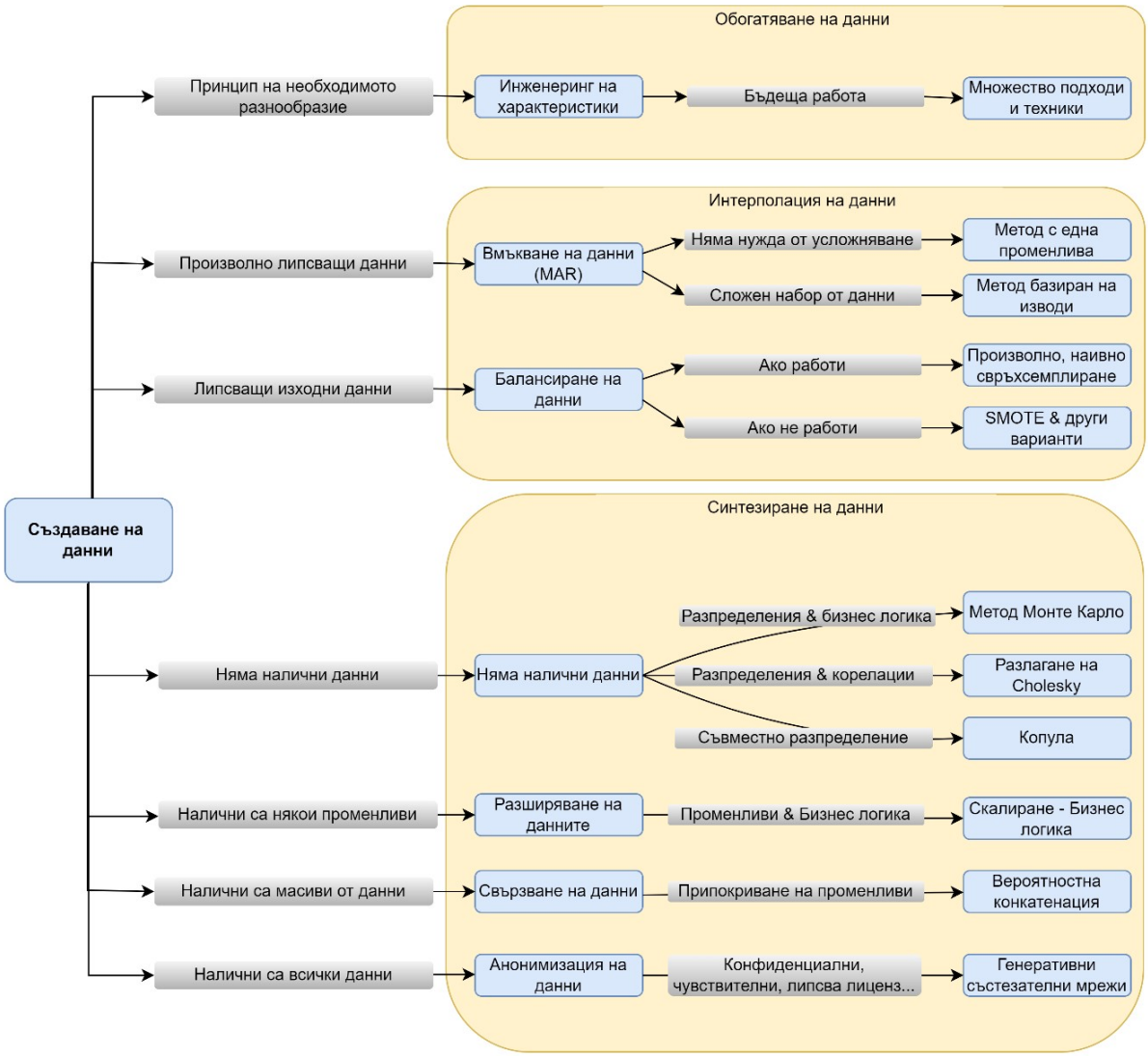
Настоящата глава обобщава методологиите, прозренията и теоретичните резултати в изследваните методи за синтез на данни, реализирани в процеса на изследване на тематиката. Синтезирани са открития от обширни анализи и моделиране, за да се предостави цялостно описание на изследователския път.

В процеса на работа е реализиран сравнителен анализ на различните подходи за генериране на данни, като се извеждат критични зависимости, които определят ефективността на моделите в реални сценарии. Този аналитичен процес не само потвърждава теоретичните хипотези, но и очертава перспективите за бъдещи иновации в областта на автоматизирания синтез.

6.2 Методологически насоки в процеса по създаване на данни

В методологичен аспект се наблюдават няколко различни ситуации, които подчертават необходимостта от прилагане на структуриран процес за

генериране на многомерен набор от данни. Тази глава се фокусира върху основните случаи, при които генерирането на данни може да се счита за неизбежна стъпка в процеса на анализ на информацията, систематизирани в методологичното дърво. То има три основни клона: обогатяване на данни, интерполация на данни и синтез на данни (фиг. 6.2-1).



Фиг. 6.2-1 Методологично дърво

6.2.1 Обогатяване на данни

Принцип на необходимото разнообразие. Тази концепция подчертава важността на наличието на разнообразни и изчерпателни входни данни, за да се гарантира ефективността и точността на анализа и моделирането на данните. Една от основните подфази в разглеждания случай е разработването на допълнителни характеристики. Тази стъпка включва избор, трансформиране и създаване на нови характеристики от сурови данни, за да се подобри производителността на моделите за машинно обучение.

Принципът притежава жизненоважна роля за извличане на смислени прозрения от данните и подобряване на прогнозиращите способности на алгоритмите.

Следващата подстъпка в процеса се характеризира с възможността за използване на множеството методи разнообразни методи в рамките на описаната задача. Този подход използва разнообразни методологии, алгоритми или процедури свързани с необходимостите за анализ и интерпретация на данните. Резултатите от изследванията могат да бъдат подсилени и направени по-възприемчиви чрез включване на тези компоненти в процеса на генериране на данни.

6.2.2 Интерполация на данни

Случайно липсващи данни. В случаите, при които се наблюдават липсващи данни на случаен принцип, могат да се приложат следните подстъпки: въвеждане на MAR данни. Това е ключова фаза за запазване на точността, особено при работа с набори от данни, при които липсващите стойности се появяват на случаен принцип.

На тази стъпка моделът се разклонява и протича в два различни потока:

- Ако няма необходимост от усложнения. Когато възникне такава ситуация, се прибягва до прилагане на методи с една променлива.
- Ако са налични сложни и многовариантни набори от данни, се прибягва до методи, базирани на изводи.

Липсващи данни за даден клас. Когато наблюдават липсваща информация за някои изходни данни, за определен клас, е необходимо да се добави нов блок към модела. Този модул се използва за балансиране на данните. Небалансираните набори от данни могат да представляват предизвикателство за моделите за машинно обучение. Моделите могат да станат пристрастни към класа на мнозинството, което води до по-ниска производителност в класа на малцинството, който често е класът, който ни интересува. В тази стъпка моделът има два изхода:

- Ако моделът работи правилно, тогава се използва случайно наивно свръхсемплиране.
- Ако моделът не работи правилно, се използва техника за свръхсемплиране на извадки при работа с небалансирани категории за класификация в набор от данни (Synthetic minority oversampling technique – SMOTE) и други варианти.

6.2.3 Синтез на данни

В някои случаи спецификата на данните, сложността и закономерността предполагат трудности при осигуряването на модела с необходимото количество и качество на информация. При липса на данни възниква необходимостта от генериране на търсените масиви. Случаите, в които е необходимо да се синтезират данни, се разделят на няколко основни подгрупи:

6.2.3.1 При липса на данни

Ако изобщо няма данни. Тази ситуация възниква в случаите, при които се работи с данни, попадащи под GDPR, Закона за банкова тайна или други закони за защита на личните данни. Ако възникне ситуация, в която няма наличен набор от данни, има три различни подхода за обработката му:

- Първият подход е базиран на разпределения и бизнес правила. Във втората фаза е предвидено да се използва методът на Монте Карло, който стъпва върху характеристиките на данните [3, 10].
- Вторият метод използва разпределения и корелационна матрица. Като втора стъпка, този метод се основава на разлагането на Холецки [1].
- Третият подход използва многовариантно съвместно разпределение. Във втората фаза на този метод се използва инверсно семплиране с помощта на копула [4].

6.2.3.2 При наличие на частични данни под формата на съществуващи характеристики

Използването на съществуващи характеристики, в съществуващ набор от данни, е от решаващо значение за генерирането на данни. Тази техника се нарича разширяване (мащабиране) чрез бизнес правила, при която новите характеристики се синтезират въз основа на съществуващите и се добавят бизнес правила за мащабиране на новите стойности. Процесът има за цел да обогати набора от данни с допълнителна информация, подобрявайки способността на модела да обобщава и да прави точни прогнози чрез интегриране на подходящи характеристики и бизнес правила.

6.2.3.3 Частични данни под формата на съществуващи вариращи набори от данни

Третият подход е вероятностно конкатениране. Той се разглежда при условие на свързване на два или повече реални набора от данни, които се съпоставят с уникален идентификационен ключ. Идентификационният ключ се създава от две или повече променливи, които присъстват и в двете матрици [4].

Когато са налични различни набори от данни, с повече от една повтаряща се характеристика във всяка, е подходящо да се приложи подходът на конкатенация на данни. Втората фаза на метода е свързана с припокриване на множеството характеристики.

6.2.3.4 Съществуващ пълен нелицензиран набор от данни

Когато всички данни съществуват. В случаите, когато се обработват лични данни и се установи, че цялата необходима информация е налична, се използват методи, насочени към анонимизиране на данните.

Случай на употреба	Фази			
	Одит за минимизиране…	Генериране от…	Хоризонтална синхронизация от…	Вертикална валидация от…
Разпределения и бизнес правила	Набор от променливи с разпределения	Симулация на Монте Карло с оглед на характеристиките	Бизнес правила	Тест на Колмогоров-Смирнов
Разпределения и корелационна матрица	Списък с корелирани характеристики	Разлагане на Холецки	Корелационна матрица	Тест на Колмогоров-Смирнов
Съвместно разпределение	Списък на променливите, включени в съвместното разпределение	Обратна копула за семплиране	Напасване на копулата	Тест на Колмогоров-Смирнов
Някои променливи вече са налични и бизнес правила	Пропорция към броя на първоначално наличните променливи	Мащабиране по функции чрез бизнес правила	Бизнес правила	Бизнес правила
Налични частични набори от данни	Подходящи набори от данни за конкатенация	Вероятностно конкатениране	Вероятностно съвпадение	Не е приложимо
Нелицензиран набор от данни	Изчислителни ресурси по време на обучение	Генеративни състезателни мрежа	Състезателно обучение	Косинусово подобие

Таблица 6.2.3.4-1 Преглед на случаите на употреба за синтез на данни, представени във фази на приложение.

Такива подходи имат основната задача да направят използваема наличната поверителна, чувствителна или информация, изискваща лиценз, използвайки анонимизиращи методи. В тази ситуация като подходящ метод се разглеждат генеративните състезателни мрежи (GAN).

6.3 Работни фази

От гледна точка на науката за анализ на данни, процесът на генериране на синтетични информационни набори има няколко основни фази. Настоящата обобщена методологична последователност на различните подходи има за цел да покаже основните насоки.

6.3.1 Одит насочен към минимизиране на синтеза на данни

Тази фаза подчертава критичната важност на провеждането на предварителни одити, насочени към минимизиране на обема данни, предназначени за синтез. Такива одити са от съществено значение за ефективното управление на ресурсите и за повишаване на осъществимостта на последващата валидация на наборите от данни. Изпълнението на предварителен одит за минимизиране на данните, предназначени за синтез, може значително да облекчи тези предизвикателства.

Генерирането на синтетични данни е процес, изискващ изчислително време, често характеризиращо се със сложни алгоритми и голяма изчислителна мощност. Чрез провеждане на предварителен одит за идентифициране и изключване на несъществени данни могат да бъдат спестени изчислителни ресурси, да бъде намалено времето за обработка, да се намалят оперативните разходи и да се увеличи производителността. Този подход е в съответствие с принципите за минимизиране на данните, които се застъпват за събиране и обработка само на данните, необходими за конкретна цел. Спазването на тези принципи не само оптимизира използването на ресурсите, но и подобрява сигурността на данните и спазването на разпоредбите за защита на данните.

Осигуряването на валидност на синтетичните набори от данни е от първостепенно значение за поддържането на тяхната полезност и надеждност. Намаленият набор от данни, постигнат чрез предварително минимизиране, опростява процеса на валидиране, като намалява сложността и обема на данните, които трябва да бъдат изследвани. Всяка допълнителна синтезирана характеристика увеличава експоненциално риска за поява на несъответствия в данните, тъй като сложността на поддържането на точни взаимовръзки и статистически свойства между характеристиките нараства с размерността на набора от данни. Това рационализиране улеснява по-ефективното

идентифициране на несъответствия или грешки, като по този начин подобрява цялостното качество на синтетичните данни. Освен това, фокусираният набор от данни позволява по-целенасочени стратегии за валидиране, гарантиращи, че критичните атрибути на данните са точно представени.

6.3.2 Генериране на данни по (избран) метод

Тази фаза включва действителния синтез на данни, използвайки избран метод, като например обратно семплиране с помощта на копула или генериране на характеристики, GAN или симулация, базирана на общи разпределения и бизнес логика и др. Изследователите използват различни техники за генериране на синтетични данни или създаване на нови характеристики от съществуващи, подобрявайки набора от данни за анализ и моделиране.

6.3.3 Симулация на Монте Карло с оглед на характеристиките

Монте Карло симулациите [5] представляват клас алгоритми за случайни изчисления. Генерирането на данни, базирано на този метод, разчита на многократно случайно вземане на проби. Случайното генериране на променлива X, базирано на дадена извадка от вероятностното разпределение D(θ). В представения случай с θ се означават параметрите на разпределението. Ако има необходимост от серия от синтетични точки от данни, описаният процес се повтаря *N* на брой пъти.

$$X_i \;\; D(\theta), i = 1, 2, \ldots, N$$

Генерираните синтетични данни се изследват в следващата фаза, за да се потвърди дали съответстват на разпределението на оригиналните данни.

6.3.4 Разлагане на Холецки

Положително определена матрица може да бъде факторизирана с помощта на разлагането на Холецки [8]. Разлагането на Холецки на положително определена матрица A представлява долна триъгълна матрица L, такава че:

$$A = L\,L^T$$

Разлагането на Холецки може да се използва за симулиране на ковариационната структура между променливите в оригиналните данни при създаване на синтетични данни.

$$X = \mu + LZ$$

Където с L се обозначава разлагането на Холецки на ковариационната матрица на X, а μ е вектор от средни стойности. Това гарантира, че ковариационната структура на оригиналните данни се запазва в генерираните синтетични данни.

Чрез споменатото уравнение може да се конструира набор от корелирани нормални променливи, $X = (X_1, X_2, \ldots, X_d)$, като се използва даден набор от независими стандартни нормални променливи $Z = (Z_1, Z_2, \ldots, Z_d)$ [1].

6.3.4.1 Обърната копула за генериране на извадки

Копулата е математическа функция, която се с характеризира колективното разпределение на група случайни променливи. Обикновено се използвана за моделиране на взаимозависимостта между разглежданите променливи. Даден е набор от данни, включващ *n* екземпляра от *d* променливи, изследователите могат да представят съвместното разпределение на тези променливи чрез копула, обозначена като $C(u_1, u_2, \ldots, u_d)$, където u_i представлява равномерни случайни променливи [11].

Техниката на обратното семплиране с копула включва напасването ѝ към първоначалния набор от данни, последвано от семплиране от тази копула, с оглед генериране на синтетични данни, които точно поддържат основната структура на зависимости, наблюдавана в оригиналния набор от данни [1].

6.3.4.2 Разширяване на данните чрез бизнес правила

Моделът използва мащабиращи фактори и коефициенти, предоставени от експерти, осигурявайки съответствие с установените бизнес правила. На всяка новогенерирана променлива се присвояват уникални тегла, определени от корелацията между съществуващите променливи и симулираните. Резултатът е нова матрица, която включва както оригиналните, така и симулираните данни.

Нека входният набор от данни D е матрица с N реда и k колони:

$$D \in R^{N \times k}$$

Дефинира се вектор от параметри *p*, който присвоява тегла (коефициенти на мащабиране) на съответните колони:

$$p=\left[p_1, p_2, \ldots, p_k\right]$$

където $p_i \geq 0$ за i-тата променлива, с $p_i = 0$за неизползвани променливи. Изчислява се претеглена комбинация x_i от колоните на набора от данни:

$$x_i = D \cdot p^T$$

Процесът на синтезиране на променливи може да включва кодиране, категоризация, мащабиране на характеристики и филтриране, базирано на правила.

6.3.4.3 Вероятностно конкатениране

Този метод включва конкатениране (свързване) на два или повече реални набора от данни, използвайки уникален идентификационен ключ, генериран от споделени променливи в двете матрици [4]. Процесът започва със създаване на идентификационния ключ от поне две променливи, присъстващи в наборите от данни, след което се преминава към осъществяване на ляво външно съединение, за да се комбинират всички записи, което води до нова матрица, съдържаща всички възможни варианти на данните. За да се гарантира, че матрицата е подходяща за моделиране, дубликатите и невалидните записи се почистват посредством премахване на повтарящи се редове от лявата матрица, като по този начин наборът от данни се редуцира до прецизирана, използваема форма.

Да приемем, че съществува споделено подмножество от колони, Z, между набора от данни X и набора от данни Y, такова че:

$$Z \subseteq \backslash\{1, \ldots, n\backslash\} \cap \backslash\{1, \ldots, m\backslash\}$$

Дефинира се операция за ляво външно съединение на X и Y въз основа на (Z), която води до конкатениран набор от данни C:

$$C = X \bowtie_Z Y$$

или подмножества със съответстващи ключове $Z_X \subseteq X$ и $Z_Y \subseteq Y$, където:

$$Z_X = Z_Y \text{ and } Z_X \neq 0$$

Изберете произволен ред от Y за всеки съответстващ ред в X, като осигурите уникалност в C.

6.3.4.4 Генеративни състезателни мрежи

GAN [6; 7] представляват клас архитектура за дълбоко обучение, която може да се използва за генериране на синтетични данни. Разглежданият метод се състои от две невронни мрежи. Тези невронни мрежи се наричат генератор, обозначен с *G,* и дискриминатор, обозначен с *D*. Те се обучават заедно в игра с нулева резултат. Генераторната невронна мрежа е отговорна за създаването на синтетични данни, докато дискриминаторната мрежа има за цел да разграничи реалните от генерираните синтетични данни [12].

6.3.5 Хоризонтална синхронизация

Хоризонталната синхронизация се отнася до уеднаквяването на новополучени данни и осигуряването на съгласуваност и кохерентност в новите записи. За тази цел се разглеждат няколко подхода: бизнес правила, копула или корелации. Процесът трябва да гарантира допустимостта на всеки отделен запис от новополучени данни.

Чрез избора на потенциално взаимозависими компоненти, техните ограничения, комбинации и невъзможни такива, се установява наборът от бизнес правила, съчетани с бизнес логиката. Комбинирането на разпределенията може да доведе до записи, които са невъзможни за постигане или изключително малко вероятни. Целта на имплементацията на бизнес логиката на модела е да елиминира нежелани, невероятни и/или невъзможни комбинации.

При синхронизацията на данни, копулите могат да помогнат за улавяне на сложни взаимовръзки между променливи, особено когато традиционните мерки за корелация може да не са достатъчни. Чрез оценка на корелацията между различните елементи от данните, анализаторите могат да разберат как промените в една променлива могат да повлияят на друга, което подпомага усилията за синхронизация.

6.3.6 Вертикално валидиране на характеристики

За да се гарантира целостта на данните и да се запази първоначалното разпределение, се използва вертикална валидация при прогнозираното разпределение на всяка характеристика или променлива.

Вертикалната валидация е ключов фактор при извличане на нови функции или колони от съществуващи данни, анонимизиране на чувствителна информация, синтезиране на изкуствени данни или конкатениране на множество информационни набори. Тя гарантира, че трансформираните или

комбинирани данни поддържат очакваните статистически свойства и зависимости, получени от началните разпределения.

Необходимостта от валидиране се разглежда като критичен аспект от процеса на генериране на данни. Съвместното разпределение на променливите и техните взаимозависимости, присъщи на оригиналния набор от данни, може да не са точно отразени в синтетичните масиви. Като следствие от това, резултатите от анализа на всяка синтетична променлива трябва да бъдат проверени спрямо първоначалното разпределение.

Съществуват няколко допълнителни подфази в процеса:

- Анализ на данните. Тази стъпка включва подробен преглед на синтезираните данни и сравнение между различни източници на информация.
- Процесът на валидиране на генерираните данни гарантира, че съответствието на новополучения набор от информация с първоначалните статистически разпределения е интегрирано в симулационния модел.

Важна фаза от процеса е фокусирането върху близките честоти за новополучените данни. Плавният преход между съседни стойности, без резки колебания, е от решаващо значение за валидирането на модела.

За да се установи дали информацията е с приемливо качество, е важно тя да се прегледа. За да се определи дали първоначалните разпределения са запазени, се извършва проверка на новите разпределения.

Съществуват различни подходи за валидиране на генерираните данни, в зависимост както от наборите от информация, така и от методите, използвани в процеса на генериране. Подходи, базирани на различни статистически методи и тестове са разгледани в литературата [2].

6.4 Заключение

Настоящата глава предоставя цялостен преглед на методологиите, теоретичните рамки и практическите съображения, лежащи в основата на създаването на многомерни синтетични набори от данни. Чрез синтезиране на изследванията, проведени от авторския екип, се подчертават ключови подходи за генериране на данни, включително вероятностно конкатениране, мащабиране чрез бизнес правила и усъвършенствани техники като инверсна копула за генериране на извадки и генеративни състезателни мрежи. Всеки метод демонстрира гъвкавостта и адаптивността, необходими за справяне с предизвикателствата, свързани с недостига на данни, поверителността и точността на представяне.

Споделените тук насоки подчертават значението на структурираните процеси, като предварителни одити за минимизиране на данните и фази на

валидиране, за да се гарантира качеството, надеждността и релевантността на синтезираните данни. Методологичното дърво и специфичните за всеки случай стратегии, очертани в тази глава, служат като ръководство за бъдещи приложения при генериране на данни, разширявайки възможностите за изследвания в области с ограничен достъп до реални набори от информация.

Използвана литература

[1] Marchev, A., Jr, Marchev, V., (2023). Automated Algorithm for Multi-variate Data Synthesis with Cholesky Decomposition, ICACS '23: Proceedings of the 7th International Conference on Algorithms, Computing and Systems, https://doi.org/10.1145/3631908.3631909

[2] Barbierato, E., Vedova, M. L. D., Tessera, D., Toti, D., and Vanoli, N., (2022). A methodology for controlling bias and fairness in synthetic data generation. Applied Sciences, 12(9):4619

[3] Marchev, A., Marchev, V., & Efremov, A. (2025). Methodology for cold-start modeling using synthetic data. AIP Conference Proceedings, 3182(1), 020002. https://doi.org/10.1063/5.0246285

[4] Marchev, A., Jr, Marchev, V., (2022). SYNTHESIZING MULTI-DIMENSIONAL PERSONAL DATA SETS. AIP CONFERENCE PROCEEDINGS; 2505 (1): 020012

[5] Del Moral, P., Doucet, A., Jasra, A., (2006). Sequential Monte Carlo samplers. J. R. Statist. Soc. B (2006) 68, Part 3, pp. 411–436

[6] Goodfellow, I., J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., (2014). Generative Adversarial Nets

[7] Honghee Ko, D., Hassan, A., Suk, J., and Choi, J., (2020). "Korean Font Synthesis with GANs," International Journal of Computer Theory and Engineering vol. 12, no. 4, pp. 92-96, 2020.

[8] Dereniowsky, D., Kubale, M., (2003). Cholesky Factorization of Matrices in Parallel and Ranking of Graphs, Parallel Processing and Applied Mathematics, 5th International Conference, PPAM 2003, Czestochowa, Poland, September 7-10, 2003

[9] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., (2016). Improved Techniques for Training GANs

[10] Marchev, V., Marchev, A., Piryankova, M., Masarliev, D., Mitkov, V., (2023), Synthesizing an anonymized multidimensional dataset featuring financial, economic, demographic, and personal traits data, VSIM, vol. 19, no. 1, 2023, ISSN 1314-0582

[11] Marchev, A., Lyubchev, D., Marchev, V., (2024). Inverse Copula Sampling for Multi-Dimensional Data Synthesis, 50th International

Conference Applications of Mathematics in Engineering and Economics, Sozopol, June 07-13, 2024, AIP Conference proceedings (in publishing)

[12] Marchev, A. Marchev, V. (2024). - Methodological Considerations for Anonymizing Tabular Data Using Generative Adversarial Networks, ICACS '24: Proceedings of the 8th International Conference on Algorithms, Computing and Systems

[13]

7 Синтезиране на многомерни набори от лични данни

Автори: Ангел Марчев, мл., Васил Марчев

Резюме. Настоящата глава представя синтезиран списък от различни методи за синтезиране на набори от синтетични данни, състоящи се от лични анонимизирани многомерни записи. В края на главата е представен кратък сравнителен анализ. Синтезирани информационни набори са приложими за различни бизнес употреби, особено в случаите, при които е идентифицирана необходимост от идентифициране и анализ на характеристики, на ниво индивид (физическо лице) като същевременно не е допустимо използване на реални информационни масиви. Като извод, който може да бъде направен, като следствие от разгледаните методи е, че е установена методология за анонимизиране и генериране на данни за случаите, при които не са налични данни или масивите, които са налични са недостатъчни.

Ключови думи: подходи за генериране на данни, бизнес логика, разпределения, корелации, конкатениране, анонимизация, валидиране на данни

7.1 Въведение

Процесът на анализ и моделиране на данни обхваща няколко основни момента. Значително място заема фазата на установяване на информация с необходимото количество и качество, от гледна точка на модела. Спецификата на разглежданите информационни набори, тяхната комплексност и юридическото обезпечаване на процеса предполагат редица трудности при осигуряване на необходимите данни. Данните, необходими за създаването на модела за оценка, попадат в рамките на европейското законодателство, GDPR, Закона за кредитните институции в частта му уреждаща въпросите за банковата тайна.

Фазата на събирането на данни посредством познатите стандартни методи за анкетиране представлява сложен и бавен процес. Като се вземат като отправна точка гореописаните трудности в последователността се идентифицира алтернативен подход за получаване на необходимото количество данни. За целта е предложен набор от подходи за симулиране на синтетична база данни, базирана на набор от различни демографски, лични, индивидуални и банкови променливи и показатели.

7.2 Симулация базирана на общи разпределения и бизнес логика

7.2.1 Методология

Първата фаза от процеса обхваща установяването на съвкупност от фактори и съответстващите им разпределения. За всяко едно от наблюдаваните разпределения е валидно, че е получено в резултат на изследване, априорни знания, правдоподобни допускания, анализ на банковата среда и др.

Основните етапи на процеса са: избор на променливи, анализ на разпределенията, бизнес логика, симулация, валидиране на процеса. Основните поделементи, на чиято база е разработен моделът на симулацията, са демографските характеристики, личните характеристики, банковите и финансовите характеристики.

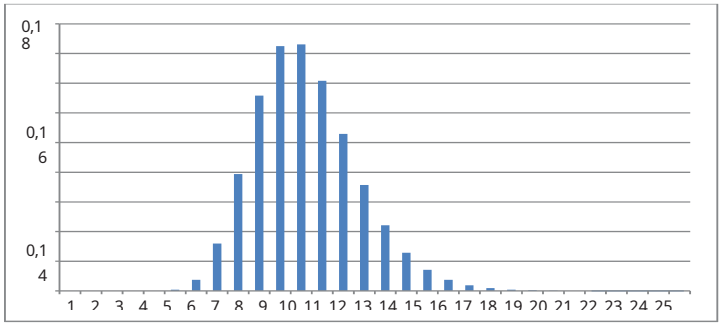
7.2.1.1 Анализ на разпределения

Съществуват три основни подхода, разглеждани в настоящата глава, при анализа на разпределенията, за целите на симулацията на данни: достатъчно данни, недостатъчно данни, правдоподобни допускания.

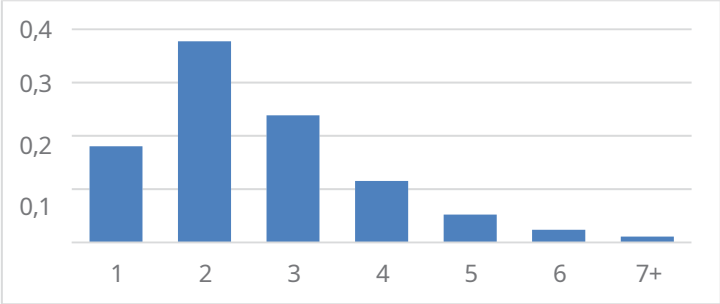
В случаите, при които са налице достатъчно данни за установяване на разпределенията, се допуска, че липсват променя на условията. Запазване на първоначалните разпределения за всеки от факторите, за които имаме достатъчно данни.

Когато не са налице достатъчно данни за създаване на разпределения, се използва информацията, която е налична за конкретната променлива, като средна стойност на данните, минимум, максимум, среднопретеглена стойност и др., и се прави частична симулация въз основа на налична предишна информация.

Правдоподобни предположения се използват в случаите, при които няма информация, посредством която да се създаде разпределение. В този случай се прави правдоподобно предположение за разпределенията, основано на априорни познания за променливите.



Фиг. 7.2.1.1-1 Среден брой банкови операции



Фиг. 7.2.1.1-2 Брой членове на домакинство

7.2.1.2 Бизнес логика

Бизнес логиката, която се използва в модела за симулация на данни, се реализира посредством подбиране на потенциално взаимозависими фактори, техните ограничения, възможни комбинации и невъзможни такива. При комбиниране на разпределенията съществува вероятност да се генерират недостижими или много малко вероятни стойности. Чрез имплементацията на бизнес логика в рамките на модела се цели премахване на малко вероятните и невъзможните комбинации.

Процесът на имплементация на бизнес логиката обхваща следните фази:

- Подбор на потенциално взаимозависими фактори
- Установяване на възможните и невъзможни комбинации
- Премахване на невъзможните комбинации

Извадков пример, съдържащ списък с правила, които са част от установената бизнес логиката може да бъде разгледан в таблица 7.2.1.2-1 Пълният списък с правила на бизнес логиката може да съдържа стотици различни норми.

Фактор	**Комбиниращ фактор**	**Информация**
Възраст	Степен на образование	Под 24 години е по-малко вероятно да има завършено висше образование
Възраст	Семеен статус	Под 24 години е по-малко вероятно да е женен/омъжена, вдовец/а. Над 64 има по-голяма вероятност да е вдовец/а.
Възраст	Брой деца под 18	От 20 до 24 години е по-малко вероятно да има повече от едно дете под 18 години. Над 65 години е по-малко вероятно да има дете под 18 години.

Таблица 7.2.1.2-1 Бизнес логика на модела (извадка)

7.2.1.3 Симулация

Методологията на работа предполага фаза за настройване на метода на работа на проектирания модел с дефинираните и предварително установени разпределения. Това представлява процес на подготовка на алгоритми в

специализиран софтуер, в който се задават съответните ограничителни условия за разпределението на факторите.

Процесът на разработване на модела започва с процедура по получаване на едномерни случайни променливи от реални данни и техните разпределения. Симулираните стойности е необходимо да съответстват на статистическите разпределения, които са получени от реалните информационни източници. Разглежданата стъпка трябва да бъде еквивалентна за всички променливи зададени в модела. При наличие на новогенерирани данни е необходимо всички симулирани вектори да бъдат обединени в една матрица. За тази матрица е дадено, че всяка колона предоставя информация за различна променлива. В процеса по генериране на данни (фиг. 7.2.1.3-1) се установява възможност за получаване на недостижими или много малко вероятни стойности. При възникване на такава ситуация се, моделът преминава през фазата на редуциране. При тази фаза редовете, които не съответстват на зададените бизнес правила биват редуцирани от новата матрица.

7.2.2 Математически алгоритъм

1) Дадено е ω_m да е едновариантна крайна случайна променлива от реални данни, обозначаващи $m-th$ (от общо s) количествени характеристики на изучаваната система S.

2) Използвайки метода на моментите, характеризиращото разпределение $f_m(\omega_m;\theta_m)$ на ω_m с набор от параметри θ_m се получава като решение на следното SSE

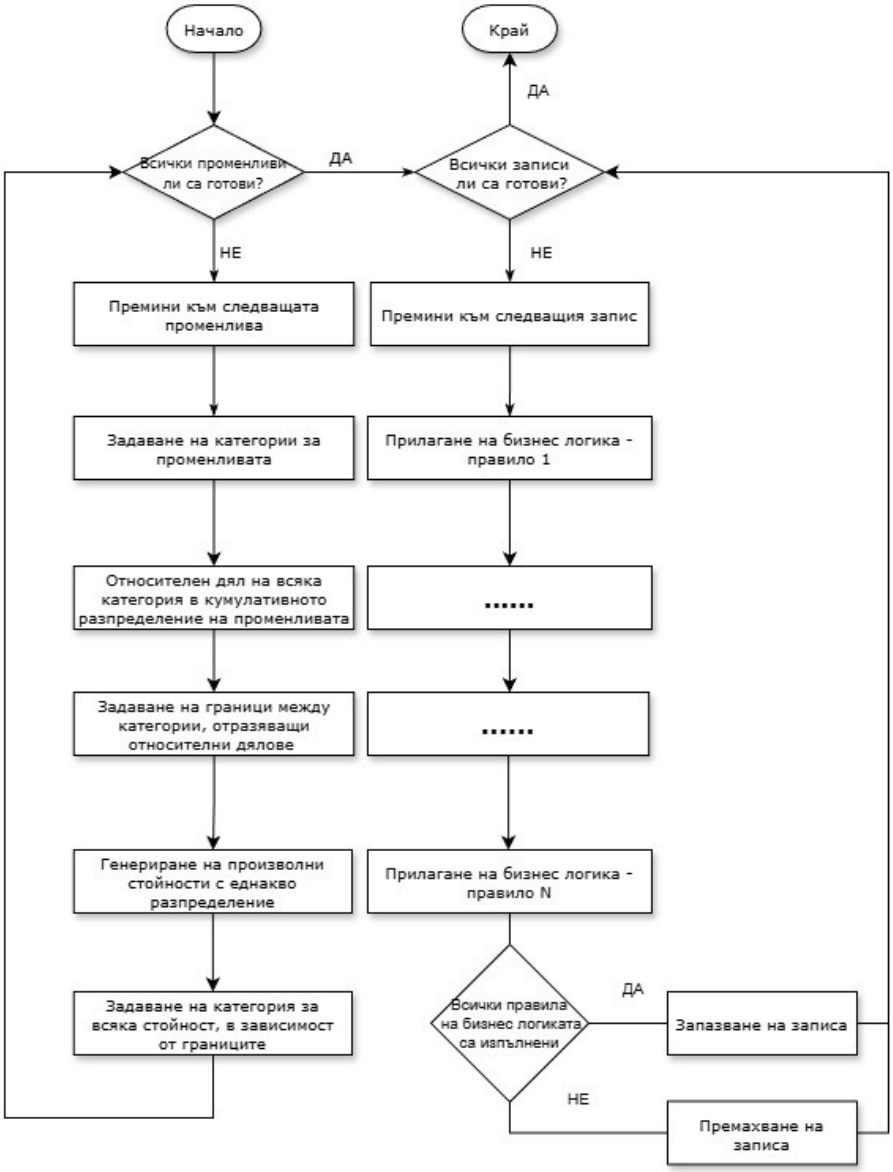
$$\hat{\mu}_{m_1}=\varphi_{m_1}\left(\hat{\theta}_{m_1},\hat{\theta}_{m_2},\ldots,\hat{\theta}_{m_k}\right)$$
$$\hat{\mu}_{m_2}=\varphi_{m_2}\left(\hat{\theta}_{m_1},\hat{\theta}_{m_2},\ldots,\hat{\theta}_{m_k}\right)$$
.
.
.
$$\hat{\mu}_{m_k}=\varphi_{m_k}\left(\hat{\theta}_{m_1},\hat{\theta}_{m_2},\ldots,\hat{\theta}_{m_k}\right)$$

Където $\varphi_{m_1}\cdots\varphi_{m_k}$ са известните функции и $\hat{\mu}_{m_1}\cdots\hat{\mu}_k$са първите k моменти на, f_m получени от ω_m.

3) Генериране на стойности от симулирания вектор x_m с дължина t, съгласно f_m, така че $\{x_m \mid f_m(x_m,\theta_m)\sim f_m(\omega_m,\theta_m)\}$ и е подреден на случаен принцип $x_m=\{x_{m_1},x_{m_2},\ldots,x_{m_t}\}$.

4) Следващата стъпка е конкатениране x_m в матрица X: $X_{t,s}=[x_1 x_2 \ldots x_s]$

5) Даден е набор от v бизнес правила, обозначени със система от неравенства $A.z \leq b$, където A е $v \times s$ е матрица, X която се редуцира до $X^{\%}$ такава, че $X^{\%}_{n,s} \subseteq X_{t,s}$ и $X^{\%} = ¿$, където $n \leq t$.



Фиг. 7.2.1.3-1 Процес на моделиране

7.3 Симулация, базирана на общи разпределение и корелации

7.3.1 Методология

Разглежданият метод използва синтезиран списък от общи разпределения, които са налични за анализираните променливи, сравнявайки корелациите между тях [1]. Този вид симулация може да бъде декомпозиран на две основни идеи.

Първата е свързана с разглежданите общи разпределения на променливите. За целта на метода е необходимо получените данни да отговарят

на статистическите разпределения. За симулираните информационни масиви се изисква да запазят първоначалното общо разпределение на променливите.

Вторият основен етап от процеса се разглежда при корелации между променливите. Алгоритъмът се основава на установени корелации между променливите. Те са в пряка зависимост от използваните променливи в процеса на симулиране на характеристиките. Ключовите елементи за този метод са необходимостта от разпределение и корелация между променливите, използвани за симулация.

Алгоритъмът започва с процедурата за получаване на едномерни случайни променливи от реални данни и техните разпределения. Тези разпределения се характеризират с метода на моментите. Когато има разпределения на всички променливи, е необходимо да се стандартизират стойностите между минус едно и едно със средна стойност нула. Генерираните стойности трябва да са в съответствие с нормалните разпределения.

Според описаните експертни познания, се получават коефициенти на корелация на новата променлива, съобразно с една от дадените променливи. Процесът преминава през фазата на симулация на новите данни, които са стандартизирани. Получените нови данни трябва да бъдат дестандартизирани с параметри за мащабиране, които също трябва да бъдат зададени. Когато има нова колона с генерирана информация, тя трябва да бъде свързана с останалата част от матрицата на симулираните информационни масиви.

7.3.2 Математически алгоритъм

1) Дадено е ω_m да е едновариантна крайна случайна променлива от реални данни, обозначаващи $m-th$ (от общо s) количествени характеристики на изучаваната система S.

2) Използвайки метода на моментите, характеризиращото разпределение $f_m(\omega_m;\theta_m)$ на ω_m с набор от параметри θ_m се получава като решение на следното SSE

3)
$$\hat{\mu}_{m_1}=\varphi_{m_1}\left(\hat{\theta}_{m_1},\hat{\theta}_{m_2},\ldots,\hat{\theta}_{m_k}\right)$$
$$\hat{\mu}_{m_2}=\varphi_{m_2}\left(\hat{\theta}_{m_1},\hat{\theta}_{m_2},\ldots,\hat{\theta}_{m_k}\right)$$
$$\vdots$$
$$\hat{\mu}_{m_k}=\varphi_{m_k}\left(\hat{\theta}_{m_1},\hat{\theta}_{m_2},\ldots,\hat{\theta}_{m_k}\right)$$

Където $\varphi_{m_1}\cdots\varphi_{m_k}$ са известните функции и $\hat{\mu}_{m_1}\cdots\hat{\mu}_k$са първите k моменти на, f_m получени от ω_m.

4) Стандартизира се ω_m стойността, така че нейната средна да е 0, а стандартното отклонение да бъде 1, като се използва $\overline{\omega}_m = \frac{\omega_m - \hat{\mu}_{m_1}}{\hat{\mu}_{m_2}}$

5) Генериране на стойности от симулирания вектор w с дължина n, съгласно нормално разпределение на $\dot{f}$, така че $\{w \vee \dot{f}(w, \theta_{m_1} = 0, \theta_{m_2} = 1)\}$.

6) При даден коефициент на корелация ρ измерван с реални данни $\overline{\omega}_m$ се симулират нови данни за стандартизирани стойности на x_1като например $\overline{x}_1 = \rho . \overline{\omega}_m + \sqrt{1 - \rho^2 . \omega}$

7) Дестандартизира се $\omega_m = \overline{\omega_m} \cdot \widehat{\mu_{m_2}} + \widehat{\mu_{m_1}} u\, x_l = \overline{x}_l \cdot \gamma_{l_2} + \gamma_{l_1}$, където γ_{l_1} и γ_{l_2} са мащабиращи параметри дадени за

8) Следващата стъпка е конкатениране на ω_m и x_l в матрица $\widetilde{X}: \widetilde{X}_{n,s} = [\omega_1 \ldots \omega_k x_1 \ldots x_l]$, където $s = k + l$ колони.

7.4 Обърната копула за подбор на извадки

7.4.1 Методология

Разглежданият метод наблюдава и описва зависимостите между променливите, като предоставя възможности за създаване на разпределения, моделиращи корелирани многовариантни данни. Алгоритъмът предоставя подход за конструиране на многовариантно разпределение, който е базиран на подход за определяне на оптималните едновариантни разпределения. Въз основа на функция на копула, алгоритъмът изработва структура на корелации между зададените променливи. Разглежданият алгоритъм предоставя възможност за използване както на двувариантни разпределения, така и на такива от по-висока размерност. В описвания алгоритъм се използва едно от свойствата на функцията на копула – напасването ѝ към многовариантното разпределение, след което се преминава към обръщане на част от резултатите, получени от копулната функция [2, 3].

Стойностите, които се получават в резултат на действието на копулата са нормализирани между 0 и 1, така че необходимо да се подходи към денормализиране, посредством мащабиране на параметрите.

Последният етап от процеса представлява конкатенация на получената матрица с денормализираните данни.

7.4.2 Математически алгоритъм

1) Дадено е $\Phi = f_1, \ldots, f_s(\omega_1 \ldots \omega_s)$ да бъде многомерно съвместно разпределение на m крайни случайни променливи ω от реални данни, обозначаващи m броя от количествени характеристики на изучаваната система

S. Тогава кумулативната функция на разпределение се дефинира като $\Phi = P\left(\omega_1 \le \widehat{\omega_1}, \omega_2 \le \widehat{\omega_2}, \ldots, \omega_s \le \widehat{\omega_s}\right)$, където P показва вероятността всички случайни променливи ω_i да бъдат по-малки или равни на кореспондиращите стойности $\widehat{\omega_i}$;

2) Използва се функция на копула C, за да се генерират симулирани нормализирани извадкови вектори $\left(x_1, x_2, \ldots, x_m\right)$ с дължина n от Φ като $C\left(\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_s\right) = \Phi\left[\omega_1 \le f_1^{-1}\left(\overline{x}_1\right), \omega_2 \le f_2^{-1}\left(\overline{x}_2\right), \ldots, \omega_s \le f_s^{-1}\left(\overline{x}_s\right)\right]$.

3) Денормализира се $x_s = \overline{x}_s . \gamma_{s_2} + \gamma_{s_1}$ където γ_{s_1} и γ_{s_2} са параметри на мащабиране, дадени за x_s

4) Конкатенира се x_m в матрица $\widetilde{X}$: $\widetilde{X}_{n,s} = \left[x_1 x_2 \ldots x_s\right]$

7.5 Симулация на данни, базирана на реални данни

7.5.1 Методология

В процесите по симулация на информационни масиви, използващи като основа набори от предишни анонимизирани съвкупности от реални данни, посредством които се генерират нови променливи, както и съответстващите им записи. Основните фази от последователността, които могат да се обособят са както следва:

Осигуряване на малка съвкупност от **реални данни.** За целите на описвания подход са необходими няколко променливи от реални данни.

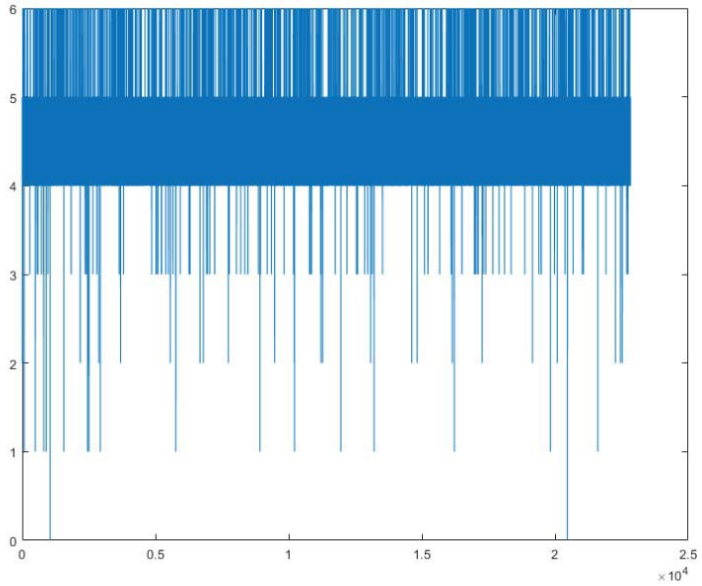
Анонимизиране на реалните информационни масиви. Тази фаза обхваща действията за криптиране на съществуващия набор от данни. Процесът на криптиране на набора от данни осигурява премахването на чувствителна информация като: имена, лични номера, имейли, адреси и др. от масива. Действията предполагат анонимизиране на целия набор от лична информация, с изключение на демографските данни. Описваната част от процеса представлява задължителна фаза, като се вземе предвид законодателната рамка (GDPR, Закона за кредитните институции, и законите за защита на личните данни на ЕС).

Проектиране на Симулационен алгоритъм за нови променливи. Фаза, в която се програмират условията, според които се генерира масива от синтетични данни.

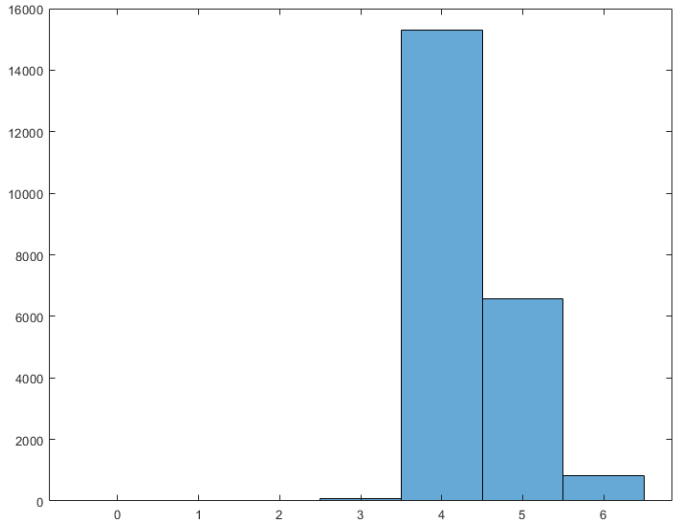
Задаване на различни тегла за основните променливи. Моделът използва като отправна точка мащабиращи фактори и коефициенти на мащабиране, получени от допитване до експерти, които отговарят на зададените бизнес правила. За всяка една от новогенерираните променливи се използват различни теглови коефициенти. От своя страна значимостта на

отделните коефициенти зависи от корелацията между наличните променливи и симулираните.

Последната фаза от процеса е **генерирането на нова матрица**, съдържаща както всички оригинални данни, така и новосъздадените променливи със съответните записи.



Фиг. 7.5.1-1 Степен на образование



Фиг. 7.5.1-2 Степен на образование – разпределение

7.5.2 Математически алгоритъм

1) Дадено е ω да бъде многомерна крайна случайна променлива от реални данни, обозначаваща k броя количествени характеристики на изследваната система S.

2) Генерират се стойности за l от симулирания вектор x_l с дължина n, съгласно $x_l = \omega . B$, където B е вектор-стълб от коефициенти, удовлетворяващи

набор от v бизнес правила, обозначени със система от неравенства $A.(\omega.B) \le b$, където A е $v \times l$ матрица

3) Конкатенират се ω и x_l в матрица $\widetilde{X}$: $^{\widetilde{X}_{n,s}=[\omega x_1 \dots x_l]}$

7.6 Вероятностно конкатениране на бази данни

7.6.1 Методология

Описаният метод се основава на конкатенацията (обединяването) на два или повече съществуващи набора от данни. Подходът се разглежда през призмата на свързване на два или повече реални набора от данни, които биват съпоставени посредством уникален идентификационен ключ. Използваният ключ се генерира посредством обединяването на две или повече променливи, за които е необходимо да присъстват и в двете матрици.

Комбинацията от съществуващи затворени набори от данни преминава през създаването на гореописания идентификационен ключ, като посредством него, моделът създава външно съединение у завършва с пълна комбинация от всички възможни варианти на записите.

В процеса по вероятностно конкатениране на информационни масиви е вероятно да се получат невъзможни или повтарящи се комбинации. В новата матрица е необходимо данните да бъдат подходящи за последваща употреба (моделиране, анализ и др.), което налага последващо почистване на базата посредством премахване на повтарящите се записи от лявата матрица. Резултатът представлява нова матрица, редуцирана посредством изтриване на редове.

7.6.2 Математически алгоритъм

Използван е математически алгоритъм, който гласи следното:

1) Дадено е ω да бъде многомерна, крайна, случайна променлива от реални данни, обозначаваща k брой от числови характеристики в изучаваната система S, докато ω включва n брой редове (така че е $n \times k$ матрица).

2) Дадено е λ да бъде друга отделна многовариантна крайна случайна променлива на реални данни, обозначаваща l брой количествени характеристики на изследвана система S, докато λ включва t брой редове (така че е $t \times l$ матрица).

3) Като се има предвид, че има пресечна точка между ω и λ от поне две колони (докато $n \le t$): $[\omega] \cap [\lambda] = [x_1 \dots x_p \vee x_1 \dots x_p \in \omega \wedge x_1 \dots x_p \in \lambda \wedge p \ge 2]$.

4) Осъществява се ляво външно присъединяване на ω с λ (строго в този ред) $\omega \bowtie \lambda = [\tau_{i,j} \vee i = n \wedge j = k = l - p]$, като се търси съвпадение на съответните

редове, по ключови стойности от $\hat{x}_1,\ldots,\hat{x}_p$, за да се получи конкатенирана матрица X:$^{X_{t,s}=[\omega\lambda]}$

5) Ако се установят съвпадащи подмножества λ и ω, $\left[\lambda \subseteq \lambda \vee \hat{x}_1,\ldots,\hat{x}_p \subseteq \lambda\right] \wedge \left[\omega \subseteq \omega \vee \hat{x}_1,\ldots,\hat{x}_p \subseteq \omega\right]$, където $\lambda > \omega$, избира произволен ред за съчетаване λ, така че X е редуцирано до $\widetilde{X}$ така, че $\widetilde{X}_{n,s} \subseteq X_{t,s}$ и $\widetilde{X}=¿$ и, където $n \leq t$.

7.7 Генеративни състезателни мрежи (GAN)

7.7.1 Методология

Същността на модела се изразява в симулация на данни, базирана на две противопоставящи се невронни мрежи – Дискриминатор и Генератор. Първата невронна мрежа има за цел да класифицира набори от различни извадки, както от реални, така и от симулирани данни по най-подходящ начин. Втората невронна мрежа генерира масиви от информация и е обучена максимално да подвежда дискриминатора [4, 5].

Генераторът представлява невронна мрежа, моделираща функцията на трансформация. Входът ѝ представлява случайна променлива, която трябва да бъде преобучена, с цел да се достигне оптималното разпределение на генерираните масиви от синтетични данни.

От своя страна Дискриминаторът приема като вход начална стойност (данни) и връща, като изход, информация за вероятността тази точка да лежи върху оптималното разпределение. Използва се параметризиран модел за изразяване както на генератора, така и на дискриминатора.

Моделът за генериране на данни има за цел да подведе дискриминатора. Генериращата невронна мрежа е обучена да увеличава максималната грешка при класификация (между реални и генерирани данни), докато, от своя страна, дискриминаторът има за цел да открие неправилно генерираните данни. Дискриминиращата невронна мрежа се обучава да минимизира крайната грешка при класификация.

Основните етапи от процеса обхващат:

- Необходимост от реални набори от данни, които вече са получени
- Въз основа на реалните данни моделът използва генеративна изкуствена невронна мрежа, която синтезира стойности, наподобяващи реалния набор от данни.
- На следващия етап моделът използва друга невронна мрежа, която е дискриминатор. Тази невронна мрежа трябва да разграничи дали наличните стойности са реални или изглеждат реални.

Разглежданият подход не представлява подходящ метод за получаване на стойности, когато няма реален набор от данни. Това е приложим алгоритъм при симулиране на синтетични данни, в случаите на налична предварителна информация за наблюдаваните реални събития. Методът може да бъде използван като подходяща отправна рамка за валидиране на данни или анонимизиране на големи масиви от информация.

7.7.2 Математически алгоритъм

1) Дадено е ω да бъде многомерна крайна случайна променлива от реални данни обозначаваща S на брой (от общо S) количествени характеристики на изучаваната система S, включваща n броя редове.

2) Използвайки генеративна изкуствена невронна мрежа се синтезират стойности от симулираната матрица $\widetilde{X}$: $\widetilde{X}_{n,s}=\left[\widetilde{X}_{n,s} \vee \widetilde{X}_{n,s} \neq \omega\right]$.

3) Използвайки дискриминативна изкуствена невронна мрежа, се дискриминира до $\widetilde{X}_{n,s} \sim \omega$

7.8 Заключение

Методите могат да бъдат разделени най-общо на две големи групи. Първата част обхваща тези алгоритми, които се използват в случаите, при които не се наблюдават твърде много променливи в реалните данни. Това са първите четири показани метода. Те биват разглеждани през динамиката на реалните разпределения на информацията или през призмата на многовариантни разпределения и налична корелационна информация.

Петият метод се разглежда като метод, който е използваем п случаите, при които са налични няколко реални или симулирани информационни набора и като следствие от това е необходимо тези масиви да бъдат обединени. Разглежданият метод представлява алгоритъм за конкатенация. От своя страна последния разгледан метод може да бъде използван при необходимост от анонимизация на реални информационни масиви, които биват използвани за публични изследвания.

Сравнителен анализ	**Общи разпределения и бизнес логика**	**Симулация, базирана на общи разпределения и корелации**	**Обратна копула за семплиране**	**Базирано на симулация на данни него/нея предишна дата**	**Вероятностно конкатениране на бази данни**	**Генеративни състезателни мрежи (GAN)**
Същност	Генериране базиран на данни	Едномерни общи разпределени	Генериране данни, базирани	Симулация - базирана на първоначалн	Конкатенация на два или повече	Самообучаваща се система -

	него/нея едномерни общи разпределения	я & корелации	на многомерни общи разпределения	и данни и коефициенти	набора от данни	базирана на първонача лен набор от информация
Задължително данни	Разпределения на променливите и бизнес логиката	Разпределения на променливи и корелации	Копули на многомерни реални данни	Частични, реални, анонимизирани данни и бизнес правила	Масиви от публични анонимизирани данни	Цял реален набор от данни
Хоризонтално генериране	-	Да	Да	Да	-	Да
Вертикално генериране	Да	Да	Да	-	-	Да
Синхронизация	Правила и бизнес логика	Корелация м-у променливите	Копула	Коефициенти Правила „ Ако/Тогава“	Вероятностно конкатениране	Конкуренция между невронните мрежи
Главни трудности	Необходимост от предположения за част от разпределенията. Редактиране постфактум на набора от данни	Нужда изясняване на корелациите	Извличане на копула от реални данни	Частични, реални, анонимизирани данни. Генериране на данни въз основа на вече генерирани	Масиви от анонимизирани данни – безплатни за използване	Труден и бавен процес обучение на невронните мрежи
Финален резултат	Симулация	Симулация	Симулация	Частична симулация	Конкатениран набор от данни	Симулация

Таблица 7.8-1 Сравнителен анализ на методите за генериране на данни

Описаният метод може да намери своето приложение като методология за валидиране на новогенерирани масиви, по отношение на това дали са подходящи при конкретна задача. Сравнението на описаните методи е достъпно в таблица 7.8-1

Използвана литература

[1] International Business Machines LLC. (2019), IBM SPSS Modeler 18.1.1 Python Scripting and Automation Guide, Simgennode properties pp. 93

[2] Nelsen, R. B. (2006). An Introduction to Copulas (Second ed.). New York: Springer. ISBN 978-1-4419-2109-3

[3] Wilson, A. G., Z. Ghahramani (2010). Copula Processes, arXiv:1006.1350
[4] Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio (2014). Generative Adversarial Nets, arXiv:1406.2661
[5] Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, Xi (2016). Improved Techniques for Training GANs, arXiv:1606.03498
[6]

IV. ПОДХОДИ ЗА ГЕНЕРИРАНЕ НА ДАННИ

8 Автоматизиран алгоритъм за многовариантен синтез на данни с разлагане на Холецки

Автори: Ангел Марчев, мл., Васил Марчев

Резюме: Следващата глава разглежда методи за генериране на синтетични данни, създадени алгоритмично за тестване и обучение на модели за машинно обучение. Вероятностните подходи, като симулация Монте Карло и генеративни състезателни мрежи (GAN), разчитат на случайно вземане на проби. От своя страна методите, които не се базират на вероятности, като обратна копула и разлагане на Холецки, запазват зависимостите и ковариационните структури. Предложеният алгоритъм се фокусира върху генериране на синтетични данни, които запазват както пределните разпределения, така и корелациите. Генерираните данни се валидират с помощта на теста на Колмогоров-Смирнов (KS). Емпиричен пример демонстрира ефективността на методологията, показвайки генерирането на синтетични данни и валидирането спрямо оригиналните разпределения. Това изследване предоставя ценна информация за генерирането на синтетични данни, помагайки на изследователи и практици в анализа на данни и разработването на модели.

Ключови думи: синтетични данни, синтез на данни, разпределения, корелации, разлагане на Холецки

8.1 Методи за генериране на синтетични данни

Генерирането на синтетични данни е мощна техника в науката за данните, използвана за множество приложения, като например увеличаване на обема на информационните масиви, запазване на поверителността, тестване на модели и други. За създаване на синтетични данни могат да се използват както методи основани на вероятностни методи, така и такива, които не са базирани на вероятностни методи, като са описани подробно по-долу.

8.1.1 Методи базирани на вероятности

8.1.1.1 Симулация Монте Карло

Монте Карло симулациите [1] са клас изчислителни алгоритми, които разчитат на многократно случайно вземане на проби, за да изчислят резултатите си. Случайна променлива X се генерира чрез вземане на проби от вероятностно разпределение $D(\theta)$, където θ означава параметрите на

разпределението. Този процес се повтаря N на брой пъти, за да се генерира серия от синтетични точки от данни:

$$X_i \ \ D(\theta), i=1,2,\ldots,N$$

След това генерираните синтетични данни се анализират, за да се потвърди, че те наподобяват разпределението на оригиналните данни.

8.1.1.2 Генеративни състезателни мрежи (GAN)

GAN [2; 3] са вид архитектура за дълбоко обучение, която може да генерира синтетични данни. GAN се състои от две невронни мрежи: генератор G и дискриминатор D, които се обучават заедно в игра с нулев резултат. Генераторът създава синтетични данни, докато дискриминативната функция се опитва да различи реални от синтетични данни

8.1.2 Методи, които не се базират на вероятности

8.1.2.1 Обърната копула за генериране на извадки

Копулата [4; 5] е функция, която описва съвместното разпределение на набор от случайни променливи и често се използва в контекста на моделиране на структурата на зависимостите между тези променливи. Като се има предвид набор от данни с nнаблюдения на d променливи, можем да обозначим съвместното разпределение на тези променливи, използвайки копула $C(u_1,u_2,\ldots,u_d)$, където u_i са равномерни случайни променливи.

Обратното семплиране на копула включва напасването ѝ към оригиналните данни и след това семплиране от тази копула, за да се генерират синтетични данни, които запазват структурата на зависимост на оригиналните данни.

8.1.2.2 Разлагане на Холецки

Разлагането на Холецки [6] представлява факторизация на положително дефинирана матрица. За дадена положително определена матрица A, разлагането на Холецки е долна триъгълна матрица, L така че:

$$A=L\,L^T$$

В контекста на генериране на синтетични данни, разлагането на Холецки може да се използва за моделиране на ковариационната структура между

променливите в оригиналните данни. Като се има предвид набор от независими стандартни нормални променливи $Z=(Z_1, Z_2, \ldots, Z_d)$, набор от корелирани нормални променливи $X=(X_1, X_2, \ldots, X_d)$ може да се генерира чрез:

$$X = \mu + LZ$$

Където μ е вектор от средните стойности, а L е разлагането на Холецки на ковариационната матрица на X. Това гарантира, че генерираните синтетични данни запазват ковариационната структура на оригиналните данни.

8.2 Синтез на данни въз основа на разпределения и корелации - дефиниране на проблема

В предишни изследвания [7] авторите на главата систематизират набор от различни случаи на употреба за синтез на данни. В настоящия текст фокусът се акцентира върху ситуация, при която разпределенията на променливите и корелациите между тях са известни.

В повечето случаи е налице възможност за получаване на разпределенията на променливите от генералната популация или аналогична подизвадка от популацията, а не пълните сурови данни. Тези разпределения биха могли да бъдат получени и от предишни проучвания, получени от малка представителна извадка или да бъдат приети въз основа на теоретични или емпирични доказателства. Това представлява най-очевидният случай на употреба. В настоящото изследване се разглежда процеса, при който са налични имаме не само разпределенията за всяка променлива, но и корелациите между всяка двойка променливи. Тези корелации биха могли да бъдат получени от предишни проучвания или посредством малка представителна извадка. Също така, подходящ метод може да бъде и приемането им въз основа на друга предварителна информация. В такива ситуации са приложими методи като разлагане на Холецки и симулация по метода на Монте Карло, за да бъдат генерирани синтетични данни, които запазват както граничните разпределения, така и корелациите. За дефинираната по-горе проблемна ситуация е приложим алгоритъм, който е подробно разписан в настоящата глава.

8.2.1 Напасване на разпределенията

Обикновено статистическото разпределение може да се обозначи като $X \sim D(\theta)$, където X е случайна променлива, D представлява типа на разпределението, а θ означава набор от параметри, свързани с разпределението.

Например, нормалното разпределение често се представя като $X \sim N(\mu, \sigma^2)$, където μ е средната стойност, а σ^2 е дисперсията. Всеки тип статистическо разпределение има свой собствен набор от параметри, които характеризират неговата форма или свойства. Тези параметри диктуват поведението и свойствата на разпределението, като например неговата централна тенденция, разпространение и асиметрия.

Методът на моментите [8] и обобщеният метод на моментите (GMM) [9] са статистически техники, използвани за оценка на параметрите на разпределение, на база извадка от данни от това разпределение.

При метода на моментите, моментите на извадката се приравняват към моментите на теоретичното разпределение и тези уравнения се решават спрямо параметрите. Например, ако $X_1, X_2, \ldots, X_n$е извадка от нормално разпределение, средната стойност на извадката $\overline{X}$ и дисперсията на извадката S^2 могат да бъдат приравнени към теоретичната средна стойност и дисперсията μ и σ^2, което води до уравненията:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i =, S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^2 = \sigma^2$$

Решенията на тези уравнения дават оценки на параметрите, получени по метода на моментите.

GMM е обобщение на метода на моментите, което позволява по-голяма гъвкавост при избора на моментни условия. Той предоставя начин за оценка на параметрите, когато има повече моментни условия, отколкото параметри, или когато моментните условия не са точно разрешими. GMM оценката $\hat{\theta}$ е решението на оптимизационната задача:

$$\hat{\theta} = argmin_{\theta \in \Theta} g(\theta)' W_g(\theta)$$

Където $g(\theta)$е вектор от моментни условия на извадката, Wе положително определена тегловна матрица, а Θе параметричното пространство. Тази оценка е особено полезна, когато моментните условия са извлечени от икономическа или поведенческа теория, както често се случва в иконометрията.

8.2.2 Формулиране на корелационна матрица

При наличие на корелациите между всяка двойка променливи, се конструира корелационната матрица. Корелационната матрица е квадратна матрица, която съдържа коефициентите на корелация за всяка двойка

променливи. Коефициентът на корелация е мярка за линейната връзка между две променливи, вариращa от -1 (перфектна отрицателна корелация) до 1 (перфектна положителна корелация). Корелационната матрица може да се използва за улавяне на зависимостите между променливите.

Например, ако $X_1, X_2, \ldots, X_n$са nпроменливи, корелационната матрица R може да бъде представена като:

$$R = \begin{vmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1m} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & \rho_{nm} \end{vmatrix}$$

Къде ρ_{ij}е коефициентът на корелация между X_iи X_j.

8.2.3 Разлагане на Холецки върху корелационната матрица

Разлагането на Холецки е метод за разлагане на положително дефинирана матрица, като например корелационна матрица, в произведение на долна триъгълна матрица и нейната транспонирана матрица. Това може да се използва за трансформиране на набор от некорелирани променливи в набор от корелирани променливи.

Като се има предвид корелационната матрица R, разлагането на Холецки може да бъде представено като:

$$R = L\,L^T$$

Където L е долна триъгълна матрица.

Приложението на разлагането на Холецки не се ограничава до корелационни матрици на нормално разпределени променливи. Според Qu, W., Liu, H. и Zhang [10], тази методология може да се разшири до други ненормални разпределения. По-конкретно, може да се започне с ненормални некорелирани случайни променливи и да се приложи серия от трансформации, подобни на разлагането на Холецки, за да се получат корелирани ненормални променливи.

Обобщението на разлагането на Холецки към ненормални разпределения може да се осъществи чрез следните стъпки:

Изчисляване на моменти: Като се имат предвид многомерната асиметрия (β_1), ексцес (β_2) и броят на променливите (p), зададени от потребителя, $\xi_j (j = 1, 2, \ldots, p)$се изчисляват третият и четвъртият подреден момент на случайната променлива. В математически термини, ако m_n означава n^{th} момент на случайна променлива, третият и четвъртият момент се представят като:

$$m_3 = E\left[\left(\xi j - E[\xi j]\right)^3\right]$$
$$m_4 = E\left[\left(\xi j - E[\xi j]\right)^4\right]$$

Генериране на стандартизирани променливи: Стандартизираните променливи ξ_jсе генерират чрез модифицирания степенен метод, образувайки вектора ξ. Ако σ_j означава стандартното отклонение на ξ_j, стандартизираната променлива се дава от:

$$\xi j' = \frac{\xi j - E[\xi j]}{\sigma_j}$$

Разлагане и трансформация на Холецки: Зададената от потребителя корелационна матрица (или ковариационна матрица) се разлага с помощта на разлагането на Холецки, за да се получи матрицата A.

8.2.4 Симулация чрез метод Монте Карло

За да използваме резултатите от разлагането на Холецки, е необходимо да присъства поне една променлива.

Монте Карло симулацията може да се използва за генериране на синтетични данни чрез извличане на случайни извадки от известните разпределения за всяка променлива. Монте Карло симулациите разчитат на многократно случайно вземане на извадки, за да изчислят резултатите си. Случайна променлива Xсе генерира чрез вземане на извадки от вероятностно разпределение $D(\theta)$, където θ означава параметрите на разпределението. Този процес се повтаря N на брой пъти, за да се генерира серия от синтетични данни:

$$X_1 \;\; D(\theta), i = 1, 2, \ldots, N$$

След това генерираните синтетични данни се анализират, за да се потвърди, че те наподобяват разпределението на оригиналните данни.

8.2.5 Матрично произведение на матрицата на Холецки

Това е процесът на прилагане на разлагането на Холецки към извадковите данни, за да се въведат желаните корелации. Ако е известна само една променлива, многократно се изчислява произведението $X_i = L X_1, i = 1, 2, \ldots, N$, а

след това отново с новите оценени променливи, докато всички променливи бъдат оценени.

В противен случай, се преминава към изчисляване на произведението $X = LZ$, където Z означава набор от извадки от реалните данни. Получените променливи $X_1, X_2, \ldots, X_n$ имат желаните корелации, както е зададено от корелационната матрица R.

Този подход ни позволява да генерираме синтетичен набор от данни, който не само съвпада с разпределенията и корелациите на променливите, но и включва нелинейни зависимости.

8.2.6 Валидиране на резултатите

Трябва да се отбележи, че синтетичните данни може да не отразяват точно съвместното разпределение на променливите или зависимостите между променливите в оригиналните данни. Следователно, резултатите от всеки анализ трябва да бъдат валидирани за всяка синтетична променлива спрямо желаното разпределение.

В разглежданата методология подходящият метод за валидиране на новополучения набор от синтетични данни е тестът на Колмогоров-Смирнов (за повече подробности – глава 4 Валидиране на синтетични данни)

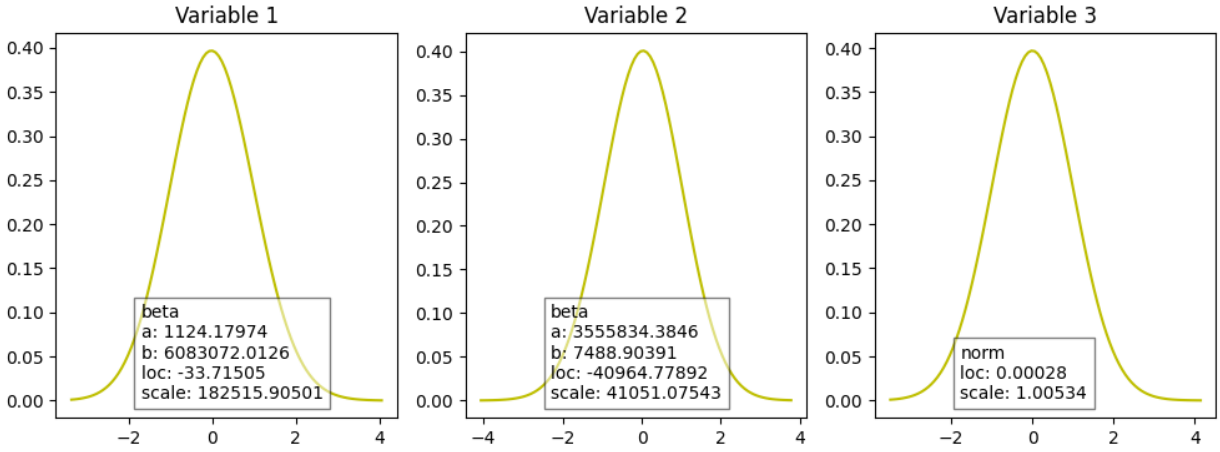
8.3 Емпиричен пример

Като илюстрация към статията, е представен кратък симулационен експеримент, следвайки зададения алгоритъм. Експериментът се състои в симулиране на три променливи от хипотетичен набор от данни. Експериментът е проведен в следната среда:

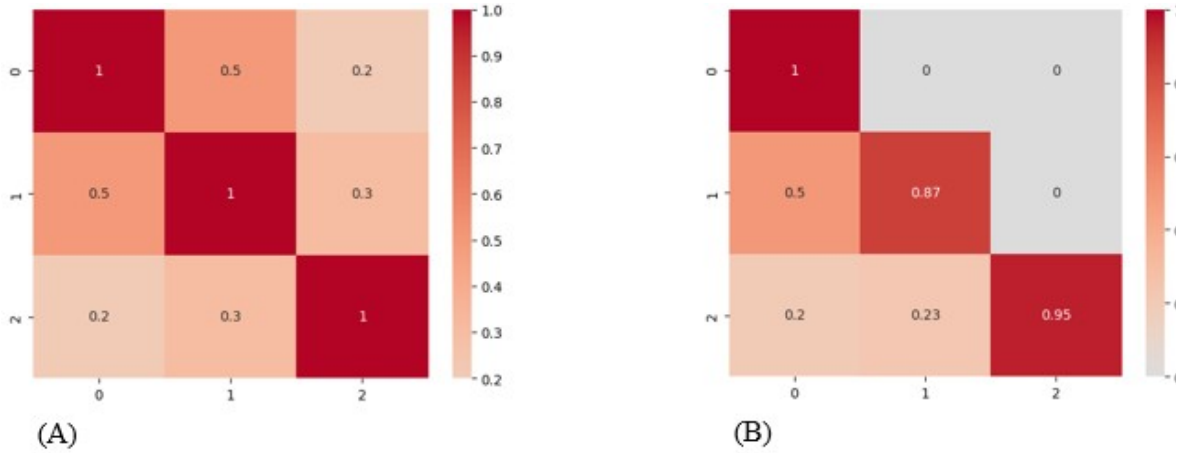
Архитектура на машината: x86_64; Операционна система: Linux 5.19.0-35-generic

Версия на Python: 3.9.13; Библиотеки: Numpy, Pandas, Scipy, fitter, seaborn, matplotlib, platform.

Процесът стартира с три случайни променливи. Като допълнителна стъпка се извежда статистическото разпределение за всяка променлива, използвайки метода на моментите фиг. 8.3-1



Фиг. 8.3-1 Получени статистически разпределения за три променливи с техните параметри

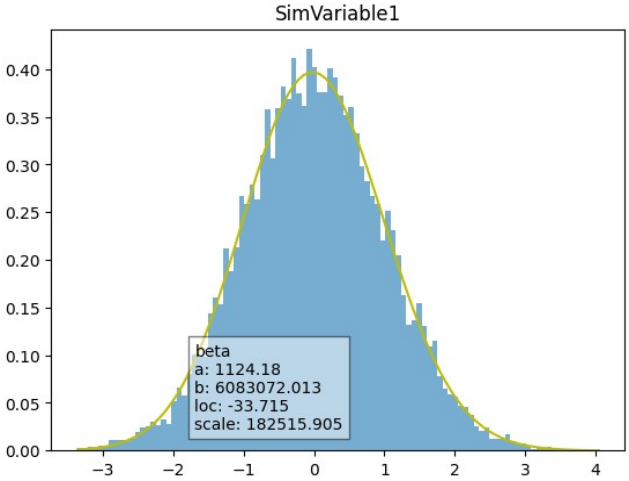


Фиг. 8.3-2 Изведена корелационна матрица (A) и изведена матрица на разлагане на Холецки (B)

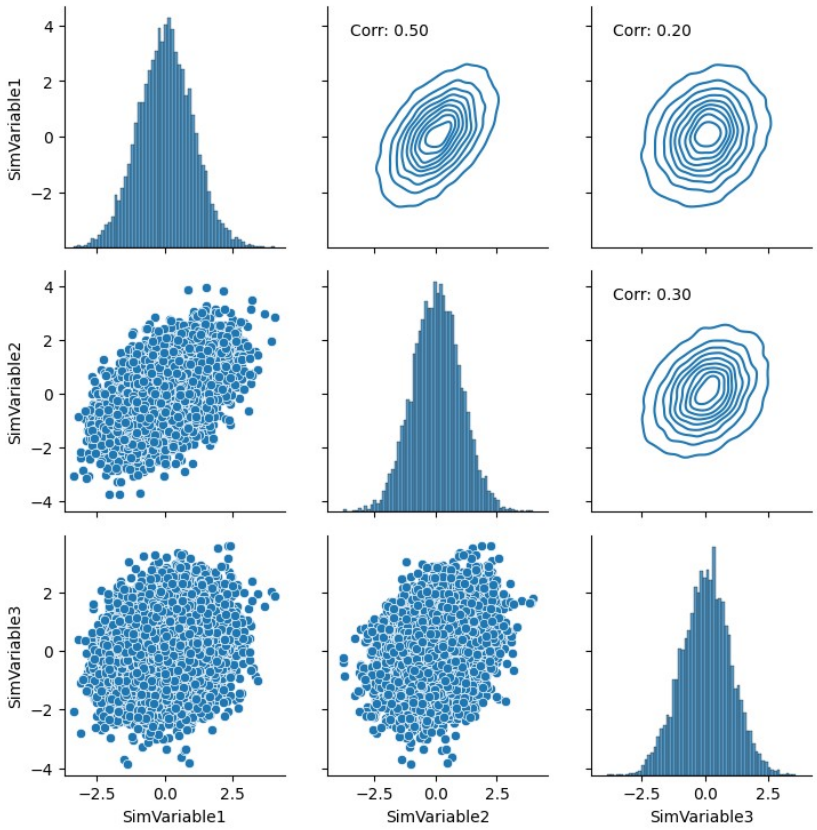
Второ (фиг. 8.3-2A), създава се корелационната матрица като необходим компонент за алгоритъма. Следващата стъпка е да се изчисли разлагането на Холецки, което води до полуконюгирана триъгълна матрица на корелационната матрица. (фиг. 8.3-2B).

Следващата стъпка е да се генерира първата симулирана променлива чрез симулация Монте Карло (фиг. 8.3-3).

След като е генерирана първата променлива се преминава към извеждане на останалите променливи чрез умножение на матрицата на Холецки. Използват се два вида визуализация – долният ляв ъгъл се намира точковата диаграма, а горният десен ъгъл представлява контурна карта на съвместното разпределение (фиг. 8.3-4).



Фиг. 8.3-3 Симулиране на променлива чрез симулация Монте Карло, следвайки оригиналното статистическо разпределение



Фиг. 8.3-4 Диаграма на всички синтетично симулирани променливи

Накрая, всяка симулирана променлива преминава през статистическо тестване спрямо оригиналното статистическо разпределение, използвайки теста на Колмогоров-Смирнов, и се проверява за значимост чрез метода на p-стойността (табл. 8.3-1).

	SimVariable1	SimVariable2	SimVariable3
ks_values	0.019996	0.016770	0.026365
p_values	0.000664	0.007133	0.000002

Таблица 8.3-1: Статистически тест на синтетичните променливи

8.4 Заключение

В настоящото изследване е разгледан подробно процесът по генериране на синтетични данни като средство за симулиране на информация от реалния свят, като целта е новополучените синтетични записи да са приложими в областта на науката за данни. Обсъдени са подходи базирани на вероятности както и такива, които не се основават на вероятностни методи за генериране на синтетични данни, като са очертани техните предимства и областите им на приложение.

Представен е предложеният алгоритъм за генериране на синтетични данни, който запазва както граничните разпределения, така и корелациите. Алгоритъмът включва напасване на разпределението за всяка променлива, която представлява интерес. Валидирани са синтетичните данни с помощта на теста на Колмогоров-Смирнов и са използвани техники като разлагане на Холецки за моделиране на ковариационната структура.

Чрез емпиричен пример са демонстрирани ефективността на предложената методология при генериране на синтетични данни. Малкият симулационен експеримент демонстрира процеса на генериране на синтетични променливи, валидиране на тяхното разпределение спрямо оригиналните данни и оценка на значимостта им с помощта на KS теста и p-стойностите.

Резултатите от експеримента потвърждават надеждността и полезността на предложения подход при генериране на синтетични данни, които запазват желаните разпределения и корелации. Процесът на валидиране с помощта на теста на Колмогоров-Смирнов повишава увереността в подобието на синтетичните данни по отношение на оригиналните масиви.

Значението на това изследване се състои в способността за преодоляване на предизвикателствата, свързани със събирането на данни, като например опасения за поверителност, етични проблеми и логистични трудности. Синтетичните масиви от информация предлагат алтернативен подход, който заобикаля тези предизвикателства, като генерира записи, имитиращи свойствата на данните от реалния свят, като същевременно не е необходим компромис с чувствителната информация, съдържаща се в оригиналните масиви.

Чрез предоставяне на цялостен преглед на различни методи за генериране на синтетични данни и предлагане на алгоритъм, който отчита както разпределенията, така и корелациите, това изследване допринася за развитието на техниките за генериране на синтетични данни. Изследователите и анализаторите могат да прилагат тези техники за справяне с различни предизвикателства, свързани с данните, включително тестване и валидиране на модели за машинно обучение, допълване на данни, запазване на поверителността и други.

Бъдещите изследователски насоки биха могли да се фокусират върху разширяване на обхвата на предложената методология чрез отчитане на допълнителни фактори, които влияят върху генерирането на данни, като например нелинейни зависимости, хетероскедастичност и зависимости от по-висок порядък. Изследването на производителността на синтетичните данни в специфични области или приложения, като здравеопазване или финанси, също би било ценно.

В заключение, синтетичните данни представляват ценно решение за генериране на данни, които могат да се използват за различни цели, от обучение на модели до запазване на поверителността. Настоящото проучване предоставя задълбочено разрез на методите за генериране на синтетични данни и предлага алгоритъм, който запазва както пределните разпределения, така и корелациите. Чрез емпирични експерименти е демонстрирана ефективността на предложения подход. Процесът на валидиране с помощта на теста на Колмогоров-Смирнов гарантира надеждността на синтетичните данни, което ги прави ценен ресурс за анализ на данни и разработване на модели. Чрез използването на синтетични данни, изследователите и практиците могат да преодолеят предизвикателствата пред събирането на данни и да подобрят разбирането и приложенията в науката за данни.

Използвана литература

[1] Del Moral, P., Doucet, A., & Jasra, A. (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 68*(3), 411–436. https://doi.org/10.1111/j.1467-9868.2006.00553.x

[2] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 27* (pp. 2672–2680). Curran Associates, Inc.

[3] Ko, D. H., Hassan, A. U., Suk, J., & Choi, J. (2020). Korean font synthesis with GANs. *International Journal of Computer Theory and Engineering, 12*(4), 92–96.

[4] Nelsen, R. B. (2006). An introduction to copulas (2nd ed.). Springer Science+Business Media.
[5] Strelen, J. C. (2009). Tools for dependent simulation input with copulas. Proceedings of the 2nd International ICST Conference on Simulation Tools and Techniques. ICST. https://doi.org/10.4108/ICST.SIMUTOOLS2009.5630
[6] Dereniowsky, D., & Kubale, M. (2004). Cholesky factorization of matrices in parallel and ranking of graphs. In R. Wyrzykowski, J. Dongarra, M. Paprzycki, & J. Waśniewski (Eds.), *Parallel Processing and Applied Mathematics* (pp. 985–992). Springer.
[7] Marchev, A., & Marchev, V. (2022). Synthesizing multi-dimensional personal data sets. *AIP Conference Proceedings, 2505*(1), 020012.
[8] Pearson, K. (1936). Method of moments and method of maximum likelihood. *Biometrika, 28*(1/2), 34–59.
[9] Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica, 50*(4), 1029–1054. https://doi.org/10.2307/1912775
[10] Qu, W., Liu, H., & Zhang, Z. (2020). A method of generating multivariate non-normal random numbers with desired multivariate skewness and kurtosis. Behavior Research Methods, 52(3), 939–946. https://doi.org/10.3758/s13428-019-01291-5

9 Генериране на данни при наличие на частични информационни набори

Автори: Ангел Марчев, мл., Васил Марчев, Александър Ефремов

Резюме. Текущата глава въвежда систематична методологична рамка за изграждане на синтетични набори от данни, за да се смекчат ограниченията, породени от недостатъчната чувствителна информация при моделирането на данни. Предложената методология включва структурирана последователност от процеси, насочени към разширяване на ограничен начален реален набор от данни чрез интегриране на взаимовръзки, извлечени от бизнес логиката. Разработен е алгоритъм за дефиниране на връзки между променливи чрез прилагане на бизнес ограничения и присвояване на тегла за мащабиране. Синтетичният набор от данни преминава през валидиране чрез серия от тестове, които оценяват качеството и съгласуваността на генерираните данни, включващи принципите на хоризонтална синхронизация и вертикална валидация. Това гарантира разработването на синтетични набори от данни, които са едновременно статистически обосновани и мащабируеми, което ги прави подходящи за приложения във финансови и икономически анализи, както и за използване в науката за данни.

Ключови думи - синтез на данни, разширяване на данни, бизнес логика, мащабиране на тегла, псевдомоделиране

9.1 Въведение

Предвид широко разпространените решения, основани на данни, търсенето на мащабни, хетерогенни и представителни набори от информация се е увеличило в множество индустрии, включително здравеопазване, финанси, изкуствен интелект и машинно обучение.

Настоящата глава се фокусира върху процеса на проучване на възможностите за преодоляване на критичната празнина в наличните набори от данни, съдържащи чувствителна лична информация, поради строги разпоредби като GDPR и законите уреждащи банковата тайна [1]. За да се преодолее тази празнина, изследването се фокусира върху създаването на синтетични многомерни набори от данни, които симулират лични черти, демографски данни и предпочитания за използване във финансите и икономиката.

Предвид сложността и опасенията за поверителност, свързани с реалните данни, синтетичните данни [2], генерирани статистически, а не директно събрани, предлагат жизнеспособна алтернатива. Те позволяват мащабно

обучение и тестване на модели с изкуствен интелект, особено когато реалните данни са оскъдни или скъпи за получаване. Синтетичните данни могат да бъдат пригодени и за балансирано представителство на класовете и специфични изследователски нужди.

От методологична гледна точка, съществува ясна нужда от структурирана рамка за изграждане на многомерни набори от данни. Тази рамка може да се визуализира като методологично дърво [3] с три основни клона: обогатяване на данни, интерполация на данни и синтез на данни. Настоящото изследване се фокусира специално върху клона на синтеза на данни, като се наблюдават следните разклонения (фиг. 9.1-1):



Фиг. 9.1-1 Методологично дърво - синтез на данни

- Когато няма налични данни поради разпоредби за поверителност, като GDPR или закони за банкова тайна, могат да се използват три основни подхода за справяне със ситуацията [2, 4, 14].
- Вероятностното конкатениране се използва за свързване на множество реални набори от данни чрез уникален ключ, базиран на споделени променливи, подходящ, когато наборите от данни имат повтарящи се характеристики [15].
- Анонимизиране на данните. Когато всички данни са налични и включват лична или чувствителна информация, се използват методи за анонимизиране, за да се направи използването на данните, като същевременно се защитава поверителността. Генеративните състезателни мрежи са ефективна техника в този контекст [1].

- Разширяването на данните чрез бизнес правила използва съществуващи данни за създаване на нови функции, прилагайки бизнес логика за обогатяване на набора от данни и подобряване на производителността на модела.

По-конкретно, настоящото изследване се фокусира върху процеса по разширяване (обогатяване) на наличните информационни масиви, като се основава на хипотезата, че синтетичните данни могат да помогнат за преодоляване на пропуските както в количеството, така и в качеството, причинени от правни и практически ограничения.

В ситуации, когато реалните данни не са налични или са недостатъчни, генерирането на нови данни става от съществено значение. Една ефективна техника за това е инженерството на характеристики за разширяване на данните, което използва атрибутите на съществуващите набори от данни и се ръководи от бизнес правила. Този подход, често наричан мащабиране чрез бизнес логика, включва създаване на нови характеристики чрез синтетично разширяване на текущите променливи. Целта е да се обогати наборът от данни, като по този начин се подобри производителността на модела чрез вграждане на специфични знания и подобряване на неговите предсказващи възможности.

Втора част от главата очертава формулировката на задачата, докато трета детайлизира предложената методология за симулация на данни. Част четвърта обсъжда процеса на валидиране, а пета предоставя кратък емпиричен пример. Накрая, заключението е представено в част шест.

9.2 Формулиране на задачата

Това изследване представя структуриран подход за генериране на синтетични набори от данни, базиран на методология, която изисква предварително дефиниране на различни параметри, като фактори, коефициенти и тегла. Тези елементи обикновено се определят с помощта на експерти в областта, за да се осигури съответствие с установената бизнес логика и правила.

Инженерингът на характеристики се извършва посредством присвояване на индивидуални тегла на всеки нов атрибут въз основа на неговата корелация с първоначалните променливи. Този подход, основан на корелацията, позволява генерирането на симулирани променливи, които са съгласувани с оригиналния набор от данни [3].

Даден е входен набор от данни D, който представлява матрица с Nредове и k колони:

$$D \in R^{N \times k}$$

Процесът включва стъпки за дефиниране на вектор от параметри p. Коефициенти на мащабиране (тегла) се присвояват на съответните променливи:

$$p=\left[p_1, p_2, \ldots, p_k\right]$$

където $p_i \geq 0$за i-тата променлива, с $p_i = 0$за неизползвани променливи.

Изчислява се претеглена комбинация x_iот колоните на набора от данни:

$$x_i = D \cdot p^T$$

Конструирането на синтетични характеристики в рамките на описания процес включва методична поредица от стъпки за подготовка и трансформация на данни, насочени към разработване на надежден аналитичен набор от данни. Тези стъпки се основават на първоначалния реален набор от данни и обикновено включват операции като кодиране на данни, дефиниране на категории, мащабиране на характеристики и селективно филтриране. Всяка от описаните процедури се изпълнява следвайки предварително дефиниран набор от бизнес правила - бизнес логика.

9.3 Генериране на данни

Методът има за цел да създаде синтетичен набор от данни чрез прилагане на серия от линейни (с експертно избрани тегла) и нелинейни (бининг) трансформации, нормализиране и след това разделяне на категории, където е необходимо.

Този метод „синтезира“ всяка нова характеристика чрез детерминистична линейна проекция на оригиналните реални характеристики, последвана от проста нормализация и, ако е необходимо, дискретизация, за да се имитират категорични нива (което ги прави нелинейни). Тегловите вектори кодират знания за областта за това кои реални характеристики трябва да управляват всяка синтетична, а праговите параметри гарантират, че полученото разпределение (непрекъснато или категорично) изглежда правдоподобно.

Първоначално трябва да дефинираме вектор на теглото, с по един запис за всяка оригинална колона с данни. Всяко число кодира колко силно трябва да допринесе тази оригинална характеристика. Векторите на теглото трябва да кодират каквато и да е бизнес или домейн логика, която е в основата на вашите синтетични характеристики:

- Експертна преценка: Някой - анализатор или експерт в областта - е решил кои реални входни данни трябва да „имат най-голямо значение“ и с колко, при синтезирането на всяка характеристика.
- Бизнес правила: Конкретната комбинация от тегла (и по същия начин перцентилни прагове) отразява изискванията или конвенциите на съответната бизнес единица (напр. „доходът трябва да определя най-вече нивото на образование, но също така и семейното положение донякъде“).
- Външна настройка: посредством автоматизирана числена процедура.

След това алгоритъмът следва следните стъпки:

1. Пресмята се линеен резултат – изчислява се претеглената сума на реалните входни данни.
2. Нормализират се стойностите [0,1] - Премащабиране в единичен интервал.
3. Праг или повторно скалиране (ако синтетичната характеристика трябва да бъде категорийна). Вектор с малки параметри определя персентилни гранични точки; стойностите под, между или над тези гранични точки се картографират в дискретни категории. Вектор с малки параметри $\left[p^1, p^2, \ldots p^{k-1}\right]$ съдържа персентилни гранични точки в $(0,1)$. След което се напасва всеки резултат i в една от k категориите, като се предприемат стъпки по установяване в кой интервал попада.
4. (По избор) добавя се шум (Гаусов или равномерен) или последваща обработка, въпреки че повечето скриптове се придържат към простия модел линеен → нормализиране → праг.

Хоризонтална синхронизация

Хоризонталната синхронизация се постига чрез вектора на теглото, който определя всяка характеристика поотделно. С други думи:

- Всяка функция получава (от експертни мнения, от бизнес логика или от някаква автоматизирана процедура) свой собствен масив от параметри.
- Все още няма споделен глобален параметър или рутина за съвместно напасване, която да налага съгласуваност между различните синтетични колони.
- Другото нещо, което свързва различните синтезирани характеристики, е, че всички те черпят от една и съща оригинална матрица от данни. Така че, ако две целеви характеристики имат голяма тежест, да речем, на „възраст“ или „доход“, те ще се окажат корелирани само поради използването на едни и същи входни данни - но това е косвен ефект, а не изрична координация.

- Бизнес логиката се използва за коригиране на праговете, така че две категорийни характеристики никога да не си противоречат.
- Генерирането на характеристики е верижно. Глобалната матрица от данни се увеличава с по една колона на всяка итерация. На свой ред, следващите скриптове използват увеличените данни като входяща информация:
- Функция 1 изчислява от оригиналните реални данни с n колони
- Характеристика 2 започва с изчитане на входните данни (сега n+1 колони) и нейният вектор на теглото par има n+1 запис, наречен характеристика 1, така че може да използва току-що създадената характеристика като вход. Генерирането на следващите характеристики следва аналогичен модел на работа.

Този итеративен модел на добавяне на характеристики и повторно използване продължава последователно, така че при по-късните променливи се предполага включването на предварително синтезирани характеристики като част от тяхното линейно оценяване.

Докато по-късните характеристики могат да използват (и надграждат) както по-ранни синтетични променливи, така и оригиналите, глобалната матрица от данни расте, но всяка нова линейна комбинация все още се свързва със суровите входни данни.

Нека $X \in R^{n \times p}$ да е оригинална матрица с реални данни (n наблюдения, p „ сурови “ характеристики).

Дефинира се разширената матрица на стъпка j от

$$A^{(0)} = X, A(j) = \left[A^{(j-1)} \mid f_j \right] \in R^{n \times (p+j)}$$

На всяка стъпка $j = 1, \ldots, m$:

Линеен резултат:

$$s^{(j)} = A^{(j-1)} w^{(j)}, w^{(j)} \in R^{p+(j-1)}$$

е векторът на теглото за характеристиката j.

Нормализиране на стойностите $[0,1]$:

$$\widetilde{s^{(j)}} = \frac{s^{(j)} - \min\limits_i s_i^{(j)}}{\max\limits_i s_i^{(j)} - \min\limits_i s_i^{(j)}}$$

Създава се нова колона f_j, като се прилага (евентуално прекъсната) карта.

$$f_j = \psi_j\left(\widetilde{s^{(j)}}\right)$$

При необходимост от непрекъсната синтетична функция, ψ_j се прилага уравнението:

$$f_j = \widetilde{s^{(j)}}$$

При необходимост от категорийна функция с Kнива, се коригират граничните точки:

$$0 = \tau_0 < \tau_1 < \cdots < \tau_{K-1} < \tau_K = 1$$

и дефинират:

$$\psi_j(u) = k \Longleftrightarrow u \in \left[\tau_{k-1}, \tau_k\right)$$

Задава се:

$$A^{(j)} = \left[A^{(j-1)} \middle| f_j\right]$$

И се преминава към $j+1$. След mстъпките се получава

$$\left[X \middle| f_1 \middle| \cdots \middle| f_m\right]$$

пълен синтетичен набор от данни.

Бележки:

Зависимост между оригинални данни и по-ранни синтезирани набори – При описване f_jкато $w^{(j)T} A^{(j-1)}$, всяка нова функция може да се основава както на суровите, Xтака и на всички преди това генерирани характеристики $f_1, \ldots, f_{j-1}$

Едноредов „разгънат“ изглед - Тъй като всеки f_jв представлява линейно изображение на $A^{(j-1)}$ и $A^{(j-1)}$ е изграден посредством наслагване на линейни изображения, може всяка синтетична характеристика да бъде представена като пряка, но съставна функция X, само на:

$$f_j = \psi_j\left(W^{(j)} X\right), \text{ с } W^{(j)} = w^{(j)}\left[w^{(j-1)T} \vee \ldots \vee w^T\right]$$

Ако бъдат обединени всички заедно, целият синтетичен набор от данни се генерира от оператора:

$$S: X \longmapsto \left(f_1, f_2, \ldots, f_m\right),$$

с рекурсия:

$$f_1 = F_1(X), f2 = F_2\left(X, f_1\right), \ldots, f_m = F_m\left(X, f_1, \ldots, f_{m-1}\right) \quad (17)$$

Тази формулировка ясно показва, че всяка нова характеристика е функция от оригиналните данни X и всяко подмножество от предварително синтезираните характеристики.

9.4 Валидиране

Проверката за надеждност на синтезираните данни се дефинира като (валидиращ) процес. Това се обуславя от недостатъчността на подхода за утвърждаване на новополучените данни посредством набор от критерии за проверка. Съществуват различни подходи за валидиране на синтезирани набори, които се основават на (неизчерпателен списък):

- Правила на бизнес логиката;
- Техники за псевдомоделиране;
- Експертни мнения;
- Автоматизирана статистическа процедура за валидиране и коригиране на данните;
- Агенти базирани на Големи езикови модели (LLM).

Последните три (експертни мнения, автоматизирана статистическа процедура и LLM агенти) са техники, разчитащи на обяснението на по-фундаменталните подходи. Следствие от гореописаното, фокусът в настоящата работа се установява върху валидирането посредством **бизнес логика** и **псевдомоделиране**.

9.4.1 Бизнес логика

Установената рамка дава възможност да бъдат идентифицирани местата, на които линейният процес нарушава зададените бизнес правила. Също така се предоставя ясна последователност, посредством индикатори, проценти на успешно преминаване и прагове за вземане на решения, за да се осигури надежден измервателен процес и валидиране посредством приложение на бизнес логиката.

От гледна точка на настоящото изследване, работното определение на понятието бизнес логика се разбира като: обобщена конструкция, включваща твърдения, правила и допускания, които служат като основни насоки за осигуряване на ефективното внедряване на рамката за синтез на данни.

Въз основа на априорни знания, свързани с разглеждания контекст, наблюдаваните променливи и взаимовръзките между тях, се инициира процес на изграждане на правила по последователен способ, посредством който се извлича набор от логически базирани ограничения, способни да откриват

случаи, в които могат да възникнат невъзможни или много малко вероятни стойности [4].

Неразделна част от тази логическа структура е подреденото взаимодействие между зависимите фактори, техните модели на корелация, свързаните ограничения и както възможните, така и невъзможните комбинации.

Крайната цел на тази процедура за изграждане на бизнес логика е да се предостави възможност за генериране на кохерентно подреден хибриден набор от данни, който обединява синтетични и емпирични данни, като същевременно систематично се премахват нежизнеспособни комбинации.

Основните стъпки, свързани със създаването на бизнес логика, могат да бъдат очертани, както следва:

- Избор на потенциално взаимозависими фактори.
- Установяване на възможни комбинации и такива, които не са.
- Елиминиране на невъзможни комбинации.

Синтезираната бизнес логика служи като обобщена аналитична рамка за разглеждане на цялостния процес на генериране на данни. Прилага се набор от предварително дефинирани правила в съответствие с характеристиките на оригиналния емпиричен набор от данни и набора от налични променливи.

Бизнес логиката представлява кодирано познание за предметната област, което управлява начина, по който обектите се създават, свързват, валидират и развиват с течение на времето, така че получените данни да отразяват точно и закономерно реалния бизнес процес.

Тук знанието за предметната област обхваща богатият ментален модел за това как работи бизнес светът, факти, език, ограничения и особености, преди тези компетентности да бъдат приложени във формална бизнес логика. На теория, добре аргументираната бизнес логика може да бъде автономна по отношение на индивидуалните експертни мнения, след като тези мнения бъдат интегрирани в официален, подкрепен от консенсус набор от правила. На практика описаната независимост не се предоставя автоматично. Настоящата глава приема използването на добре установена, бизнес логика без да се задълбочава в процеса на нейното извличане.

В табл. 9.3-1 всеки ред очертава бизнес правило, което свързва независима характеристика със зависим резултат, под формата на неравенство, използвайки специфични стойности и контекстуална логика, описани в колоната „Забележка“. Като илюстрация на подхода са разгледани две бизнес правила:

- Ако възрастта е под 18 години, банкова сметка не е възможна. Това означава, че лица под 18 години не могат да имат разплащателна сметка.

- Ако професията е във финансовия спектър или администрацията и е на лице предишен инвестиционен опит над 0 години, тогава се наблюдава по-висока склонност към инвестиции.

Идентификатор	Независима функция	Стойност на независимата функция	Зависима функция	Стойност на зависимата характеристика	Забележка
1	Възраст	<18	Банкова сметка	Н	Лица под 18 години не могат да имат разплащателна сметка
2	Професия	Финансови и административни	Предишен инвестиционен опит	>0	По-вероятно е да притежават банкова сметка
3	Ниво на образование	Висше	Доход	>27601 годишно	По-високата степен на образование, предполага по-високи доходи
4	Възраст	<25	Ниво на образование	<Висше	Под 24 години е по-малко вероятно да имат висше образование
5	Възраст	<25	Брой деца под 18 години	<2	От 20 до 24 години е по-малко вероятно да имат повече от 1 дете под 18г.

Таблица 9.4.1-1 - Бизнес логика - пример

9.4.2 Валидиране чрез бизнес логика

За да бъде валидиран синтезиран набор от данни, използвайки правила за бизнес логика, е необходимо всяко правило да бъде представено като условно изявление под формата на неравенство (например табл. 9.3-1) След това всеки синтезиран запис от генерирания информационен набор се проверява дали удовлетворява съответното правило. Така че потенциално има $l \; x \; r$ проверки, колкото е броят на записите, където l представлява броят на записите, а r е броят на правилата.

Проверката на всички записи представлява процес характеризиращ се с необходимост от големи времеви периоди, изискващ множество изчисления и компютърно време. Поради тази причина, вместо пълна проверка, се прилага метод за множество проверки посредством бизнес логика по отношение на голям набор от генерирани записи, като се цели да се достигне до статистически значим брой на проверките. Следващата стъпка е да се установи

процентът на записите, които удовлетворяват правилото и сравняването му с определен праг (например 95%), за успешна валидация на набора от данни.

Важно е да се отбележи, че наборът от данни се създава отляво надясно, след първите няколко колони с реални данни, така че не е необходимо колоните отляво, които са създадени от бизнес логиката, да се тестват спрямо логиката на следващите колони отдясно. Освен това, според методологията, към част от колоните е добавен шум.

За всеки запис $i=1,\ldots,n$ и всяко правило $j=1,\ldots,R$ се дефинира индикатор

$$I_{ij}=\begin{cases}1, if\ i\ satisfies\ j\\ 0, \wedge otherwise\end{cases}$$

За всяко правило j се изчислява емпиричен процент на успешно преминаване

$$P_j=\frac{1}{n}\sum_{i=1}^{n} I_{ij}$$

Установява се валидността на набора от данни съгласно правилото j ако $P_j \geq T_j$ (напр. $T_j=0.95$). Приложимо е и дефинирането на общ процент на успешно преминаване

$$P_{\text{overall}}=\min_{j=1\ldots R} P_j$$

при изискване $P_{\text{overall}} \geq T_{\text{global}}$.

В процеса на работа се обобщават някои проблеми, свързани с валидирането чрез метода на бизнес логиката:

- Не е налично вградено прилагане на ограничения. Тъй като всяка функция е изградена със собствен набор от тегла и гранични точки, няма твърдо правило, което да гарантира:

$$\text{Ако } X_i(\text{Age})<18 \Longrightarrow f_i(\text{BankAccount})=\text{“N”} \quad (21)$$

В най-добрия случай, изборът на $w^{(j)}$ и прагове $\{\tau_k\}$ може да доведе до разпределяне на резултатите в правилните категории, но нищо не налага твърдо правило.

- Независим избор на тегло. Тъй като всяко тегло $w^{(j)}$ е изрично зададено без оглед на зададения набор от правила, различните скриптове могат да изместят синтетичната „Възраст“ на един човек нагоре, докато „Банкова

сметка“ се насочва към противоположната категория - което води до директни нарушения на правилата.

- Добавянето на шум нарушава монотонността. Необходимо е да се отбележи, че „някои колони имат добавен шум“. При последваща обработка на категориен или непрекъснат резултат, е възможно да бъде направено прехвърляне на част от записите през границата на правилото. Това е възможно и посредством потвърждение чрез генератора на данни, в който няма добавен допълнителен шум. В случай че правилата в конвейера без добавен шум не са противопоставящи се.
- Асиметрия отляво надясно. Бизнес логиката често изисква проверка на всички функции спрямо всички правила, но разглежданият конвейер гарантира само, че по-късен скрипт може да види по-ранни синтезирани колони, а не обратното. Тази специфика показва, че правило, което включва променливи като „Доход“ и „Брой деца“, ще бъде нарушено, освен ако не се премине към изрична последваща оценка.

9.4.3 Валидиране чрез псевдомоделиране

Този подход може да настройва векторите на параметрите въз основа на предсказваща точност на съответствие с реална характеристика. За целта се използва една от колоните с реални данни като целева променлива (същата колона не може да се използва при генериране на данни) и всички останали реални и синтезирани данни се прилагат като входни данни в псевдомодел, който проверява различните стойности на параметрите.

Нека $X \in R^{n \times p}$ да бъдат оригиналните реални характеристики.

Избера се една запазена колона $Y = \in R^n$ (или категорийна $Y \in \backslash\{1, \ldots, K \backslash\}^n$) и се премахва от входните данни за синтез.

Генерира се набор от синтетични характеристики $f_1, \ldots, f_m$, като се използва генератор стъпващ върху останалите $p-1$ реални колони.

Предсказващи фактори за стека:

$$Z = \left[X_{\backslash h} \middle| f_1 \middle| \cdots \middle| f_m \right] \in R^{n \times (p-1+m)}$$

Напасване на псевдомодел:

$$\hat{\beta} = \arg\min_{\beta} L(Y, Z\beta) + \lambda \|\beta\|^2$$

Където:

- За непрекъснат вектор Y, $L(Y, Z\beta) = |Y - Z\beta|^2$ Ридж регресия [5, 6].
- За двоични/категорийни системи Y се използва логистична или мултиномна загуба.

Използва се полученият резултат $\hat{\beta}$ като нов параметричен вектор за която и да е синтетична променлива (или се разпределят нейните записи в няколко $w^{(j)}$).

Ридж регресия (Ridge solution) (непрекъснат YY)

$$\hat{\beta} = \left(Z^\top Z + \lambda I\right)^{-1} Z^\top Y$$

Логистична регресия (бинарна $Y \in \{0,1\}$)

$$\hat{\beta} = \arg\min_{\beta} - \sum_{i=1}^{n} \left[Y_i \log \sigma\left(z_i^\top \beta\right) + \left(1 - Y_i\right) \log\left(1 - \sigma\left(z_i^\top \beta\right)\right) \right] + \lambda |\beta|^2 .$$

Кръстосана валидация [9, 10] за избор λ

$$\lambda^{*} = \arg\max_{\lambda} CV \text{ CV - score}\left(Y, Z\hat{\beta}(\lambda)\right)$$

Наблюдават се някои вътрешни логически проблеми на псевдомоделиращия подход:

- Риск от цикличност в случай, че по време на синтеза вече е използвано изключението. Необходимо е избраната реална колона Y да не е включена в нито един от оригиналните тегловни вектори $w^{(j)}$. В противен случай се получава настройва към собствения синтетичен шум на съответната характеристика.
- Мултиколинеарност на синтезираните входни данни [7, 8]. Предсказващите фактори ще включват както суровите X, така и предварително синтезираните $f_1, \ldots, f_m$. Всяка променлива f_j представлява линейна комбинация от X, от това следва, че проектираната матрица е силно колинеарна. В случай на използване на метода на най-малките квадрати ще се получи нестабилен краен резултат (огромна дисперсия в $\hat{\beta}$).
- Пренастройване към една характеристика [11]. Чрез специфично настройване на теглата за прогнозиране на Y, могат да бъдат нарушени пределните или съвместните разпределения на други синтетични характеристики, като по този начин се жертва интерпретируемостта на бизнес логиката по отношение на нейната предсказваща сила.
- Добавянето на шум замърсява сигнала. В случай че някой от сигналите f_j е с добавен шум, псевдомоделът бива частично съобразен със случайността.

Това може да насочи новооценените тегла към обяснение на шума, а не към истинската структура.

- Мащабируемост. В процеса на работа се наблюдава необходимост от отделен цикъл на съпротивление за всяка реална колона, която следва да бъде валидирана, като всеки от тях дава отделен набор от βs. Обратното им съгласуване в един кохерентен набор от параметрични вектори представлява сложен процес.

Като възможно усъвършенстване за справяне с мултиколинеарността и отчасти с пренастройването към една характеристика, може да бъде посочен методът на ортогонализацията [12] на синтетични характеристики X, преди настройването на модела. Ортогонализирането на всяка синтетична характеристика спрямо суровите данни X означава „проектиране“ на всеки компонент от f_j, който се намира в обхвата на X, така че това, което остава, да е напълно некорелирано с всяка колона на X. След това се преминава към настройване на псевдомодела върху:

$$Z=\left[X\middle|\widetilde{f_1},\ldots,\widetilde{f_m}\right]$$

Където всека

$$\widetilde{f_j}=\left(I-P_X\right)f_j, \text{ и } P_X=X\left(X^\top X\right)^{-1}X^\top$$

представлява проекционната матрица върху пространството на колоните на X.

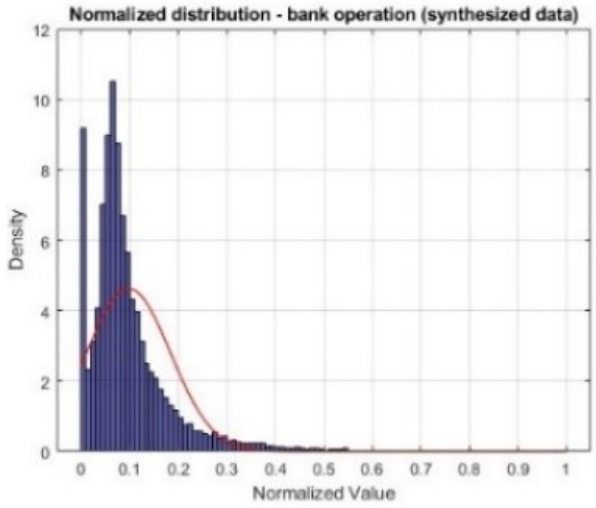
9.5 Емпиричен пример

В рамките на настоящото изследване е проведен кратък симулационен експеримент, следвайки зададения алгоритъм за разширяване на данните. Този експеримент включва генериране на малък набор от синтетични променливи обединени в хипотетичен набор от данни, с цел се илюстрира как работи процесът по разширяване на данните. Примерът е само илюстрация на метода и не разглежда подробно последващия резултат от обработката на синтезираните данни или оригиналния източник на информация. Симулацията в настоящото изследване е проведена на работна станция със следните характеристики: Windows 10 Pro, процесор Intel(R) Xenon(R) W-2145, RAM 64 GB, софтуер - MATLAB R2022b [13], следвайки следната процедура:

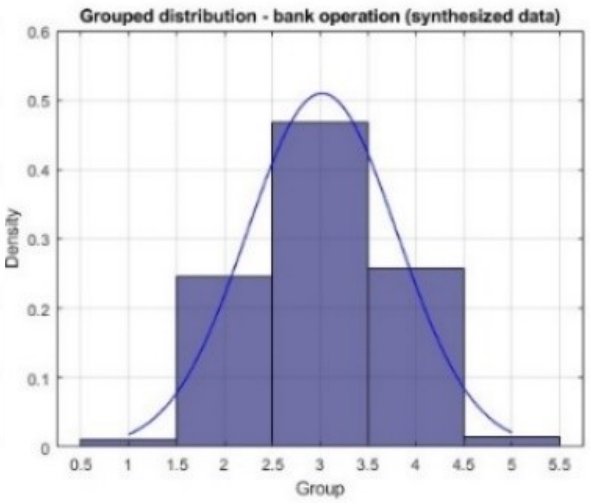
Първата стъпка е да се събере малък набор от реални данни, който включва необходимите основни променливи. Тези променливи могат да се различават в зависимост от данните, необходими за отделните случаи на употреба.

Втората стъпка от процеса представлява присвояване на тегла за мащабиране към всяка променлива, въз основа на зададената бизнес логика, в зависимост от връзката им със синтезираните данни.

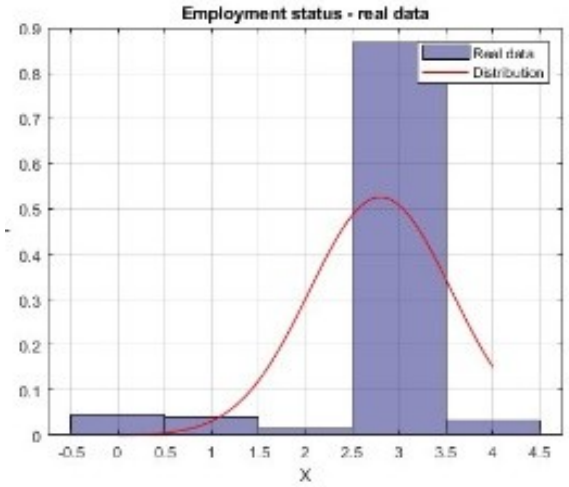
Третата стъпка включва създаването на нормализиран набор от стойности, организирани според специфичната бизнес логика (фиг. 9.5-1 и фиг. 9.5-1).



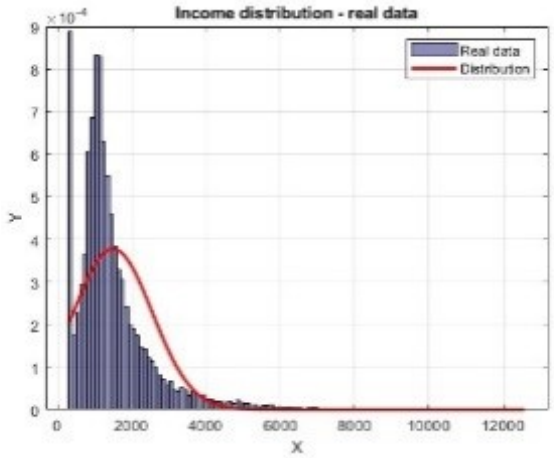
Фиг. 9.5-1 Синтезирани данни – нормализирана стойност



Фиг. 9.5-2 Синтезирани данни – групирана стойност

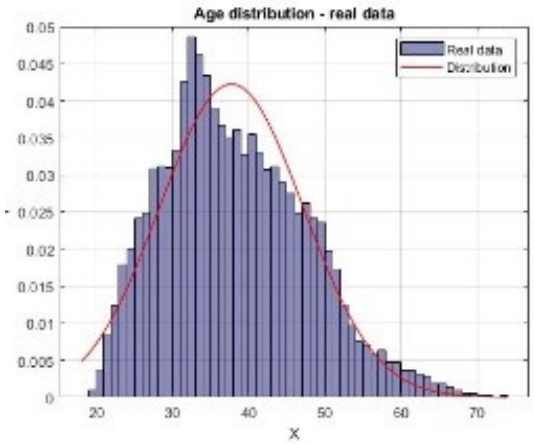


Фиг. 9.5-3 а) Трудов статус – разпределение на реални данни

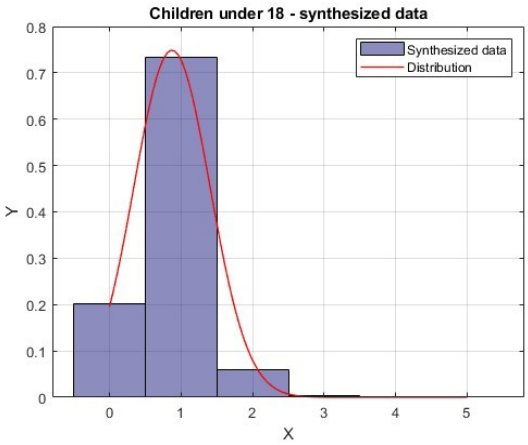


Фиг. 9.5-3 б) Доход – разпределение на реални данни

В **четвъртата стъпка,** посредством инженеринг на характеристики, може да се премине към използване както на оригиналния набор от данни, така и на предварително синтезираните променливи, като основни променливи.



Фиг. 9.5-3 в) Възрастово разпределение – реални данни



Фиг. 9.5-3 г) Синтезирани данни за деца под 18 години въз основа на доход, възраст и образователна степен

Създаването на нова променлива изисква наличието на множество вектори данни, като са необходими минимум три, за да се осигури достатъчно информация за извеждане (фиг. 9.5-3 а, б, в, г) на новополучена информация. В по-сложни сценарии процесът може да включва използването на десетки взаимосвързани вектори, необходими за улавяне на основните модели и зависимости в структурата на данните.

Индекс	sex	age	empl_stat	prof_ind	prof_stat	income per mont	gen_edu	pers_exp	taxes_exp	house_exp	transp_telec_exp	chil_u_18	house_memb	Статус на записа
173	1	39	3	22	2	4,328	6	2	2	2	2	2	4	Успешен
513	2	35	0	34	0	2,055	5	1	1	1	1	1	3	Успешен
793	1	24	3	32	2	7,844	6	3	3	3	3	4	4	Неуспешен
1332	2	23	3	23	2	12,560	6	4	4	4	4	5	7	Неуспешен
1845	2	45	2	14	4	1,150	4	1	1	1	1	1	3	Успешен
20619	1	24	3	32	2	6,475	6	3	3	3	2	3	4	Неуспешен

Таблица 9.5-1 Резултат от симулацията след внедряване на бизнес логиката

След като успешно са конструирани необходимите променливи заедно със съответните им записи, се преминава към прилагане на бизнес логиката. Тя се използва за системно идентифициране и елиминиране на стойности, които са логически неосъществими или статистически много слабо вероятни. Обобщена извадка от прецизираните резултати след прилагането на бизнес логиката може да бъде намерена в таблица 9.5-1, която са отбелязани несъответствията.

9.6 Заключение

Това проучване е провокирано от практическия недостиг на реални данни, които да са съвместими с поверителни информационни набори, причинен от регулации като GDPR и законите за банковата тайна. В отговор на това ограничение, цел е да се създаде повтаряем метод за разширяване на ограничени емпирични данни в големи и реалистични информационни масиви, подходящи за употреба, анализ, моделиране и оповестяване на резултатите.

За постигането на тази цел са изпълнени три взаимосвързани задачи:

- Проектиран е прозрачен алгоритъм, който вгражда експертиза в областта чрез мащабиране на теглата;
- Осигуряване на хоризонтална и вертикална съгласуваност чрез последователен синтез на характеристики и тяхното валидиране, базирано на правила и псевдомодел;
- Демонстриран е цялостния процес на разработка върху реалистичен набор от финансови микроданни.

Алгоритъмът за мащабиране на бизнес логиката трансформира матрица с реални данни в по-голяма и допълнена матрица, чиито синтетични колони са детерминистични проекции, последвани от дискретизация по избор. Те биват параметризирани чрез експертно избрани тегловни вектори и персентилни гранични точки. Чрез внедряване на процес на валидиране се осигурява разумен брой извадкови записи, спазващи наложения набор от правила. Същевременно обратното тестване на псевдомодели възстановява задържаните променливи с приемлива грешка, потвърждавайки, че генерираните данни запазват основната статистическа структура на източника.

Новостта на описания подход се състои в обединяването на три елемента, които не са били комбинирани преди това в единна рамка за синтез. Първо се извършва разширяване на характеристиките отляво надясно, така че всеки нов синтетичен вектор да може да използва повторно по-ранни колони, създавайки контролирани взаимодействия от по-висок порядък, без да се прибягва до генеративни модели тип „черна кутия“. Второ, формализира се хоризонталната синхронизация (имплицитна корелация, индуцирана от споделени входни данни) и бива допълнена с вертикална валидация, която проверява завършените записи спрямо експертни правила. Трето, въвежда се двустепенна процедура за осигуряване, която комбинира проверка на правилата с псевдомоделиране по ридж/логистика, давайки на изследователите бърз количествен сигнал за преминаване/неуспех.

Могат да се обособят няколко предимства. Методът е интерпретируем - всяка синтетична характеристика е претеглена сума от входове; мащабируем - сложността нараства линейно с броя на характеристиките и икономичен от гледна точка на изчислително време - изисква само матрични операции, които работят на обикновен компютър.

В същото време, изследването разкрива три основни недостатъка, които ще бъдат разгледани в бъдещи проучвания:

Строгите бизнес ограничения се насърчават, но не се прилагат математически, така че нарушенията на правилата могат да продължат;

Изборът на тегло в момента зависи от експертни мнения, които могат да бъдат субективни;

Ранното въвеждане на шум в процеса или последователната асиметрия могат да разпространят несъответствия надолу по веригата.

Въпреки тези ограничения, предложената рамка за мащабиране на бизнес логиката постига заявената цел, преодолявайки разликата в недостига на данни със синтетични записи, които са едновременно статистически обосновани и правдоподобни за дадена област. Като последваща работа се обособява автоматизираното настройване на теглата и интеграцията на LLM агенти за усъвършенстване на откриването и валидирането на правила в голям мащаб, като по този начин допълнително ще се разшири приложимостта на рамката в чувствителни области с ограничени данни.

Използвана литература

[1] Marchev, V., Marchev, A. (2025). Anonymizing Personal Information Using Distribution-based Data Synthesis. In: El Khatib, M., Alzoubi, H. M., Angelova, Y. (eds) Sustainable Innovation for Engineering Management (SIEM). ISCME 2024. Advances in Science, Technology & Innovation. Springer, Cham. https://doi.org/10.1007/978-3-031-89889-1_38.

[2] Marchev, A., Jr., & Marchev, V. (2023). Automated algorithm for multi-variate data synthesis with Cholesky decomposition. In *ICACS '23: Proceedings of the 7th International Conference on Algorithms, Computing and Systems* (pp. 54–59). Association for Computing Machinery. https://doi.org/10.1145/3631908.3631909

[3] Marchev, A. A., Jr., & Marchev, V. S. (2026). Methodological findings from generating multidimensional synthetic datasets. *ARPHA Proceedings*, *10*, 301–314. https://doi.org/10.3897/ap.10.e0301

[4] Marchev, V., & Marchev, A., Jr. (2019). Simulation of a multi-criteria database for banking services: Algorithm and business logic. In *New information technologies and big data: Opportunities and perspectives in analyzes and*

management decisions in business, economy, and social sphere (pp. 179–190). UNSS Publishing House.

[5] Hilt, D. E., & Seegrist, D. W. (1977). Ridge, a computer program for calculating ridge regression estimates (General Technical Report NE-28). U. S. Department of Agriculture, Forest Service, Northeastern Forest Experiment Station. https://doi.org/10.5962/bhl.title.68934

[6] Gruber, M. H. J. (1998). Improving efficiency by shrinkage: The James–Stein and ridge regression estimators (pp. 7–15). CRC Press.

[7] Leamer, E. E. (1973). Multicollinearity: A Bayesian interpretation. *The Review of Economics and Statistics, 55*(3), 371–380. https://doi.org/10.2307/1927962

[8] Goldberger, A. S. (1964). Econometric theory. John Wiley & Sons.

[9] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological), 36*(2), 111–133. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x

[10] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2* (pp. 1137–1143). Morgan Kaufmann Publishers Inc.

[11] Claeskens, G., & Hjort, N. L. (2008). Model selection and model averaging. Cambridge University Press.

[12] Löwdin, P. (1970). On the nonorthogonality problem. *Advances in Quantum Chemistry, 5*, 185–199.

[13] MathWorks, Inc. (2024). *MATLAB* (Version R2022b) [Computer software]. https://www.mathworks.com

[14] Marchev, A., Jr., Lybchev, D., & Marchev, V. (2025). Inverse copula sampling for multi-dimensional data synthesis. *AIP Conference Proceedings, 2505*(1), 020012. (In press).

[15] Marchev, A., & Marchev, V. (2022). Synthesizing multi-dimensional personal data sets. *AIP Conference Proceedings, 2505*(1), 020012. https://doi.org/10.1063/5.0100615

10 Валидиране на синтетични данни за поведенческо моделиране

Автори: Ангел Марчев, мл., Васил Марчев, Александър Ефремов

Резюме: Следващата глава цели да бъде въведена унифицирана рамка за валидиране на синтетични набори от данни, използвани в процеса на поведенческо моделиране при ограничения за поверителност и студен старт, като се разграничава статистическата правдоподобност от бизнес допустимостта. Формализират се три независими валидатора: (1) Одит за съгласуваност от експерти (HECA), който дава рекордни проценти на преминаване и надеждност между оценки за решения „продължава/не продължава“, съобразени със съответните политики; (2) Програмен одит, базиран на големи езикови модели LLM (LLM-PA), който мащабира проверката на правилата посредством структурирани за целта агенти, калибрирани към малък набор от експертни правила; и (3) Автоматизирано калибриране „Leave-One-Feature-Out“ (LOFO-AC), итеративен цикъл за самовъзстановяване, който намалява грешката при реконструкция на задържане на записите, използвайки регуляризирани регресии и актуализации на теглото в стил Якоби. Извършено е сравнение на методите по отношение на управление, мащабируемост и техническа прецизност: HECA дава приоритет на отчетното съответствие, LLM-PA позволява високопроизводително сортиране, когато моделите са структурирани и са приложени контроли за поверителност, а LOFO-AC подобрява вътрешната съгласуваност на генератора, без да се предполага външен реализъм. Изследването посочва критериите за приемане за всеки генератор и предоставя някои практически насоки относно правилата за вземане на извадки, инструкции (промпт), регуларизация и спиране. Цялостната рамка е демонстрирана с илюстративни примери. Показва как всеки валидатор улавя различни рискове и предоставя структурирани насоки за избор на метод, калибриране и справедливост.

Ключови думи и фрази: валидиране на синтетични данни; поведенческо моделиране; експертен одит; одит, базиран на LLM

10.1 Въведение

Настоящата глава представлява продължение на предходната [16], в която е очертан подход за синтез на данни чрез иновативна евристика - разширяване на данните. Фокусът на настоящата точка е върху въпросите, които повдига процеса на валидиране, като част от по-големия процес за синтез на данни.

Съществуващите изследвания върху качеството на синтетичните данни се фокусират върху статистически показатели или върху процеса по генериране. Проучванията отбелязват, че оценката често се основава на прости сравнения на разпределението (напр. съвпадение на средни стойности, дисперсии или други статистически разстояния). Някои автори [18] описват основните статистически мерки (средна стойност, медиана, дисперсия и др.) и тяхното използване за оценка на качеството на синтетичните данни. Други автори [19] отбелязват, че повечето съществуващи оценки на синтетични данни се фокусират върху „статистическата прецизност“ на генерираните данни. Други изследвания се фокусират предимно върху разработването на методи за генериране (напр. GAN, VAE). В [20] се предоставя систематичен преглед на техниките за генериране на таблични синтетични данни.

За разлика от това, в настоящото проучване се разглежда възможността за директно включване на набор от правила и експертна/LLM преценка в процеса на валидиране. Също така е добавен нов итеративен цикъл за калибриране (LOFO-AC), който не е разглеждан в предишни изследвания по темата. С цел да бъдат запълнени описаните пропуски, настоящата статия се стреми към достигане на следните приноси:

- Формулиране на три валидатора на синтетични данни (HECA, LLM-PA, LOFO-AC) с ясни критерии за вземане на решения (процент на успешно преминаване, k прагове, намаляване на грешките).
- Изясняване на ролята на всеки метод (напр. HECA за съответствие, LLM-PA за високопроизводителни проверки, LOFO-AC за статистическа настройка).
- Практическо ръководство за внедряване (стратегии за вземане на проби, бързо проектиране, настройване на модела, правила за спиране и оразмеряване на извадката).
- Илюстративни примери върху конкретни примери, демонстриращи резултата и взаимодействието на всеки метод.

10.2 Концептуален преглед

10.2.1 Поведенчески данни

Поведенческите данни се отнасят до информация за потребителите и взаимодействието им с даден продукт или услуга. Те улавят техните действия, предпочитания и модели на употреба. Демографската информация е полезна за стратегии за сегментиране на потребителите, но не предлага същата дълбочина за анализ както поведенческите данни. Поведенческите данни предоставят по-задълбочена и по-персонализирана информация, за това как даден човек взаимодейства с определен продукт, предоставяйки перспектива, която е по-

подходяща за подобряване на монетизацията [4]. Те предоставят информация за поведението на потребителите, като съдействат за създаването на цялостен и динамичен поглед върху предпочитанията на потребителите, подчертавайки както контекста на техните действия, така и честотата на тяхното използване на съответните услуги. Разбирането на поведенческите данни изисква не само описание на самото поведение, но и отчитане на контекста, в който то се случва. Чрез анализ на подробни поведенчески записи, компаниите могат да получат ценна информация за предпочитанията на клиентите и да адаптират своите стратегии, с цел по-адекватно отразяване на клиентските необходимости [5].

Поведенческите данни, събрани от потребителите на конкретен продукт включват сложен и времеемък процес за събиране на необходимата информация. Този процес включва проследяване на потребителски файлове за влизане в системата, които често съдържат чувствителна информация и изискват допълнителен лицензионен режим. Това предполага използването на различни методологии за генериране на необходимите данни в количествен и качествен аспект.

Чрез анализ на тези набори от данни могат да се идентифицират модели, които може да не са очевидни, когато се разглеждат само демографски данни. Това по-задълбочено разбиране на потребителското поведение позволява вземане на решения, основани на данни, в различни области.

Съществува широк спектър от приложения за поведенчески данни:

Персонализирането на потребителските модели е от съществено значение за разработването на индивидуализирани пакети със съдържание, препоръки и оферти, които са съобразени с предпочитанията и историческото поведение на всеки потребител. Този подход е подходящ в различни сектори, включително електронна търговия, финанси и медицински услуги. Чрез използването на многомерни поведенчески данни, тези модели могат да бъдат непрекъснато актуализирани и усъвършенствани, за да се адаптират към променящите се интереси на всеки потребител. Това ниво на персонализация води до повишена ангажираност и удовлетвореност сред клиентите на услугата [6].

Оптимизация на продукта. Поведенческите данни идентифицират области, в които потребителите изпитват затруднения или им липсват структурирани насоки. В резултат на идентифицирането на такива области се правят целенасочени подобрения в прилаганата методология. Такива подобрения, основани на данни, повишават използваемостта и насърчават по-дълбокото ангажиране на потребителите.

Маркетингова ефективност. Използването на поведенческа информация служи за създаване на по-целенасочени продукти и дейности. Чрез

поведенчески анализ се прави подробен напречен разрез на предпочитанията на клиентите и въздействието, което различните персонализирани услуги или продукти оказват върху тях. Тези информационни масиви могат да се използват за прецизно и персонализирано таргетиране на потребителите.

Сегментиране на клиентите. Данните за поведението позволяват идентифицирането на по-смислени клиентски сегменти въз основа на техните действия и предпочитания, отколкото да се разчита единствено на основни демографски данни. Чрез сегментиране на потребителите според тяхното поведение, компаниите могат да персонализират комуникацията и услугите си за всяка група. Този подход гарантира, че стратегии като кампании за задържане на клиенти, или кръстосани продажби, са съобразени с характеристиките и нуждите на всеки сегмент.

Прогнозната аналитика включва анализ на минало поведение, за да се прогнозират бъдещи тенденции и проактивно да се отговори на нуждите на потребителите. Чрез изследване на историческите модели на употреба, бизнесите могат да разработят модели, които предвиждат резултати, свързани с вероятното напускане на потребители на съответния пазар или повишаване на търсенето на определени продуктови линии. Тези прогнозни модели, базирани на поведенчески данни, дават възможност на организациите да вземат информирани решения, да намалят рисковете и да се възползват от възможностите, свързани с очакваното бъдещо поведение.

С развитието на техниките за анализ на данни, синтетичните масиви и поведенческият анализ се превръщат във все по-ценни инструменти. Синтетичните данни предоставят безопасен и гъвкав начин за разработване и тестване на модели при строги режими на поверителност, докато поведенческите данни предлагат задълбочен поглед върху предпочитанията на потребителите и моделите на употреба. Заедно тези подходи позволяват на изследователите да изграждат точни, персонализирани и съвместими решения [7].

10.2.2 Студен старт и синтетични данни

Синтетичните информационни структури се генерират посредством математическа апроксимация на статистическите зависимости, целящи репликация на емпиричните разпределения при пълна анонимизация. Синтетичните записи се създават с помощта на усъвършенствани алгоритмични архитектури, които улавят и запазват корелационни зависимости, наблюдавани в първичните източници на информация. Като резултат от това се създават предпоставки за прецизно тестване и обучение на

модели. Приложението на синтетични масиви от данни приложимо при недостатъчност на извадката или наличие на конфиденциални параметри. Синтезираните посредством алгоритми данни могат да се използват за подобряване на наличните информационните масиви, както и за провеждане на симулационни сценарии, чрез които се осъществява анализ или за разработване на модели [1].

Студеният старт описва трудностите, които се срещат при наличие на оскъдна или изцяло липсваща историческа информация за разработване на модел в нова бизнес среда. Тази ситуация обикновено възниква, когато една компания навлиза на непознат пазар или въвежда нов продукт, което води до недостатъчно количество данни за изграждане на надежден модел.

За да се преодолее студеният старт и предизвикателствата свързани с него, както и ефективната работа в нова среда, от изключително значение е да се осигури налична подробна информация за клиентите и пазара. Събирането на изчерпателни данни за потенциалните клиенти, включително техните предпочитания, поведение и демографски профили, е от съществена важност. Освен това, получаването на информация за пазарната среда, конкурентите, развитието на индустрията и други контекстуални фактори е необходимо за разработването на надеждни модели и вземането на информирани бизнес решения. Традиционно изграждането на модели за подобен тип ситуации изисква период на изчакване, докато се съберат достатъчно реални данни. От своя страна, синтетичните информационни масиви предлагат алтернативен подход към събирането и моделирането на данни в сценарии на студен старт. Чрез придобитите масиви от синтетична информация могат да се създадат изкуствени потребителски профили, атрибути на артикули или записи на взаимодействия, като по този начин се генерира необходимото количество и качество на информацията, която е потребна за изграждането на модели.

Важно е да се отбележи, че качеството на моделите, разработени с помощта на синтетични данни, зависи до голяма степен от разумните предположения, направени за целевата област. Тези модели могат да бъдат подобрени с течение на времето, когато реалните данни станат достъпни, като постепенно заместват синтетичните данни. Следователно, създаването на синтетични набори от данни представлява целесъобразна стратегия за преодоляване на първоначалния недостиг на информация.

Последователното заместване на синтетични данни с реални наблюдения позволява на модела да улавя по-точно автентичната динамика на клиентите и съответния пазар. Тази поетапна промяна подпомага откриването на специфични за областта закономерности, което от своя страна засилва точността и надеждността на модела [3].

Предсказващият модел може да бъде обозначен като: $g(X)$, с X се обозначават данните, а с $\hat{Y}$- прогнозата [3]:

$$\hat{Y} = g(X)$$

При наличие на реални данни (напр. j наблюдения), те подлежат на постепенно включват в модела. Това може да бъде представено като претеглена комбинация от синтетични и реални данни [3]:

$$X = \begin{cases} D_{real} \\ D_{synthetic}, n > j \end{cases}$$

Като съществен момент в процеса трябва да се отбележи, че поетапното въвеждане на реални данни създава възможност за итеративно актуализиране и усъвършенстване на модела. Вниманието към подходящото време и скорост на интегриране е от съществено значение. Чрез постепенното включване на реални данни, преходът остава структуриран и контролиран, като по този начин се избягват внезапни промени в поведението и се улеснява плавния процес на адаптация.

10.2.3 Значение на валидацията на синтетични данни

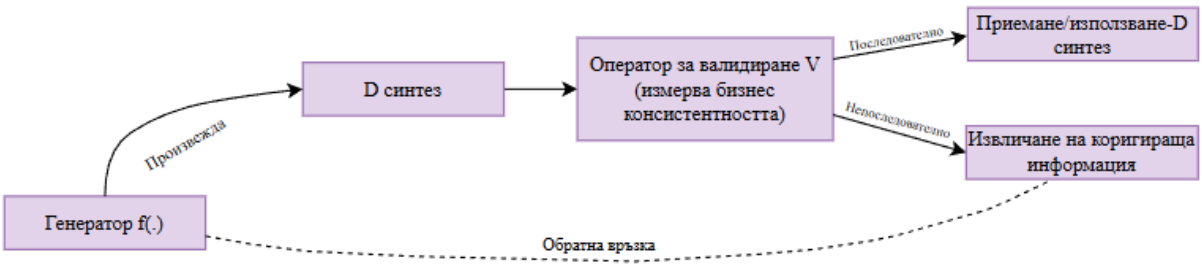
Синтетичните данни представляват практичен инструмент за анализ с реда характеризираща се с ограничения, свързани с поверителност, недостиг на данни и при студен старт. От казаното по-горе се установява, че генераторът $f(\cdot)$ разпределя реални информационни масиви D_{real} (или добре обосновани априорни прогнози) в синтетичен набор от данни $D_{\text{synthetic}} = f(D_{\text{real}})$, който запазва разпределенията, важните асоциации и връзките между данните с достатъчно голяма точност, за да позволи моделиране, тестване и подкрепа на решения. Този аргумент е убедителен само ако синтетичните данни са подходящи за бизнес употреба - т.е. те имат поведение, което се доближава в максимална степен до поведението на реалните данни, когато са изправени пред оперативни правила, експертен контрол и модели надолу по веригата.

Последните изследвания разглеждат валидирането на синтетични данни, използвайки показатели за сходство в разпределението, компромиси между поверителност и полезност и състезателно тестване. Santangelo et al. [21] въвеждат рамка, оценяваща синтетични таблични данни по показатели за „сходство“ (разпределение), за полезност и такива разглеждащи поверителността. Разглежданата рамка включва анализи в състезателен стил:

изводи за принадлежност (MIA) и изводи за атрибути (AIA) върху синтетични данни, като задачата е да се измери изтичането на поверителност. Резултатите на Yan и колектив [22] подчертават ясен компромис между поверителност и полезност в синтетичните данни от електронните здравни досиета: подобряването на полезността често влошава поверителността. Sarmin и колектив [23] преразглеждат компромиса между поверителност и полезност и показват, че на практика синтетичните данни могат да постигнат по-благоприятен компромис от традиционната анонимизация.

За разлика от посочените трудове, настоящото изследване се фокусира върху валидирането, основано на специфични за дадена област правила и процеси на одит. В следващите части от настоящата глава се разглежда въпроса: как да бъдат валидирани синтетични данни за изработване на бизнес решения?

Този принос е пряко продължение на по-ранен труд [16], в който механизмът на синтез е специфициран като поредица от линейни проекции и дискретизации, ръководени от априорни знания за предметната област. В тази по-ранна работа централното предложение е, че внимателно изработените тегловни вектори и точки на разрез могат да кодират бизнес логика и по този начин да генерират правдоподобни многомерни записи. В настоящото проучване се разширява цялостната програмата: от стъпките на конструиране към фазите на осигуряване, като са предложени, формализирани и операционализирани три допълващи се подхода за валидиране.



Фиг. 10.2.3-1: Процес на валидиране на синтетични данни

10.3 Предложени методи

Тази част предлага практически, стъпка по стъпка насоки за прилагане на рамката за валидиране в реални условия.

10.3.1 Одит на съгласуваност между експерти (HECA)

Одит на съгласуваност между експерти (HECA) - формализира експертната проверка в измерими показатели - проценти на успешно преминаване на ниво запис, като се разглежда надеждността между оценители

(κ). HECA формализира това, което експертите вече правят: количествено се остойностяват експертните оценки както с процент на успешно преминаване, така и между оценители κ. Това дава прозрачен показател за съответствие, основан на правилата на политиката. HECA отговаря на въпроса: „Ще приеме ли информиран експерт тези записи като вътрешно съгласувани, в съответствие с установените политики?“. Разглежданият метод фиксира процесът на валидация в предварително установените норми на организацията и създава следа от решения и обосновки за евентуален неуспех, която подлежи на последващ одит. Докато експертните одити са често срещани, новостта на HECA се състои в структурирането на прегледа, базиран на правила, с количествени проценти на успешно преминаване и между оценители κ в контекста на синтетичните данни.

Цел

Използват се преценки на експерти в областта, за да се определи дали отделните синтетични записи се подчиняват на фиксиран набор от бизнес правила и ограничения, основани на здравия разум и априорна информация. След това се преминава към количествени определяне както на процента на успешно преминаване, така и надеждността между оценителите.

Процедура

1) Нека $D=\{x_i\}_{i=1}^{n}$ е синтетичният набор от данни и $R=\{r_1,\ldots,r_R\}$ е набор от правила (всяко правило е условно неравенство). Начертава се извадка за преглед $S \subseteq D$ с размер s.
2) Всеки експерт $k=1,\ldots,K$ етикетира записа $i \in S$ като последователен или непоследователен $C_{i,k} \in \{0,1\}$:.
3) Обобщава се с мнозинство от гласовете

$$C_i=I\left(\sum_{k=1}^{K} C_{i,k} \geq \lceil K/2 \rceil\right), P_{\text{HECA}}=\frac{1}{s}\sum_{i \in S} C_i$$

Тъй като преценките са субективни, необходимо е да се разгледа надеждността между оценители. За двама оценители се използва метода на Коен κ [8]; за $K>2$ се използва метода на Флайс κ [9]. И двата метода коригират наблюдаваното съгласие за случайност и са стандартни в поведенческите и социалните измервания. Експертите трябва да маркират кое правило не е наред, така че нарушенията да се съпоставят със специфични характеристики. Това позволява цикъл на възстановяване и разширява този метод в LOFO-AC (вижте по-долу).

Критерий за приемане на набора от данни: $P_{\text{НЕСА}} \geq T_{\text{Н}}$ и $\kappa \geq \kappa_{min}$ с типични прагове $T_{\text{Н}} \in [0.90, 0.99]$ и $\kappa_{min} \in [0.6, 0.8]$ в зависимост от критичността на приложението.

10.3.2 Програмен одит, базиран на Големи езикови модели (LLM-PA)

Програмният одит, базиран на Големи езикови модели (LLM-PA), мащабира разглеждания контрол. Чрез набор от запитвания (промптове) приведени към големи езикови модели, с точния набор от правила и всеки запис, се получават бързи, възпроизводими преценки и структурирани обосновки. LLM-PA е надграждане спрямо НЕСА, като частично замества експертите с агенти като подчертава противоречиви модели в голям мащаб. От съществено значение е необходимостта от калибриране на LLM-PA спрямо малък, зададен от експерт набор с правила, за да бъде осигурена съгласуваност преди групово приложение. LLM-PA адаптира техниките „LLM-as-judge“ към валидиране на бизнес правила, като проектира прецизни, базирани на правила, подкани и калибрира към зададено от експерт подмножество.

Последните публикации формализират използването на LLM като автоматизирани оценители (т. нар. „LLM-as-a-Judge“). Zheng и колектив [13] въвеждат MT-Bench, използвайки GPT-4 като „съдия“ на LLM резултатите; те съобщават, че преценките на GPT-4 съвпадат с човешките предпочитания в над 80% от случаите. Прадхан и колектив [24] изрично определят GPT-4 като LLM-as-judge в RAG задача в правната област, като формулират въпроса „Можем ли да се доверим на LLM да служи като надежден съдия?“.

LLM-PA използва тези скорошни постижения (напр. LLM-as-judge), като преобразува изрични бизнес правила в подкани. Новост е калибрирането на LLM спрямо малък експертен - „златен“ набор, за да се приведат машинните преценки в съответствие с експертните норми. LLM-PA връща структуриран JSON (pass/fail + failed_rules) за всеки запис, което прави резултатите му директно сравними с човешките етикети на НЕСА.

Цел

Използват се един или повече големи езикови модели (LLM) като мащабируеми „съдии“, за да се оцени съгласуваността на правилата за всеки запис, с подкани, които изрично включват R. Последните изследвания показват, че LLM-като съдия може да се доближи до човешките преценки, когато е внимателно проектиран и калибриран [13;14].

Процедура

1) За извадка S и набор от агенти $A=\{a_1,\ldots,a_K\}$ (модели и/или варианти на заявки), се прави заявка към всеки агент със запис x_i и правила R, анализирайки двоична команда с преценка $L_{i,a}\in\{0,1\}$ и се агрегира както следва:

2) $C_i=I\left(\sum_{a\in A} L_{i,a}\geq\kappa\right), P_{\text{LLM}}=\frac{1}{|S|}\sum_{i\in S} C_i$

3) По избор, събират се индикаторите за всяко правило $L_{i,a}^{(j)}$, с цел да се изчислят процентите на успешно преминаване на ниво правило $P_{\text{LLM}}^{(j)}$.

Калибриране и предпазни мерки

- **Заземяване:** необходимо е да се включи точния текст на правилото в заявката, като се изискват структурирани JSON изходи.
- **Обединяване и консенсус:** използване $\kappa=\lceil K/2\rceil$ или меко гласуване за смекчаване на пристрастията, свързани с един агент [14].
- **Набор за проверка от експерти:** измерване на съответствието между LLM и хора върху малка извадка от експертни оценки, използвайки κ [13].

Критерий за приемане на набора от данни съгласно LLM-PA

Приема се набора от данни, ако $P_{\text{LLM}}\geq T_{\text{L}}$ и LLM-човешкото съгласие κ за калибровъчния набор надвишава κ_{min} [13].

10.3.3 Автоматизирано калибриране с пропускане на една променлива

Автоматизираното калибриране с пропускане на една характеристика (Leave-One-Feature-Out Automated Calibration - LOFO-AC) превръща валидационните сигнали в приложими корекции. Третирайки всяка оригинална характеристика x_j като временна цел, се преминава към напасване посредством останалите реални характеристики и текущия синтетичен блок, след което следва синхронно актуализиране на теглата на синтеза (стил на Якоби [17], по избор със заглушаване). Това запазва духа на мащабирането на бизнес логиката, като същевременно систематично намалява грешките при реконструкцията и нарушенията на правилата - без да превръща системата в универсална, самообучаваща се невронна мрежа. Използват се техники като настройка по Ридж еластична мрежа и по избор ортогонализация на синтетични характеристики спрямо X директно акцентиращи върху колинеарността, известен страничен продукт на линейно конструирани, синтезирани променливи. Забележка: Категорийни или двоични характеристики - могат да бъдат преобразувани във фиктивни променливи или да се използват подходящи модели (напр. логистична регресия) вместо ридж/PLS. Това гарантира, че всички типове характеристики могат да бъдат обхванати.

Цел

Всяка оригинална характеристика x_j се третира като временна цел, съчетава се с останалите реални характеристики и текущия синтетичен блок, използват се съчетаните коефициенти като актуализирани тегла и се преминава към итериране, докато се стабилизира. Това е двуетапен, подобен на кръстосана валидация цикъл на валидиране плюс калибриране, а не невронна мрежа.

Нотация

$X=[x_1,\ldots,x_p]\in R^{n\times p}$: реални характеристики.

$F(X;W)=[f_1,\ldots,f_m]$: синтетични елементи, конструирани със зададено тегло W.

За целта x_j, предиктори $Z_j=[X_j \mid F(X;W)]$.

Базово съответствие

Тъй като F е изграден от X, Z_j са колинеарни. Използват се регуляризирани или компонентно-базирани модели:

Ридж (ridge) (трябва да е стабилен при мултиколинеарност) [10]:

$$\widehat{\beta}_j=arg\min_{\beta}|x_j-Z_j\beta|_2^2+\lambda|\beta|_2^2$$

Еластична мрежа (използва се за разредени и стабилни случаи) [11]:

$$\widehat{\beta}_j=arg\min_{\beta}|x_j-Z_j\beta|_2^2+\lambda_1|\beta|_1+\lambda_2|\beta|_2^2$$

PLS регресия (проектиране към компоненти, максимизиращи ковариацията с x_j) оценява се малък брой от латентни фактори от Z_j и регресия x_j върху тях [12].

Ортогонализиране на синтетични характеристики

Преди напасването се осъществява измерване на остатъците за всяка синтетична колона спрямо разстоянието от X:

$\widetilde{f}_l=f_l-X(X^\top X)^{-1}X^\top f_l, l=1,\ldots,m$ като се използва $Z_j=[X_{-j} \mid \widetilde{F}]$. Това премахва колинеарността на сурови синтезирани записи посредством конструкция и стабилизира оценките.

Актуализация на Якоби

Нека β_i да са настроените коефициенти от регресията на характеристика от реални данни x_i върху останалите реални характеристики, като към тях се прибавят текущите синтетични характеристики. След това се преминава към актуализиране на синтетичното тегло W_iс $W_i\leftarrow W_i+\alpha.\beta_i$, където $\alpha\in(0;1]$ е коефициент на затихване. Операторът за картографиране Ф взема настроените коефициенти β_i за всяка целева характеристика x_i и актуализира теглата на

синтеза. Тази актуализация на вектора се прилага синхронно за всички характеристики (стил Якоби).

Използването на Якоби (синхронизиран), вместо Гаус-Зайдел (подвижен), премахва обратната връзка основана на вътрешна итерация; затихването $\gamma<1$ допълнително потиска осцилациите, известни от класическите итеративни методи [15]. Препоръчва се спиране на LOFO итерациите, когато подобрението в агрегираната грешка падне под малък праг (напр. <1%) или след фиксиран брой итерации (напр. 5-10), което от двете настъпи първо. Това осигурява изчислителна проследимост и избягва пренастройване.

Критерий за приемане на набора от данни съгласно LOFO-AC

$$P_{\text{LOFO}}=1-\frac{1}{p}\sum_{j=1}^{p}\frac{L\left(x_j,\widehat{x}_j\right)}{L_0\left(x_j\right)}\geq T_{\text{C}}$$

където L_0 е загубата на наивна базова стойност (напр. само средна стойност).

Схема на итерация

Синтез на характеристики: $F^{(t)}=F\left(X;W^{(t)}\right)$

Напасват се всички j паралелно: $\widehat{\beta_j^{(t)}}=arg\min_{\beta}L\left(x_j,Z_j^{(t)}\beta\right),Z_j^{(t)}=\left[X_{-j}F^{(t)}\right]$

Актуализация на Якоби: $W^{(t+1)}=\Phi\left(\widehat{\beta_1^{(t)}},\ldots,\widehat{\beta_p^{(t)}}\right)$

Недостатъчно отпускане: $W^{(t+1)}\leftarrow\left(1-\gamma\right)W^{(t)}+\gamma W^{(t+1)},\gamma\in\left(0,1\right]$

Правило за спиране $\begin{cases}\sum_{j=1}^{p}L\left(x_j,\widehat{x}_j\right)\\|W^{(t+1)}-W^{(t)}|_F<\varepsilon\end{cases}$

10.4 Емпирични резултати

Емпиричното изследване описва подробно оценката на симулиран набор от данни и резултатите от трите проучвания: Одит за съответствие от експерт (HECA), Одит на политиката за големи езикови модели (LLM-PA) и Одит с пропускане на една функция (LOFO-AC). Симулираният набор от 100 реда е структуриран умишлено да съдържа 60 съответстващи записа и 40 записа, нарушаващи три специфични бизнес правила, свързани с възраст, нисък доход и висок брой активни заеми. Ключовите „скрити“ характеристики на данните включват много силна корелация между дохода и кредитния лимит, която установява „ефект на богатството“ в съответстващите записи, както и силно

изкривени Active_Loans и ниска минимална възраст, която може да бъде маскирана от средната възраст.

Проучването HECA, в което участват двама оценители (експерти), довежда до нисък показател за процент на успешно преминаване, но висока надеждност между оценители, което показва, че генераторът на синтетични данни не е успял да генерира съответстващи данни, но одиторите до голяма степен са се съгласили с оценките си. За разлика от това, проучването LLM-PA демонстрира значителни предимства по отношение на мащабируемостта пред HECA, с линейно мащабиране и цена под 0,12 за 10 000 записа, което далеч надвишава осъществимостта на човешкия одит от гледна точка на времето и разходите. Проучването LOFO-AC се фокусира върху корекцията на данните, като успешно увеличава крайния показател LOFO от 0,66 на 0,90 (статус „Преминава“) и намалява отклонението с етикет „Статистически невероятно“, със 71%.

10.4.1 Симулиран набор от данни за апробация

Състав на набора от данни (100 реда):

- Редове 1-60 (Съвместими): Моделирани по правилна бизнес логика (напр. По-висок доход ~ По-висока граница).
- Редове 61-75 (Правило за неуспех 1): Непълнолетни (age < 18) с положителен кредитен лимит.
- Редове 76-90 (Правило 2 за неуспех): Лица с ниски доходи (inocome < 20000) с опасно високи лимити.
- Редове 91-100 (Правило за неуспех 3): Потребители с прекомерно задлъжнялост (active_loans > 5) с високи лимити.

„Скрити“ характеристики на набора от данни

- **Двигателят „Доход“**: Корелацията между доход и кредитен лимит е изключително висока (0,95). Това потвърждава, че за „Съвместимите“ 60 записа съществува „Ефект на богатството“ (по-богатите потребители получават по-високи лимити).
- **Отклоненията в „Заем“**: Характеристиката Active_Loans е силно асиметрична (1,72). Средната стойност е ниска (1,8), но максималната е 9. Това са 10-те записа, предназначени да не отговарят на правило R3, което представлява малък, но опасен „риск от опашка“ в данните.
- **„Скритите“ непълнолетни**: Минималната възраст е 14 години, което доказва, че наборът от данни съдържа непълнолетни. Средната възраст (34) обаче изглежда като добър показател в това отношение, поради което една проста проверка за „средна“ възраст не би могла да открие нарушението.

Преглед на набора от данни ($N = 100$**)**

В таблица 10.4.1-1 са обобщени статистическите свойства на набора от данни, използван за симулациите HECA, LLM-PA и LOFO-AC.

Статистика	**Възраст (години)**	**Доход (годишен)**	**Активни заеми (брой)**	**Кредитен лимит**
Брой	100	100	100	100
Средна стойност	34.3	43 108 долара	1.8	6 224 долара
Стандартно отклонение	12.8	29 728 долара	2.2	4 634 долара
Минимум	14 *(Малък)*	$0 *(без доход)*	0	200 долара
25% (първо тримесечие)	24	17 875 долара	0	2500 долара
Медиана	33	43 500 долара	1.0	4 750 долара
75% (трето тримесечие)	45	65 000 долара	2.0	10 000 долара
Максимално	62	110 000 долара	9 *(Свръхзадлъжнял)*	16 500 долара
Асиметрия	0,31 (Нормално)	0,22 (Нормално)	1,72 (Изкривено надясно)	0,56 (Умерено)
Корелация (с ограничение)	0.83 (Силно)	0,95 (Много силна)	0,64 (Умерено)	1.00

Таблица 10.4.1-1 Статистически свойства на използвания набор от данни

10.4.2 Проучване на HECA

Идентификационен номер на записа	**Ценности**	**Оценка на оценител А**	**Оценка на оценител Б**	**Причина за несъгласие**
059	50 години, Заеми 3 (Пропуск)	Успех	Неуспех	Грешно положителен резултат: Оценител Б смята, че 3 заема са твърде рисковани за този конкретен профил, въпреки че технически това е разрешено (<5).
061	Възраст 17, Лимит 1000 (Неуспешен R1)	Успех	Неуспех	Човешка грешка**:** Оценител А е пропуснал реда и е пропуснал нарушението на възрастта (умора).
076	Доход 18 хиляди, Лимит 3 хиляди (Неуспешен R2)	Неуспех	Успех	Тълкуване: Оценител Б е счел 18k за „достатъчно близо“ до прага от 20k, за да позволи ограничение от 3k.
091	Доход 60 хиляди, Заеми 8	Неуспех	Успех	Пристрастие: Оценител Б е допуснал високия брой заеми,

Идентификационен номер на записа	Ценности	Оценка на оценител А	Оценка на оценител Б	Причина за несъгласие
	(Неуспешен R3)			тъй като доходът е бил висок (60 000), игнорирайки строгото правило.
100	Доход 70 000, Заеми 9 (Неуспешен R3)	Неуспех	Успех	Пристрастие: Същото както по-горе; Оценител Б е дал приоритет на дохода пред ограничението на броя на заемите.

Таблица 10.4.2-1 Дневник на несъгласията („човешкият“ елемент)

	Оценител Б: Преминал (1)	Оценител Б: Неуспешен (0)	Общо
Оценител А: Успешен (1)	**59** (Съгласуван пропуск)	**2** (Несъгласие)	**61**
Оценител А: Неуспешен (0)	**3** (Несъгласие)	**36** (Съгласен неуспех)	**39**
Общо	**62**	**38**	**100**

Таблица 10.4.2-2 Матрица на грешката (оценител А срещу оценител Б)

Процент на успеваемост

- Процент на успешно издържан оценител А: 61%
- Процент на успешно издържан оценител В: 62%
- Процент на успешно преминаване на теста с консенсус (строг, и двете страни трябва да са съгласни): 59%

Надеждност между оценители ($\varkappa$)

$$\varkappa = \frac{P_0 - P_e}{1 - P_e} = \frac{0.95 - 0.5264}{1 - 0.5264} = \frac{0.4236}{0.4736} \approx 0.894$$

Заключение и решение на одита

Показател за успеваемост: $P_{НЕСА} \approx 60\,\%$

Статус: **неуспешен**. Прагът T_H обикновено е 0,90–0,99. Процент на успешно преминаване от 60% означава, че генераторът на синтетични данни не е подходящ за употреба.

Показател за надеждност: $\varkappa = 0.89$

Статус: **успешен**. Прагът K_{min}, обикновено, е 0,6–0,8. Резултат от 0,89 показва „почти перфектно“ съответствие.

10.4.3 LLM-PA пробен период

Системният прозорец за заявка (конфигурация на одит)
Системна роля: Вие сте одитор по съответствие с изискванията за риск.
Контекст: Поведенчески данни за кредитен продукт на BNPL.

Правила:

- R1: Ако възрастта < 18, то кредитният лимит трябва да е 0.
- R2: Ако доходът е < 20000, то кредитният лимит трябва да е < 500.
- R3: Ако активните заеми > 5, то кредитният лимит трябва да бъде < 1000.

Задача: Оценка на входния JSON. Върнете строго валиден JSON с ключове: " pass_all " (boolean), " failed_rules " (list) и "rationale".

Функция	**100 записа (пилот)**	**1000 записа (партида)**	**10 000 записа (продукция)**
LLM-PA Време	~6 секунди	~1 минута	~10 минути
Цена на LLM-PA	< 0,01 щ. д.	~ $0.01	~ $0.12
Коефициент на мащабиране	Линеен$(O(N))$	Линеен$(O(N))$	Линеен$(O(N))$
Пречка	Латентност	Умора / Продължителност на вниманието	Ценова и времева осъществимост
HECA време (човешко)	~50 минути	~8,3 часа (1 пълен работен ден)	~83 часа (2+ работни седмици)
Цена на HECA (цена 20/час)	~ $16.60	~ $167.00	~ 1660 долара
Оценка	Препоръчано от HECA Използвайте за калибриране на политики	Точка на преход Вземане на проби от експерти става необходимо.	Препоръчва се LLM-PA HECA е прекалено скъпа

Таблица 10.4.3-1 Сравнение: Мащабируемост на LLM-PA спрямо HECA

10.4.4 LOFO-AC Trial

LOFO метрика:

$$P_{LOFO}=1-\frac{L_{final}}{L_0}=1-\frac{1\,800\,000}{18\,500\,000}=1-0.097\approx 0.903$$

Метричен	**Стойност**	Тълкуване
Първоначални	0.66	Суровите данни имат умерена структура, но значителен шум (40-те лоши записа).
Крайни	0,90	Пас (¿T_C). Наборът от данни притежава силна вътрешна статистическа съгласуваност.
Намаляване на	71%	Калибрационният цикъл успешно намалява „статистически

Метричен	Стойност	Тълкуване
MSE		много слабо вероятната“ дисперсия със 71%.
Въздействие върху бизнеса	Частично поправяне	Въпреки че LOFO-AC поправя статистическите аномалии (например, намалява лимита за запис №091 от 12 хиляди на 5 хиляди), процедурата не гарантира перфектно спазване на правилата (например, лимитът за възраст 17 години може да бъде намален до $150, а не строго до $0).

Таблица 10.4.4-1 Обобщение на резултатите

	Оригинал: Не преминава	**Оригинал: Преминава**
LOFO: Неуспешен	**32** (Остатъчни грешки)	**0** (Изкривявания)
LOFO: Успешен	**8** (Пропуски)	**60** (Запазени)

Таблица 10.4.4-2 Класификационна матрица на LOFO-AC

10.5 Дискусия върху практически предложения и ограничения

10.5.1 Генериране на извадки

Препоръчва се строг (дисциплиниран) конвейер, който започва със стратифицирано вземане на извадки от одитния набор S: разделя се синтетичната популация по важни демографски данни/сегменти и по високорискови предшественици на правилата (напр. „Възраст < 18“), така че да бъдат представени редки, но критични за безопасността комбинации; размерност S, така че процентите на успешно преминаване на ниво правило да допускат тесни граници на доверие (напр. интервали на Уилсън за желаната граница на грешка).

10.5.2 Правила

Нанася се една единствена, машинно-четима рубрика: кодира се всяко правило като недвусмислен IF-THEN оператор с имена на променливи, мерни единици, конвенции за закръгляване/прагове и изрично обработване на липсващи стойности; добавя се кратка таксономия от кодове за грешки, така че както HECA, така и LLM да могат да връщат структурирани причини.

10.5.3 LLM-PA заявка

Тъй като всеки запис изисква едно (или повече) API извиквания, LLM-PA води до линейни разходи с броя на записите и вариантите на подканите (напр. мащабира се линейно в $O(N)$ както времето, така и разходите). За големи набори от данни се препоръчва прилагането му към стратифицирани извадки или използване на по-малки, локално хоствани LLM, за да се намалят разходите за извод.

Добавя се точния списък с правила, създава се повторяемост на тествания запис като компактен JSON обект, задава се детерминистично декодиране (температура ≈0) и се изисква JSON отговор като {"pass_all": true, "failed_rules": ["R3", ...]}.

LLM-PA се мащабира добре, но остава чувствителен към заявките и склонен към отклонение (дрейф): малки промени във формулировката или актуализации на модела могат да променят заключенията, така че изследователите трябва да коригират заявките, да фиксират версиите на модела и рутинно да проверяват отново съгласуваността между LLM и човек върху стационарен набор от примерни елементи, като наблюдават за внезапни спадове в κ, които сигнализират за съществено разминаване в калиброванието [13,14].

10.5.4 Настройване на LOFO-AC

LOFO-AC предлага количествено самовъзстановяване, но оптимизира вътрешната консистентност, а не външния реализъм, така че без ограничения може да се отклони от бизнес семантиката, която е мотивирала синтетичните характеристики; на това може да се противодейства посредством регуларизация (Ридж/еластична мрежа), ортогонализация на синтезираните характеристики спрямо X – по избор, ограничения за знак/кутия върху актуализациите на теглото и консервативно затихване за намаляване на осцилациите и капаните на локалния минимум [10, 11, 12, 15].

За избор на LOFO-AC модел, се стартира с регресионен модел на Ридж $\alpha=0$ за скорост (бърза, затворена форма, стабилна при мултиколинеарност) и настройване на λ чрез вложена или обобщена кръстосана валидация. Недостатъчната релаксация $\alpha<1$ (напр. 0.1-0.5) често осигурява стабилна конвергенция. По-високите размерности може да изискват много малко PLS компоненти. При необходимост от разреденост се преминава към еластична мрежа. Ако стойностите на m са големи и силно зависими – проектира се Z_j към няколко PLS компонента преди регресия [10, 11, 12].

Подобрява се числената стабилност в LOFO цикъла с Jacobi (синхронни) актуализации и недорелаксация $W^{(t+1)} \leftarrow (1-\gamma)W^{(t)}+\gamma W_{\text{new}}$ с $\gamma \in [0.3, 0.7]$ за всяка итерация, задава се правило за спиране, базирано на изчакване и се поддържа одитна следа на всички $W^{(t)}$ показатели за приемане за управление и възпроизводимост [15].

10.5.5 Мащабируемост и разход

Времевите и финансовите разходи на LLM-PA нарастват с броя на записите (и броя на използваните модели/подсказки). Времето за изпълнение на LOFO-AC нараства с броя на характеристиките (напасване на една регресия на характеристика на итерация) и точките от данни. Възможно е ограничаване на LLM-PA до подмножество за валидиране или да се използва по-малък брой заявки, когато N е с голяма стойност. По подобен начин, за LOFO-AC, задава се нисък лимит за максимална итерация или се намалява набора от характеристики, ако p е с големи стойности.

HECA одитите предоставят нормативна основа, но са изправени пред ограничения, свързани с разходите, обхвата и умората; смекчават се с фокусирани, стратифицирани извадки, кратки сесии, обучение по рубриката и рутинно отчитане на надеждността между оценители, така че заинтересованите страни да могат да пренебрегнат шумните панели, вместо да преувеличават суровите проценти на успешно преминаване.

10.5.6 Поверителност и справедливост

Изпращането на данни към външни API (LLM-PA) може да доведе до изтичане на чувствителни атрибути. С цел предотвратяване на изтичане на информация се прилага редакция и се осигурява съответствие с политиките за обработка на данни. Разглеждат се вариантите за използване на локална LLM система или защитен анклав, за да избегнете излагане на данни на трети страни. За записи, съдържащи чувствителни полета, се препоръчва прилагане на редакция, маскиране или минимизиране на бързи стъпки. Когато е възможно, организациите трябва да използват локални LLM системи или защитени API, за да избегнат излагане на данните на външни опасности.

Относно справедливостта - всички методи могат да усилят пристрастията, присъстващи в правилата или данните. Препоръчваме одитиране в различни демографски сегменти: изисква се валидаторите да отговарят на праговете за най-лошия случай. Например, приемането на входа се основава на най-лошия сегмент, а не само на общите показатели (както е отбелязано в Заключението).

10.5.7 Ръководство за рамката за изработване на решения

Ако HECA и LLM-PA не са в съответствие (напр. процентът на успешно преминаване на HECA е висок, но LLM-PA е нисък), консервативната политика в този случай, предполага да се изиска всички валидатори да отговарят на

поставените прагови критерии, преди да се премине към приемане на данни. Като алтернатива може да се използва претеглен подход. Изрично оповестяване дали се използва правило „И“ (всички проверки са успешни) или правило „ИЛИ“ (всяка проверка е успешна). Ако валидаторите дадат различни резултати (напр. HECA одобрява, докато LLM-PA отхвърля), препоръчва се една от следните политики:

- **Консервативен сценарий**: Изисква се всички валидатори да преминат преди приемане (правило „И“).
- **Претеглен сценарий:** Присвояват се тежести на важност на всеки валидатор и се дефинира съставен резултат.
- **Сегментиран сценарий:** Използва се HECA за високорискови срезове от данни, LLM-PA за рутинен мониторинг и LOFO-AC за настройка.

Изборът зависи от критичността на домейна, регулаторните изисквания и наличността на ресурси. В бъдеще може да се разработи официална стратегия за обединяване на решения.

И трите метода могат неволно да усилят отклоненията, ако правилата кодират исторически неравенства или ако одитните извадки не представят достатъчно уязвимите подгрупи. Важно е да се приоритизират разрезите както при извадките, така и при отчитането като приемането да се основава на представянето на най-лошия разрез, а не само на общите средни стойности.

Примерите в тази статия са илюстративни и служат за демонстриране на методологичните механизми. По-мащабните валидации върху реални или високоразмерни синтетични набори от данни представляват интерес за бъдеща работа.

10.6 Заключение

Настоящото проучване оценява три независими подхода за валидиране на синтетични данни за поведенческо моделиране - Одит за съгласуваност между хора и експерти (HECA), Програмен одит, базиран на LLM (LLM-PA) и Автоматизирано калибриране с изключване на една характеристика (LOFO-AC) - и подчертава разликата между статистическа правдоподобност и бизнес приемливост. Не се наблюдава един доминиращ метод; всеки един от разглежданите подходи служи за различна цел в процеса на валидиране при ограничения, свързани с поверителност, недостиг и управление.

HECA (експертен одит). HECA операционализира експертния преглед в областта в измерими показатели: процент на успешно преминаване на съответните правила, на ниво запис и надеждност между оценители (експерти) (напр. на Коен/Флайс κ), с изрично маркиране на правила за проследяване на

неуспехите до конкретни ограничения. Като самостоятелен инструмент за спазване на политиките, HECA е тясно съгласувана с организационните норми и създава одитируема следа - подходяща за регулирани контексти, където неверните потвърждения носят оперативен или одитен риск. Ограниченията са свързани с цена, умора на проверяващите и ограничено покритие; за надеждност са необходими стриктно вземане на извадки и ясни, машинно/човешки четими рубрики. Приемането се основава на спазване на праговете за процент на успешно преминаване, определени от критичността на приложението.

LLM-PA (програмен одит). LLM-PA мащабира проверката на правилата, като използва заявки към модели с точния набор от правила и към всеки запис, връщайки структурирани решения и обосновки. Използван самостоятелно, разглежданият подход е подходящ за високопроизводителен скрининг, където правилата са ясни, а малък, маркиран от човек „златен“ набор позволява калибриране. Приемането на съответния запис изисква както целеви процент на успешно преминаване, така и съгласие между LLM и човека (κ), надвишаващо предварително зададените прагове. Ключовите рискове са: чувствителността на заявките, отклонението на модела/версията и потенциалното излагане на поверителност; преодоляването на тези рискове включва: фиксирани заявки, стабилизиране на модела, редактиране/минимизиране и периодично повторно калибриране към стационарен златен набор. LLM-PA допълва (но не замества) човешкото управление, когато обяснимостта и отчетността са от първостепенно значение.

LOFO-AC (автоматизирано калибриране). LOFO-AC е самовъзстановяващ се цикъл: третира всяка оригинална характеристика като временна цел, напасва се от останалите реални характеристики плюс текущия синтетичен блок (Ридж/еластична мрежа/PLS), актуализира синхронно теглата на синтеза (Jacobi) със затихване и се итерира, за да се намалят грешките при реконструкцията и нарушенията на правилата. Използван самостоятелно, LOFO-AC е най-подходящ за техническо калибриране на механизъм за синтез, като приемането е обвързано с подобрения спрямо наивни базови линии и стабилност на итеративната процедура. Фокусът му е върху вътрешната съгласуваност; без ограничения може да се отклони от бизнес семантиката, така че се препоръчват регуларизация, ортогонализация и консервативно затихване.

Общо сравнение.

- Управление/високи залози – продължава/не продължава- HECA.
- Мащабно/оперативно приоритизиране с измеримо съответствие с експерти: LLM-PA.

- Подобряване на точността на генератора към структурата на реалните данни: LOFO-AC.

Предлага се използване на HECA за критични проверки за съответствие, LLM-PA за групово сортиране по фиксирани правила и LOFO-AC за настройване на генератора (напр. подобряване на статистическата точност) - или да се използват в комбинация, ако е уместно. Заедно, предложените валидатори разглеждат централния въпрос, повдигнат във въведението на текущата работа: Как да се определи дали синтетичните данни са подходящи за вземане на оперативни решения?

В различните методи, се отчита производителността по разрези (особено сегментите, критични за обективността) и преминаване към следващия етап за най-лошите резултати от разрезите, не само за средните стойности. Това съгласува валидирането с реалните рискове при вземане на решения, като същевременно запазва поверителността и позволява напредък от типа „студен старт“ до „зрели“ модели. В тази статия разглежданите примери имат илюстративен характер, но бъдещата работа трябва да тества мащабируемостта върху по-големи или реални набори от данни. Бъдещата работа има за задача да изследва оптимизирани правила за вземане на решения и оценки в по-голям мащаб.

Използвана литература

[1] Marchev, A., Jr., & Marchev, V. (2022). Synthesizing multi-dimensional personal data sets. In Proceedings of the 47th International Conference “Applications of Mathematics in Engineering and Economics” (AIP Conference Proceedings, Vol. 2505, No. 1, 020012). AIP Publishing. https://doi.org/10.1063/5.0100615

[2] Marchev, A., Jr., & Marchev, V. (2024). Automated algorithm for multi-variate data synthesis with Cholesky decomposition. In Proceedings of the 7th International Conference on Algorithms, Computing and Systems (ICACS '23) (pp. 1–6). Association for Computing Machinery. https://doi.org/10.1145/3631908.3631909

[3] Marchev, A., Jr., Marchev, V., & Efremov, A. (2024). Methodology for cold-start modeling using synthetic data. In *Proceedings of the 49th International Conference “Applications of Mathematics in Engineering and Economics”*. AIP Publishing. https://doi.org/10.1063/5.0246285

[4] Seufert, E. B. (2014). Quantitative methods for product management. In *Freemium economics: Leveraging analytics and user segmentation to drive revenue* (pp. 47–82). Elsevier. https://doi.org/10.1016/B978-0-12-416690-5.00003-8

[5] Liu-Thompkins, Y., & Malthouse, E. C. (2017). A primer on using behavioral data for testing theories in advertising research. *Journal of Advertising, 46*(1), 213–225.
[6] Dehghan, P., Alashwal, H., & Moustafa, A. A. (2022). Applications of machine learning to behavioral sciences: Focus on categorical data. *Discovery Psychology, 2*, Article 22. https://doi.org/10.1007/s44202-022-00027-5
[7] Motiwalla, L., Deokar, A. V., Sarnikar, S., & Tao, L. (2019). Leveraging data analytics for behavioral research. *Information Systems Frontiers, 21*, 735–742. https://doi.org/10.1007/s10796-019-09928-8
[8] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104
[9] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378–382. https://doi.org/10.1037/h0031619
[10] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634
[11] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x
[12] Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems, 58*(2), 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1
[13] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., Zhang, H., Gonzalez, J., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-bench and chatbot arena. *Advances in Neural Information Processing Systems 36* (NeurIPS 2023). https://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html
[14] Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., Wang, Y., Gao, W., Ni, L., & Guo, J. (2024), A survey on LLM-as-a-judge. arXiv. https://doi.org/10.48550/arXiv.2411.15594
[15] Heath, M. T. (c. 2013-2015). Iterative methods for linear systems [Lecture notes]. Department of Computer Science, University of Illinois. https://courses.grainger.illinois.edu/cs554/fa2015/notes/10_iterative.pdf
[16] Marchev, A., Jr., Marchev, V., & Efremov, A. (2025). Data expansion by business-logic scaling - an applied approach for data synthesis. In *Proceedings of the 2025 ICARAI Conference* (pp. 1–7). IEEE. https://ieeexplore.ieee.org/document/11137819/
[17] Du, B., Qian, K., Claudel, C., & Sun, D. (2022). Jacobi-style iteration for distributed submodular maximization. *IEEE Transactions on Automatic Control, 67*(9), 4687–4702.

[18] Figueira, A., & Vaz, B. (2022). Survey on synthetic data generation, evaluation methods and GANs. *Mathematics, 10*(15), 2733. https://doi.org/10.3390/math10152733

[19] Zamzmi, G., Subbaswamy, A., Sizikova, E., Margerrison, E., Delfino, J. G., & Badano, A. (2025). Scorecard for synthetic medical data evaluation. *Communications Engineering, 4*, Article 130. https://doi.org/10.1038/s44172-025-00450-1

[20] Hernández, M., Epelde, G., Alberdi, A., Cilla, R., & Rankin, D. (2022). Synthetic data generation for tabular health records: A systematic review. *Neurocomputing, 493*, 28–45. https://doi.org/10.1016/j.neucom.2022.04.053

[21] Santangelo, G., Nicora, G., Bellazzi, R., & Kalra, D. (2025). How good is your synthetic data? SynthRO, a dashboard to evaluate and benchmark synthetic tabular data. *BMC Medical Informatics and Decision Making, 25*, Article 89. https://doi.org/10.1186/s12911-024-02731-9

[22] Yan, C, Yan, Y., Wan, Z., Zhang, Z., Omberg, L., Guinney, J., Mooney, S. D., & Malin, B. A. (2022). A multifaceted benchmarking of synthetic electronic health record generation models. *Nature Communications, 13*, Article 7609. https://doi.org/10.1038/s41467-022-35295-1

[23] Sarmin, F. J., et al. (2025). Synthetic data: Revisiting the privacy-utility trade-off. *International Journal of Information Security, 24*(4), Article 156. https://doi.org/10.1007/s10207-025-01072-6

[24] Pradhan, A., Ortan, A., Verma, A., & Seshadri, M. (2024). *LLM-as-a-judge: Rapid evaluation of legal document recommendation for retrieval-augmented generation*. arXiv. https://doi.org/10.48550/arXiv.2509.12382

[25]

11 Генериране на данни при наличие на части от масиви (ProbCon)

Автори: Васил Марчев, Ангел Марчев, мл.

Резюме: В настоящата глава се разглеждат възможностите за комбиниране на два или повече налични набори с данни по предварително установени повтарящи се характеристики във всеки един от наборите. Фокусът е поставен върху реализирането на хибриден модел за вероятностно конкатениране на записи, който минимизира грешките при сливане на хетерогенни структури от данни. Специално внимание се отделя на методиката за бинарна трансформация на ключовите атрибути, осигуряваща висока изчислителна устойчивост и стандартизация на променливите. Представеният итеративен алгоритъм създава предпоставки за последователно надграждане на масива чрез инкорпориране на различните характеристики. В допълнение е дефиниран стриктен процес на декомпозиция и филтриране, насочен към елиминиране на техническия шум и защита на анонимността на респондентите. Така синтезираният набор от данни притежава необходимата детайлност за прецизно статистическо моделиране на изследваните зависимости.

Ключови думи: вероятностна конкатенация, обединяване на данни, бинарно кодиране, интеграция на хетерогенни масиви

11.1 Същност

Различните методи за симулация на данни се основават върху съществуващи концепции, прилагани при липса на необходимото количество и качество на информацията, които за целите на настоящия текст са пречупени през призмата на обсъжданата тематика. Разглежданият в тази глава подход се базира на априорни знания, анализи и наблюдения осъществени от авторите на настоящия труд. Методът е реализиран за целите на изследването и е изцяло съобразен с характеристиките на разработваните модели.

Същността на разглеждания метод се състои в обединение на два, или повече съществуващи реални масиви от данни. Подходът стъпва върху основите на научното изследване на автоматизираното свързване на данни, които се базират върху фундаменталните принципи за машинна интеграция на записи при отсъствие на единни идентификатори [1]. Подходът се разглежда в условието на комбиниране на два или повече реални набора от данни, които се съчетават посредством уникален идентификационен ключ. Той се изгражда от две или повече променливи, които присъстват и в двете матрици. Алгоритъмът

изисква анализ на независимите масиви от информация. При разбора се извършва наблюдение на съществуващите реални набори от данни и техните променливи. В случай че се установи необходимост от анонимизиране на масивите от информация, тя се извършва като фаза, следваща тази на анализа.

Комбинацията на съществуващи затворени масиви от информация преминава през създаването на уникален идентификационен ключ, базиран на две или повече променливи. Въз основа на този ключ алгоритъмът осъществява външно добавяне на информация и завършва с цялостна комбинация от всички варианти на записите.

В процеса на вероятностна конкатенация на база данни, някои от генерираните стойности могат да бъдат невъзможни или дублирани. Методът изисква използваните масиви да са приведени в подходящ и удобен за моделиране вид. В тази връзка се осъществява редуциране на новогенерираните данни, като се премахват дублираните записи от лявата матрица. Резултатът е нова матрица, намалена чрез изтриване на редове.

11.2 Математически алгоритъм

Използван е математически алгоритъм, който гласи следното:

Дадено е ω да бъде многомерна, крайна, случайна променлива от реални данни, обозначаваща k брой от числови характеристики в изучаваната система S, докато ω включва n брой редове (така че е $n \times k$ матрица).

Дадено е λ да бъде друга отделна многовариантна крайна случайна променлива на реални данни, обозначаваща l брой количествени характеристики на изследвана система S, докато λ включва t брой редове (така че е $t \times l$ матрица).

Като се има предвид, че има пресечна точка между ω и λ от поне две колони (докато $n \leq t$): $[\omega] \cap [\lambda] = \{x_1 \dots x_p \vee x_1 \dots x_p \in \omega \wedge x_1 \dots x_p \in \lambda \wedge p \geq 2\}$.

Осъществява се ляво външно присъединяване на ω с λ (строго в този ред) $\omega \bowtie \lambda = \{\tau_{i,j} \vee i = n \wedge j = k = l - p\}$, като се търси съвпадение на съответните редове, по ключови стойности от $\hat{x}_1, \dots, \hat{x}_p$, за да се получи конкатенирана матрица $X : {}^{X_{t,s} = [\omega \lambda]}$

Ако се установят съвпадащи подмножества λ и ω, $\{\lambda \subseteq \lambda \vee \hat{x}_1, \dots, \hat{x}_p \subseteq \lambda\} \wedge \{\omega \subseteq \omega \vee \hat{x}_1, \dots, \hat{x}_p \subseteq \omega\}$, където $\lambda > \omega$, избира произволен ред за съчетаване λ, така че X е редуцирано до $\widetilde{X}$ така, че $\widetilde{X}_{n,s} \subseteq X_{t,s}$ и $\widetilde{X} = ¿$ и, където $n \leq t$.

11.3 Последователност и приложение на метода

11.3.1 Приложение на метода

Процесът по генериране на необходимите за модела данни предполага интегриране на повече от един подход за синтезиране на информация. За целта, освен гореописания подход за симулация на демографски данни, е набавена информация относно предпочитанията на физическите лица при инвестиране, информация за инвестиционните им цели, индивидуална склонност към риск и данни за психологическите и личностни характеристики. Осигуряването на информация относно посочените групи се осъществява посредством въпросник и допълнителни набори от информация.

Получените съвкупности от данни е необходимо да бъдат обединени в един общ информационен източник. След изготвяне на гореописания масив от информация се преминава към комбинация на данните с наличните бази данни от въпросниците и допълнителните набори от данни. По този начин се цели да се получи пълен набор от информация за съответните записи.

Методът, който се използва за генериране на обобщен масив от информация се нарича вероятностна конкатенация и представлява процес по обединение на два, или повече съществуващи реални масива от данни.

Обединяват се информационните масиви за демографски характеристики (предварително симулираните данни) и данните получени от изследванията на личностните черти за определяне на темперамента на лицата. За целта се генерира уникален идентификационен ключ, базиран на две или повече променливи. В конкретния случай това са пол и години от лявата матрица с данни (симулираните демографски данни) и отново пол и години от дясната матрица (личностни черти), което изглежда по следния начин:

```
left_on = ["sex", "age_sim"]
right_on = ["gender", "age"]
```

Фиг. 11.3.1-1 Променливи използване за генериране на „ключ“

Използват се посочените две променливи поради липса на сходство между останалите характеристики в двата набора от данни. Така приложения ключ предоставя широк списък от възможни комбинации, което дава възможност за установяване на по-голям случаен момент при избор на съответстващ запис.

След обособяването на ключ се преминава към съпоставяне на двете матрици, поради големия брой възможности за комбинации се добавя случаен компонент и компонент проследяващ качеството на потенциалната

комбинация. Ако не е удовлетворено изискването за пригодност на новия запис – отпада от селекцията.

Функцията, използвана за напасване на двете матрици е както следва:

```
match_table = fm.link_table(demogr, big5, left_on, right_on, left_id_col='ind', right_id_col='id')
```

За всеки един запис от лявата матрица се предоставя набор от възможни комбинации от дясната матрица (таблица 11.3.1-1).

	__id_left	__id_right	match_score	match_rank	sex	gender	age_sim	age
0	10001	27163	0.072145	1	2	2	52	52
1	10001	32407	0.071044	2	2	2	52	52
2	10001	31056	0.068877	3	2	2	52	52
3	10001	33594	0.068146	4	2	2	52	52
4	10001	28854	0.067393	5	2	2	52	52
5	10001	24299	0.067339	6	2	2	52	52

Таблица 11.3.1-1 Възможни комбинации между лява и дясна матрица

След генериране на масив от възможни комбинации се преминава към същинския процес по вероятностно конкатениране, което се осъществява посредством следната функция, съдържаща случаен компонент:

```
matched_results = fm.fuzzy_left_join(demogr, big5, left_on, right_on, left_id_col='ind', right_id_col='id')
```

11.3.2 Подробно описание на последователността на метода

Подробното описание на процеса по конкатенация на данни обхваща изчерпателен технически анализ на софтуерен алгоритъм, проектиран за консолидиране на разнородни информационни масиви чрез последователно приложение на детерминистични и вероятностни методи за свързване на записи.

Процесът на съпоставяне преминава през няколко съществени фази, които могат да бъдат обособени в посочената последователност.

11.3.2.1 Подготовка и първоначално вероятностно конкатениране

Първата фаза в процеса е дефиниране на работната среда и импортиране на специализирани библиотеки за манипулация на структурирани данни и размито напасване на зашумени данни [2]. Първоначалните два набора данни — демографски профили и психометрични индекси са подложени на ляво съединение. Тъй като ключовете за свързване (пол и възраст) могат да съдържат вариации или шум, се прилага функция за размито ляво съединение. Като това се повтаря за всеки чифт записи, като съединението се осъществява само ако отговаря на всички предварително зададени критерии. След сливането се

извършва стриктна редукция на размерността посредством премахване на технически метаданни и излишни идентификатори.

11.3.2.2 Нормализация посредством бинарна трансформация

След първоначалното напасване, се въвежда иновативен подход за стандартизация на ключовите променливи. Стойностите на атрибутите „пол“, „възраст“ и „образование“ биват конвертирани в осембитови бинарни низове. Посредством прилагане на специализиран форматен оператор, всяка целочислена стойност се преобразува в уникална бинарна последователност. Приложената методология за бинарна стандартизация на променливите се опира на съвременните изисквания за осигуряване на високо качество и съгласуваност на връзките в сложни масиви от данни [7]. Посредством разгледания процес се цели да се минимизира изчислителната неяснота и служи като хеш-функция за подготовка на данните за следващите етапи на интеграция. Операцията се повтаря рекурсивно за всички нови масиви, като по този начин се осигурява хомогенност на ключовете.

11.3.2.3 Итеративна многослойна интеграция

Алгоритъмът продължава с последователно добавяне на следващите характеристики, като се прилага последователен модел на свързване следвайки логиката на итеративната многослойна интеграция. Итеративният подход представлява последователност от повторяеми действия, с цел да се постигне изцяло запълване на възможните записи [3].:

- **Съединение на рискови профили:** Използват се вече генерираните бинарни кодове за сравняване на синтезирания до момента масив със съответните рискови профили.
- **Повторна инициализация на ключовете:** След всяко съединение софтуерът преизчислява бинарните вектори, за да гарантира, че евентуални липсващи стойности или нови записи са коректно индексирани.
- **Краен резултат:** Последното свързване интегрира таблицата с инвестиционни нагласи, използвайки разширеното множество от бинарни идентификатори (в това число: пол, възраст, образование).

11.3.2.4 Финализиране и почистване на масива

В заключителната фаза се прилага агресивен алгоритъм за почистване на данни. Премахват се общо 39 колони, които биха могли да внесат шум или да компрометират анонимността на участниците в изследването. Този процес на декомпозиция елиминира всички междинни изчислителни променливи,

географски маркери и специфични демографски детайли, предоставяйки филтриран набор от променливи, готов за последваща обработка и приложение в модела.

11.3.2.5 Анализ на процеса на редукция на данните

В рамките на предоставения алгоритъм, премахването на променливи не е произволно. Процесът следва строго зададена последователност за осигуряване на статистическа валидност и защита на личните данни, като от съществено значение е съответствието с GDPR. Съществен аспект от последователността на модела представлява необходимостта от анонимизация на записите, рисковете съпътстващи евентуалната деанонимизация са разгледани от редица изследвания предоставящи богат набор потенциални рискове и възможности за справяне с тях [4]. В таблица 11.3.2.5-1 е представена систематизирана оценка на действията по премахване на информационни масиви, в контекста на настоящата глава.

Променливи	**Категория**	**Забележка**
best_match_score, __id_right, __id_left, Unnamed: 0, id	Технически елементи	Премахване на служебни индекси и вероятностни оценки от посредством функция за размито присъединяване, които нямат аналитична стойност за крайния модел.
binsex, binage_sim, binlv_educ, bingender, binage, binedu_level	Трансформационни вектори	Разгледаните бинарни низове служат за свързващо звено. Запазването им може да доведе до изкуствено увеличаване на размерността на масива.
nation, town, prof, position, industry	Социално-демографски променливи	Тези променливи носят висок риск за де-анонимизация. Премахването им гарантира фокус върху психологическите конструкции вместо върху локални специфики.
montly_income, income, houshold_people	Икономически променливи	Елиминиране на потенциална мултиколинеарност. Доходите често корелират силно с нивото на образование, което вече е включено в модела.
shares_left, crypto_left,	Специфични променливи	Редукция на хетерогенните

gov_bond, inv_fund, cash_left		финансови променливи до по-високи нива на абстракция (напр. общ рисков профил вместо отделни инструменти).
gender_left, age_left, gender_right, age_right	Дублирани променливи	След успешно напасване, тези дублирани колони са идентични с основните sex и age_sim. Премахването им е задължително за избягване дублиранията и намаляване на необходимостта от изчисления.

Таблица 11.3.2.5-1 Премахване на информационни масиви

11.3.2.6 Резултати от проведената конкатенация

Процесът по съпоставка и комбинация между различните набори от данни завършва с генериране на обобщен масив от информация и с подробен доклад относно проведеното конкатениране и съответните особеностите.

Сдвояването между новополученият набор от данни и информацията, достъпна от проведените анкети се осъществява по аналогичен на описания начин. Като в края на целия процес се получава общ набор, в който са интегрирани симулираните демографски данни, информацията за личностните характеристики, инвестиционните преференции (инвестиционни възможности, хоризонт, срок на преразглеждане на инвестицията и др.) и индивидуалната склонност към риск от проведените изследвания. Синтезираният краен масив от информация предоставя теоретично издържана и математически обоснована рамка, чрез която се създават предпоставки за извличане на значими икономически зависимости, съобразени с иконометричните изисквания за комбиниране на данни [6].

11.4 Предимства и потенциални ограничения на методологията

11.4.1 Предимства

Прилагането на описания алгоритъм осигурява няколко стратегически предимства, които повишават качеството и надеждността на научния анализ:

- **Методологична устойчивост при хетерогенни данни:** Хибридният модел, комбиниращ размито напасване с бинарно кодиране, позволява прецизно интегриране на масиви, които нямат общ уникален идентификатор. Това преодолява ограниченията на традиционните методи за релационно свързване.

- **Повишена изчислителна прецизност:** Трансформацията на променливите в 8-битови вектори стандартизира сравнителния процес, елиминирайки грешки, произтичащи от вариации в изписването или мащаба на числовите данни. Този подход действа като филтър, който уеднаквява логическата структура на данните преди тяхната финална обработка.
- **Оптимизация на информационния обем:** Систематичната декомпозиция на излишните атрибути и техническия шум подобрява статистическата мощ на модела. Чрез редуциране на размерността се предотвратява рискът от пренапасване и се гарантира, че крайните резултати се базират само на теоретично обосновани ковариати [3].
- **Автоматизирана консистентност:** Скриптовият характер на обработката гарантира, че първоначално установения набор от правила се прилагат идентично върху всички итерации на данни. Това минимизира субективния фактор и позволява пълна прозрачност и повторяемост на описаната процедура.
- **Високо ниво на сигурност и етика:** Процедурата по окончателно почистване на данните служи като вграден механизъм за анонимизация. Премахването на квази-идентификаторите гарантира, че крайният масив не създава предпоставки за риск от нарушаване на поверителността на участниците.

11.4.2 Потенциални недостатъци

Въпреки високата ефективност на предложената методология за вероятностно конкатениране на информационни масиви, следва да бъдат отчетени няколко потенциални методологични ограничения, произтичащи от спецификата на автоматизираната обработка на данни:

- **Вероятностни грешки при напасването:** Използването на метрики за размито сходство носи присъщ риск от грешки от вид фалшиво положителни съответствия, както и от вид пропуснати реални връзки [5]. Това може да доведе до отклонения в детайлите на крайния масив при наличие на екстремен шум в първичните данни.
- **Резолюция на бинарното кодиране:** Трансформацията на променливите в предложения бинарен формат налага преобразуване на непрекъснатите стойности в дискретни. Макар това да оптимизира процеса на свързване, то би могло да ограничи разделителната способност при анализиране на променливи с изключително голям диапазон от уникални стойности.

- **Итеративно натрупване на грешките:** Тъй като процесът е последователен, всяко неточно напасване в първия етап се пренася в следващите фази на интеграция. Това изисква стриктен предварителен контрол на качеството на входните масиви.
- **Информационна редукция:** Агресивното премахване на 39 променливи в заключителната стъпка е наложено от нуждите на модела, но съществува теоретичен риск от елиминиране на латентни ковариати, които биха могли да имат допълнителна обяснителна стойност.

11.5 Заключение

В настоящата глава е осъществен подробен преглед на авторски изчислителен алгоритъм, осигуряващ систематичен и научно обоснован подход за интегриране на многоизмерни масиви от данни. Използваната синергия между вероятностното размито напасване и детерминистичната бинарна трансформация създава предпоставки за постигане на прецизна синхронизация на респондентите, при липса на унифицирани идентификатори. Разглежданата методология механизъм спомага както за оптимизиране на структурата на данните, така и за минимизиране на системните грешки и техническия излишък, като създава предпоставки за стабилна основа за последващи статистически анализи и моделиране.

Крайният продукт представлява синтезиран информационен масив с подчертан изследователски ресурс, който успешно балансира между високата информационна плътност и стриктните изисквания за анонимност в кореспонденция с най-добрите практики обособени в GDPR и Закона за кредитните институции, в частта уреждаща защитата на банковата тайна. Приложената методология за филтриране и премахване на част от характеристиките обезпечава фокусът на изследването да бъде концентриран върху съществените психологически и финансови корелации, изчистени от изкривявания. Предложеният алгоритъм се утвърждава като ефективен инструмент за обработка и конкатениране, посредством вероятностна компонента, на сложни социално-икономически данни, предоставяйки надеждност и методологична прозрачност.

Използвана литература

[1] Newcombe, H. B., Kennedy, J. M., Axford, S. J., & James, A. P. (1959). Automatic linkage of vital records. Science, 130(3381), 954-959. https://doi.org/10.1126/science.130.3381.954

[2] Christen, P. (2012) Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer Science and Business Media, Berlin.

[3] Winkler, W. E. (2006). Overview of record linkage and current research directions. *Bureau of the Census*, *25*(4), 603-623.

[4] Vatsalan, D., Christen, P., & Verykios, V. S. (2013). A taxonomy of privacy-preserving record linkage techniques. *Information Systems*, *38*(6), 846-869. https://doi.org/10.1016/j.is.2012.11.005

[5] Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). Data Quality and Record Linkage Techniques, New York: Springer Science+Business Media. 234 pp. US$44.95. ISBN: 978-0-387-69502-0. DOI: 10.1007/s11336-007-9040-1

[6] Ridder, G., & Moffitt, R. (2007). Chapter 75 The Econometrics of Data Combination. In J. J. Heckman & E. E. Leamer (Eds.), Handbook of Econometrics (Vol. 6B, pp. 5469–5547). Elsevier, DOI: 10.1016/S1573-4412(07)06075-8

[7] Harron, K., Goldstein, H., & Dibben, C. (Eds.). (2015). Methodological developments in data linkage. John Wiley & Sons. DOI:10.1002/9781119072454

12 Генериране на синтетични данни при наличие на пълни набори от данни. Анонимизация чрез генеративни състезателни мрежи (GAN)

Автори: Ангел Марчев, мл., Васил Марчев

Резюме: Анонимизацията представлява ключов процес в науката за данните, особено в случаите, при които се работи с чувствителна информация, която е предмет на законите за защита на личните данни. Тази глава изследва използването на генеративни състезателни мрежи (GAN) като подход за анонимизиране на данни. GAN използват генератор, за да създават нови масиви от информация, които наподобяват оригиналния набор от данни, както и дискриминатор, който разграничава реалните от генерираните данни.

Ключовите фази на процеса на анонимизация, базиран на GAN, включват подготовка на данни, проектиране на модел на генератор и дискриминатор, състезателно обучение, генериране на синтетични данни и окончателно прецизиране на данните. Този подход позволява създаването на синтетични данни, които запазват статистическите характеристики на оригиналния набор от данни, като същевременно гарантират защитата на личната поверителност. Изследването предоставя технически подробности за архитектурите на невронните мрежи, функциите за активиране и процедурите за обучение, които са от решаващо значение за успеха на тази техника за анонимизация. Като се възползват от възможностите на GAN, заинтересованите страни могат да получат ценна информация от чувствителни данни, без да компрометират личната поверителност.

Ключови думи: синтетични данни, синтез на данни, анонимизация, невронна мрежа, GAN

12.1 Въведение

Нарастващият брой регулативни закони (като GDPR и Закона за банковата тайна) и повишаващата се прецизност на институциите по отношение на поверителността на данните, засилват необходимостта от ефективни техники за анонимизация. Традиционните методи често не успяват да балансират адекватно полезността на данните и поверителността. Това изследване разглежда потенциала на генеративните състезателни мрежи (GAN) за създаване на синтетични набори от данни, които да се използват за реални

финансови услуги. Тези набори от данни запазват статистическите свойства на оригиналните данни, като същевременно гарантират индивидуалната поверителност, като по този начин се обръща внимание на критична празнина в настоящите практики за анонимизация на данни. Допълнителен акцент в настоящата глава е фокуса върху доказване на ефективността на метода при работа с различни области на данни – като демографски данни, финансови услуги и др., включително редки области като психологически и лични предпочитания.

Настоящата работа разглежда изследователския въпрос свързан с анонимизирането на масиви от данни, посредством разработване на цялостна рамка, базирана на GAN. Това включва проектиране и обучение на GAN с генератор за създаване на синтетични данни и дискриминатор за разграничаване на реални от синтезирани масиви. Процесът разглежда в подробности различните фази, като подготовка на данни, проектиране на модел, състезателно обучение, генериране на синтетични данни и последваща обработка, като се гарантира, че новосъздадените масиви запазват основните статистически характеристики на оригиналния информационен набор.

Получените резултати показват, че генерираните посредством GAN синтетични данни отразяват точно статистическите свойства на оригиналния набор от информация. Способността на генератора да създава реалистични данни се валидира посредством серия от тестове посредством дискриминатора, чиято цел е да направи коректно разграничаване между реалните от синтетични данни. Анонимизираните данни запазват високи нива на полезност по отношение на аналитичните цели, като същевременно ефективно се намалява риска от нарушения на поверителността. По време на разработването на методологията е приложена новаторска идея за валидиране на генерираните, посредством GAN, информационни набори, използваща косинусово подобие, което установява нивата на разликите във формата между синтетичните и реалните данни.

Проведен е обширен набор от проверки за устойчивост, които потвърждават, че предложения метод константно произвежда висококачествени синтетични данни в различни информационни масиви и подмножества, демонстрирайки надеждност и гъвкавост.

Фундаменталната работа за разработване на GAN от Goodfellow et al. [5] полага основите за използване на състезателни мрежи в задачи за генериране на данни. Посоченото изследване подчертава потенциала на GAN за генериране на данни, неразличими от реалните, което е от решаващо значение за приложението на метода в процеса на анонимизация на данните. Последващите

изследвания разширяват тази концепция, разглеждайки различни GAN архитектури и техните приложения в генерирането на синтетични данни.

Друга критична област в литературата включва техники за синтез на данни, запазващи поверителността. В проучване на Yoon и колектив [15] са представени методи като ADS-GAN, специално разработени за анонимизиране на лични данни, като същевременно се запазва тяхната полезност по отношение на последващи анализи. Посоченият труд подчертава важността на балансирането между полезността на данните и поверителността, което е и централна тема в настоящото изследване.

Използването на GAN за генериране на синтетични данни в контекста на машинното обучение и изкуствения интелект претърпява значителен напредък. Изследванията на Radford и колектив [8] върху Deep Convolutional GAN (DCGAN) демонстрират ефикасността на GAN при създаването на висококачествени синтетични данни за различни приложения. Тези постижения осигуряват солидна основа за подхода, който е обект на настоящото изследване, към анонимизация на данни с помощта на GAN.

Настоящото проучване допринася за развитието на тази област, като представя ново приложение на GAN - за анонимизиране на данни, съчетаващо теоретични и практически характеристики. Предоставена е подробна методология за внедряване на разглеждания инструмент в посока за генериране на синтетични данни, които запазват статистическата цялост на оригиналните информационни набори. Тази работа преодолява разликата между поверителността на данните и тяхната полезност, предлагайки надеждно решение за обработка на чувствителни данни в съответствие със строгите закони за защита на данните.

Следващите части се задълбочават в техническите аспекти на предложената рамка за анонимизация, базирана на GAN. Обсъдени са стъпките за подготовка на данните, архитектурата на моделите на генератор и дискриминатор, както и спецификите на процеса на обучение. Представена е експериментална настройка, резултати и цялостен анализ на ефективността на разглеждания подход. Настоящата глава завършва с дискусионен панел относно последиците от направените открития и потенциалните бъдещи насоки на изследване.

12.2 Теоретичен преглед

Анонимизацията е неразделна част от науката за анализ на данни, особено в случаите когато се работи с чувствителна информация или такава, която попада в обхвата на законите за защита на личните данни, като GDPR,

Закона за банковата тайна и други. Този процес включва премахване или замяна на личната информация по такъв начин, че данните да не могат да бъдат проследени до отделни лица.

Подходите за справяне с този сложен проблем използват набор от различни методи, базирани на статистически парадигми. Един такъв подход е генерирането на синтетични данни, когато действителни данни не са налични. Синтетичните данни представляват изкуствено създадени данни, които изглеждат като реални данни и имат свойства, наподобяващи данните от реалните масиви, но не съдържат никаква чувствителна информация. Друг подход е анонимизиране на съществуващи набори от данни чрез заглушаване или обобщаване на полета с данни, които могат да идентифицират лица. Такива подходи позволяват гъвкавост при работа с информация, която трябва да бъде защитена [1, 3, 4].

Анонимизирането на наборите от данни може да реши множество проблеми, включително поверителност, недостиг на информация и спестяване на разходи. Това е ефективна техника за анализ на данни, която позволява на бизнеса да получава масиви от чувствителни данни, без да прави компромис с поверителността.

Авторите на настоящата разработка систематизират [4] набор от различни случаи на употреба на синтез на данни. Също така е описан подход за анонимизиране на данни, използващ синтез на информация, базиран на разпределение [1]. Настоящото проучване се фокусира главно върху следната ситуация – налице са съществуващи набори от данни, които трябва да бъдат анонимизирани, за да могат да бъдат използвани, като същевременно се запазват взаимовръзките и разпределенията на оригиналните набори от данни.

12.2.1 Основни моменти при генеративните състезателни мрежи

Генеративните състезателните мрежи (GAN) са дълбоки невронни мрежи, които функционират като генеративни модели на данни. Този модел представлява симулация на данни, базирана на две състезателни невронни мрежи. Ключовите компоненти на подхода са: генеративен модел и дискриминативен модел, които са обучени в игра с нулева сума. Генераторът е обучен да заблуждава дискриминатора във възможно най-висока степен [6, 7, 10].

Генераторът се използва за създаването на нови примери, които наподобяват данните за обучение, докато дискриминаторът оценява генерираните примери, за да разграничи реалните от фалшивите данни. Тази настройка създава конкурентна динамика, при която генеративната функция се

стреми да създаде реалистични данни, за да заблуди дискриминатора. Същевременно, дискриминаторът се стреми към коректно разграничаване на реалните от генерираните данни.

Дискриминаторът приема като вход началната стойност, а за изход се приема вероятността тази точка да лежи върху оптималното разпределение. Той е обучен да изучава характеристиките и моделите на реалното разпределение на данните, така че да може ефективно да идентифицира всякакви отклонения или несъответствия в генерираните извадки. За изразяване на генератора и дискриминатора се използва параметризиран модел.

Моделът за генериране на данни има за цел да заблуди дискриминатора. Невронната мрежа, използвана за създаване на информация, е обучена да увеличи максималната грешка при класификация (между реални и новополучени данни). Основната задача на дискриминатора е да открива неправилно зададените новосинтезирани данни и да навигира (поставяне на етикети) коректно случаите, в които данните са генерирани или реални; следователно, дискриминаторът е обучен да минимизира крайната грешка при класификация [2, 5, 8, 9, 11, 18].

Основните етапи на описания метод са:

- Първата стъпка е да се предоставят реални набори от данни, които ще бъдат използвани както в процеса на обучение, така и при валидиране на новополучените данни.
- Новополученият синтетичен набор от данни се генерира от генеративната изкуствена невронна мрежа въз основа на реалните набори от данни от стъпка първа. Новите редове с данни наподобяват на реалните набори от информация.
- Дискриминаторът е отделна невронна мрежа, която приема наличните данни като вход и извежда вероятност или резултат, показващ вероятността извадката да е реална или да изглежда реална.

12.2.2 Математически алгоритъм

- Дадено е ω като многомерна крайна случайна променлива от S реални данни, обозначаваща броя (от общо S) на количествените характеристики, на изследваната система S и n броя на редовете.
- Използвайки генеративна изкуствена невронна мрежа се генерират стойности на симулираната матрица:

$$\widetilde{X}_{n,s}=\left\{\widetilde{X}_{n,s}\middle|\widetilde{X}_{n,s}\neq\omega\right\}$$

- Използвайки дискриминативна изкуствена невронна мрежа се извършва дискриминиране:

$$\widetilde{X}_{n,s} \quad \omega$$

12.3 Внедряване на генеративни състезателни мрежи

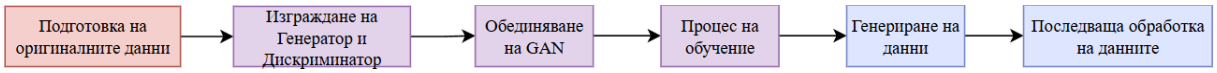
Анонимизирането на данни е ключов процес, насочен към защита на личната неприкосновеност в наборите от данни, особено когато тези данни са предназначени за публично публикуване или използване от трети страни. Подходът, използващ генеративни състезателни мрежи (GAN) за анонимизиране на данни, предоставя усъвършенстван метод за генериране на синтетични данни, като запазва статистическите характеристики на оригиналния информационен масив и същевременно гарантира, че поверителността на лицата не е компрометирана. Този процес включва няколко отделни фази, всяка от които е критична за осигуряване на ефективността на анонимизирането, като същевременно се запазва полезността на данните [15, 16].

В следващите точки е очертан структуриран подход към анонимизация на данни с помощта на GAN, разделен на следните ключови фази:

- **Подготовка на оригиналните данни.** Първоначалната подготовка на данните включва почистване, нормализиране и кодиране на данните, за да се гарантира, че са в подходящ формат за обучение на GAN. Тази фаза е от решаващо значение за съгласуване на характеристиките на входните данни с изискванията на невронните мрежи.
- **Изграждане на генератор и дискриминатор**. Това са основните компоненти на GAN. Генераторът се обучава за създаване на синтетични набори от данни, наподобяващи реални такива, докато дискриминаторът се обучава да установява разлика между реални и синтетични масиви.
- **Компилиране на GAN.** Тази стъпка включва настройване на GAN с подходящи функции за загуба и оптимизатори, интегриране на генератора и дискриминатора в унифицирана рамка за състезателно обучение.
- **Процес на обучение.** GAN се обучава в итеративен процес, при който дискриминаторът и генераторът се подобряват последователно въз основа на тяхната производителност един спрямо друг, подобрявайки качеството на генерираните данни с течение на времето.
- **Генериране на данни.** След като бъде обучен, генераторът се използва за генериране на нови синтетични данни от произволни шумови входове, като ефективно имитира разпределението на оригиналния набор от данни.
- **Последваща обработка на данни.** Крайните синтетични данни често изискват последваща обработка. Преобразуването им в полезния формат на

оригиналните данни е необходимо, за да бъдат приложени обратни трансформации за мащабиране и кодиране.

Тези фази са от решаващо значение за успешното внедряване на GAN в процеса по анонимизиране на данни. Процесът е графично изобразен в следната диаграма, обхващаща стъпките на алгоритъма:



Фиг. 12.3-1 Внедряване на GAN за анонимизиране на данни

12.3.1 Подготовка на данни за анонимизация с помощта на генеративни състезателни мрежи (GAN)

Подготовката на данните е ключова стъпка в прилагането на генеративни състезателни мрежи (GAN) с цел анонимизиране на данните. Тази фаза гарантира, че данните, въведени в модела, са в подходящ формат за обучение, като същевременно създава предпоставки за последващо привеждане на анонимизираните данни обратно в първоначалния им вид, запазвайки както полезността, така и поверителността им. Ефективната предварителна обработка не само улеснява процеса на моделиране, но и гарантира, че анонимизираните данни запазват статистическите характеристики на оригиналния набор от данни, без да се компрометира личната поверителност.

- Проверка за NaN стойности. Преобразуването на колони, съдържащи NaN стойности, в цели числа води след себе си грешка. Необходимо е правилното обработване на NaN стойностите преди преобразуване. Обработката на липсващи стойности е извън обхвата на тази статия, така че тук се приема, че всички стойности (за всички характеристики и записи) съществуват.
- Запазване на целостта на характеристиките, които не са цели числа. Тази стъпка поддържа целостта на определени характеристики в набора от данни, като запаметява кои от тях трябва да са цели числа, така че да могат да бъдат правилно преобразувани обратно, след различните стъпки от процеса на обработка на данните. Това се прави, като първо се идентифицира кои колони трябва да са цели числа. След това се съхраняват оригиналните типове данни на тези колони за по-късна употреба.
- Кодиране на категорийни променливи: Категорийните променливи изискват кодиране, за да се трансформират във формат, който може да се използва ефективно от алгоритми за машинно обучение. Използва се кодиране с етикети, което преобразува всяка категория в уникално цяло число. Този процес е дефиниран за всяка категорийна колона в набора от данни, както следва:

$$X^{(j)}_{encoded} = f^{(j)}_{encoding}\left(X^{(j)}\right), \text{for all } j \text{ where } X^{(j)} \text{ is categorical}$$

Където $f^{(j)}_{encoding}$ е функцията за кодиране на етикети за j -тата категорийна характеристика, $X^{(j)}$ е оригиналният вектор от категорийни характеристики, а $X^{(j)}_{encoded}$ е кодираният вектор от характеристики. Функцията $f^{(j)}_{encoding}$ за кодиране се дефинира чрез съпоставяне:

$$f^{(j)}_{encoding}: \text{categories of } X^{(j)} \rightarrow 0, 1, 2, \ldots, N_j - 1$$

Където N_j е броят на уникалните категории в $X^{(j)}$.

- Мащабиране на характеристиките: За да се гарантира, че всяка характеристика допринася еднакво за процеса на обучение на модела се пристъпва към нормализиране на данните до общ мащаб, използвайки техниката за мащабиране MinMax. Всяка характеристика се мащабира до диапазона, използвайки формулата:

$$x_{scaled} = \frac{x - min(x)}{max(x) - min(x)}$$

Където $min(x)$ и $max(x)$ са съответно минималната и максималната стойност на характеристиката. Тази трансформация се прилага към всяка характеристика в набора от данни. Параметрите за мащабиране се съхраняват за всяка характеристика, за да се даде възможност за обратна трансформация по време на реконструкцията на данните. Това е от решаващо значение за поддържане на целостта на процесите на анонимизиране на данните.

12.3.2 Конструиране на генератор за анонимизация на данни

Генераторът в генеративна състезателна мрежа (GAN) служи като ключов компонент при синтезирането на данни, които са статистически подобни на оригиналния набор от данни, но не разкриват индивидуалните идентичности. Архитектурата на генератора е проектирана така, че постепенно да учи и да имитира сложните разпределения на входните данни. В следващите редове подробно са описани структурата и функциите за активиране на изкуствената невронна мрежа на генератора. Дизайнът на генератора е от основно значение за гарантиране, че синтезираните данни са едновременно разнообразни и представителни за оригиналното разпределение на данните, което е критичен аспект за успешното анонимизиране на чувствителни информационни масиви.

Генераторът е структуриран като невронна мрежа с директно подаване, напълно свързана, състояща се от множество слоеве с нарастваща сложност, за

да улови сложните модели в данните. Архитектурата е дефинирана както следва:

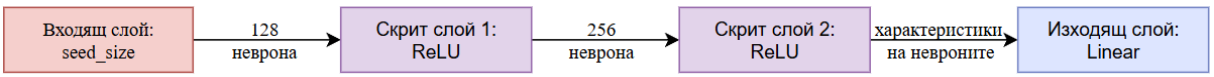
- Входящ слой: Входящият слой получава случаен шумов вектор с размер `seed_size`, който действа като латентно представяне, от което се генерират данни.
- Скрити слоеве: Използват се два скрити слоя за подобряване на способността за обучение на мрежата, като първият слой се състои от 128 неврона, а вторият съдържа 256 неврона.
- Изходящ слой: Изходящият слой има брой неврони, равен на броя на характеристиките в оригиналния набор от данни, което гарантира, че всеки генериран екземпляр има съответстваща стойност за всяка характеристика.
- Изборът на активиращи функции в генератора е от решаващо значение за моделиране на нелинейни сложности:
- Активиране на ReLU: Функцията Rectified Linear Unit (ReLU) се използва в скритите слоеве. Тя въвежда нелинейност в процеса на обучение, позволявайки на модела ефективно да изучава сложни модели. Функцията ReLU се дефинира като:

$$\mathrm{ReLU}(x) = max(0, x)$$

- Линейна активация: Изходящият слой използва линейна активационна функция, която позволява на модела да извежда стойности, които са линейна комбинация от входните му данни. Това е особено полезно за регресионни задачи и се дефинира като:

$$\mathrm{Linear}(x) = x$$

Тези активиращи функции са неразделна част от оформянето на динамиката на обучение на генератора, улеснявайки генерирането на нови точки от данни, които много наподобяват характеристиките на оригиналния набор от данни.



Фиг. 12.3.2-1 Архитектура на генератора

12.3.3 Конструиране на дискриминатор за анонимизация на данни

В контекста на генеративна състезателна мрежа (GAN), дискриминаторът играе критична роля в разграничаването между реални и генерирани данни. Ефективността на GAN зависи до голяма степен от способността на

дискриминатора да идентифицира точно синтетични данни, което от своя страна оказва натиск върху генератора да подобри производството на данни. Тази част от изследването описва подробно архитектурата и функционалността на изкуствената невронна мрежа на дискриминатора. Чрез този структуриран подход, дискриминаторът действа като критичен компонент в рамката на GAN, подобрявайки способността на генератора да произвежда висококачествени анонимизирани данни [12, 13].

Дискриминаторът е проектиран като плътна невронна мрежа с предварителна обратна връзка, която оценява автентичността на йерархията от данни. Архитектурата му е организирана по следния начин:

- Входящ слой. Приема вектор от входни данни с размерност, равна на броя на характеристиките в информационния набор.
- Скрити слоеве. Налични се два скрити слоя, които обработват входните данни, като са въведени и отпадащи слоеве, за да се предотврати преобучаване.
- Изходящ слой. Един неврон със сигмоидна активираща функция за извършване на двоична класификация, като извежда вероятност, която показва вероятността входящите данни да са част от реалната съвкупност.
- Активационни функции и регуларизация:
- Активиране на ReLU. ReLU се използва в скритите слоеве, за да въведе нелинейност, подобрявайки способността на мрежата да улавя сложни модели в данните.

ReLU се дефинира като:

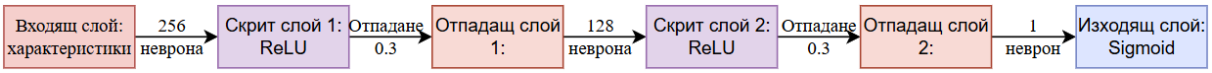
$$\mathrm{ReLU}(x) = max(0, x)$$

- Сигмоидна активация: Изходният слой използва сигмоидна активационна функция, която е идеална за задачи за двоична класификация. Тя компресира изхода в диапазон между 0 и 1, представляващ вероятността входните данни да са реални.

Сигмоидната функция се изразява като:

$$\mathrm{Sigmoid}(x) = \frac{1}{1+e^{-x}}$$

- Регуларизация на отпадането: За да се намали рискът от преобучаване, след всеки скрит слой се включват слоеве с отпадане. Отпадането се извършва произволно, като задава част от входните единици да са равни на 0, при всяка актуализация по време на обучение, което подпомага стабилността на модела. Процентът на отпадане е зададен на 0,3 (30% от невроните се нулират произволно).



Фиг. 12.3.3-1 Архитектура на дискриминатора

12.3.4 Компилация на генеративната състезателна мрежа

След като архитектурите на генератора и дискриминатора са дефинирани, следващата критична стъпка включва компилирането на тези две мрежи в една обща Генеративна състезателна мрежа (GAN). Този дял изяснява интеграцията на генератора и дискриминатора, заедно с обосновката за избора на метрики, функция на загуба и оптимизатор.

Интегриране на генератор и дискриминатор. GAN моделът обединява генератора и дискриминатора, като настройва първият модул да генерира нови данни от шум, които след това се подават директно към дискриминатора. Дискриминаторът оценява дали данните са реални (от действителния набор от данни) или фалшиви (генерирани). Процесът на компилация е проектиран така, че теглата на дискриминатора да са регулируеми при обучението на самия дискриминатор, като същевременно биват замразени при обучението на генератора. Тази настройка гарантира, че генераторът се учи да произвежда по-добри данни, без дискриминаторът да се подобрява в същото време.

Функция на загубата. Бинарната функция на загуба на кръстосана ентропия се използва както за дискриминатора, така и за комбинирания GAN модел. Тази функция на загубата е особено подходяща за бинарни класификационни задачи. Тя измерва производителността на модел, чийто изход е вероятностна стойност между 0 и 1. Загубата се увеличава с отклонението на прогнозираната вероятност от действителния етикет. Функцията се дефинира като:

$$L(y,\hat{y}) = \frac{-1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right]$$

Където y е реалният етикет, $\hat{y}$ е прогнозираната вероятност и N е броят на наблюденията.

Adam оптимизатор. Adam оптимизаторът [14, 17] се използва както за генератора, така и за дискриминатора. Adam е съкращение от Adaptive Moment Estimation (Адаптивна оценка на момента) и комбинира предимствата на две други разширения на стохастичния алгоритъм с градиентно спускане. По-конкретно, той изчислява адаптивните скорости на обучение за всеки параметър. Adam е ефективен по отношение на изчислителни усилия и има малки изисквания за памет. Този метод е устойчив и не подлежи на промяна спрямо диагоналното премащабиране на градиентите и е подходящ за

значителни проблеми по отношение на данни и/или параметри. Оптимизаторът Adam е разширение на стохастичния градиентен спускаем алгоритъм, което получава по-широко приложение при дълбоко обучение. Adam комбинира предимствата на две други разширения на стохастичния градиентен спускаем алгоритъм: Adaptive Gradient Algorithm (AdaGrad) и Root Mean Square Propagation (RMSProp). Adam изчислява експоненциална пълзяща средна на градиента и квадрата на градиента, както и за параметрите β_1 и β_2, контролирайки скоростите на затихване на тези пълзящи средни. Правилата за актуализиране за Adam са както следва:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$$
$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$$
$$\widehat{m}_t = \frac{m_t}{1-\beta_1^t}$$
$$\widehat{v}_t = \frac{v_t}{1-\beta_2^t}$$
$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\widehat{v}_t} + \epsilon} \widehat{m}_t$$

Където:

g_tе целевият градиент спрямо параметрите θ във времевата стъпка t,

m_t и v_t са оценки на първия момент (средната стойност) и втория момент (нецентрираната дисперсия) на градиентите, съответно, правилно класифицирани прогнози

$\widehat{m}_t$ и $\widehat{v}_t$ са коригирани с отклонение версии на m_t и v_t,

η е скоростта на обучение,

ϵ е малък скалар, използван за предотвратяване на деление на нула в имплементацията.

Параметрите β_1 и β_2 обикновено се задават близо до 1.0 при проблеми с разреден градиент (напр. проблеми с NLP и компютърно зрение). Версиите на m_t и v_t, коригирани за отклонение, предоставят по-добри оценки на моментите от по-нисък порядък и са особено важни по време на началните времеви стъпки и когато скоростите на затихване са зададени на стойности близки до 1.

Метрика за точност. Индикация за това, до каква степен дискриминаторът е обучен да прави разлика между реални и генерирани информационни масиви. Точността е един от най-ясните и широко използвани показатели за ефективност при класификационни задачи, включително при обучението на дискриминативни модели в рамките на генеративни

състезателни мрежи (GAN). Точността измерва дела на коректните резултати (както истински положителни, така и истински отрицателни) сред общия брой изследвани случаи. За дискриминатора в GAN, който класифицира входните данни като реални или фалшиви, точността може да бъде математически дефинирана като:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} = \frac{TP + TN}{TP + TN + FP + FN}$$

Където:

-TP (Коректни положителни резултати) е броят на правилните предвиждания, че даден случай е реален,

TN (Коректни отрицателни) е броят на правилните предвиждания, че даден случай е фалшив,

FP (Грешни положителни резултати) е броят на неправилните прогнози, че даден случай е реален,

FN (Грешни отрицателни) е броят на неправилните прогнози, че даден случай е фалшив.

12.3.5 Обучение на генеративната състезателна мрежа

Процесът на обучение на генеративната състезателна мрежа (GAN) включва едновременно обучение на два модела: генератор и дискриминатор. Генераторът има за цел да създава данни, които изглеждат възможно най-близки до реалните, докато дискриминаторът се опитва да различи реалните данни от тези, генерирани от генеративната функция. Тази част описва подробно параметрите и итеративния процес на обучение на GAN, което е от решаващо значение за разбирането на динамиката и конвергенцията на състезателното обучение.

Преди започване на процеса на обучение се определят няколко ключови параметъра:

- Размер на извадката. Размерността на входния шумов вектор за генератора. Той определя количеството случайни събития, въведено в процеса на генериране на данни. Тази стойност е произволна и може да се коригира въз основа на специфични характеристики на данните или желаната сложност на генератора.
- Брой характеристики. Този параметър представлява размерността на изходните данни от генератора, която трябва да съответства на броя характеристики в реалния набор от данни.

- Епохи и размер на партидата. Броят на пълните преминавания през набора от данни за обучение (епохи) и броят на обработените проби преди актуализиране на модела (размер на партидата) са критични хиперпараметри за обучението. Тези стойности са избрани, за да се балансира между скоростта на обучение и качеството на конвергенция на модела.

Обучителният цикъл на GAN редува обучението на дискриминатора и генератора:

Обучение на дискриминатора. Дискриминаторът се обучава върху партида реални данни с етикети, зададени с единица (което показва реални данни). Едновременно с това, дискриминаторът се обучава върху партида фалшиви данни, генерирани от генератора, с етикети, зададени с нула (което показва фалшиви данни). Загубите от тези две стъпки се сумират, за да се актуализира дискриминаторът. Математически, загубата на дискриминатора, за реални и фалшиви данни, се определя като:

$$L_D = \frac{-1}{2}\left[\log\left(D(x)\right) + \log\left(1 - D\left(G(z)\right)\right)\right]$$

Където $D(x)$ е оценката на дискриминатора за вероятността данните x да са реални, а $D\left(G(z)\right)$ е вероятността генерираната извадка $G(z)$ да е реална.

Обучение на генератора. Генераторът е обучен да заблуждава дискриминатора, като се целта е да бъдат генерирани данни, които дискриминаторът да класифицира като реални. Загубата на генератора се изчислява въз основа на това колко добре той заблуждава дискриминатора:

$$L_G = -\log\left(D\left(G(z)\right)\right)$$

Генераторът подобрява своите тегла, въз основа на това дали дискриминатора е заблуден от генерираните извадки.



Фиг. 12.3.5-1 Процес на обучение на GAN

12.3.6 Генериране на данни посредством обучена GAN

След като Генеративната състезателна мрежа (GAN) е адекватно обучена, генераторът може да се използва за синтезиране на нови данни. Този дял обяснява процеса на генериране на данни с помощта на обучения генератор на GAN, включително генериране на случаен шум, последваща обработка за подходящо мащабиране на данните и преобразуването им в използваем формат. Обсъждат се също стъпки, посредством които се гарантира, че данните се придържат към оригиналните разпределения, посредством техники като winsorization и обратна трансформация на кодировките на етикетите.

Генераторът започва с произволен зашумен вектор като вход. Този вектор служи като начален елемент, от който генераторът създава нови екземпляри от данни. Процесът може да бъде дефиниран както следва:

- Брой проби: $N = 1500$

  Където N е броят на генерираните инстанции от данни.
- Размер на извадката: $S = 100$

  Съвпадение с входната размерност на генератора, която е размерът на вектора на случаен шум.

Случайните шумови вектори се генерират с помощта на нормално разпределение, центрирано в нула, със стандартно отклонение от единица:

$$z \sim N(0, I)$$

Където z е векторът на шума, а I е единичната матрица (показателно за стандартно нормално разпределение).

12.3.7 Последваща обработка на данни, генерирани с GAN

След генериране и първоначална обработка на данните, с цел съобразяване с характеристиките на оригиналния масив, може да са необходими допълнителни уточнения, за практическо приложение или анализ. Това включва коригиране на типа данни на всяка характеристика, за да съответства на предвидения формат, преобразуване на типовете с плаваща запетая в цели числа, където е приложимо, като се избягва променя на колони, които не са цели числа. Настоящата част обяснява стъпките, свързани с прецизиране на типовете данни по отношение на генерираните характеристики.

След като GAN генерира данни, в общия случай, се изисква последваща обработка, за да се гарантира, че те съответстват на мащаба и разпределението на оригиналния набор:

- Обратно мащабиране. В случай че оригиналните данни са мащабирани до диапазон с помощта на `MinMaxScaler`, се прилага обратна трансформация, за да се преобразуват генерираните мащабирани данни обратно в оригиналния мащаб. Това е от решаващо значение за уеднаквяване на синтетичните данни с мащаба на действителните:

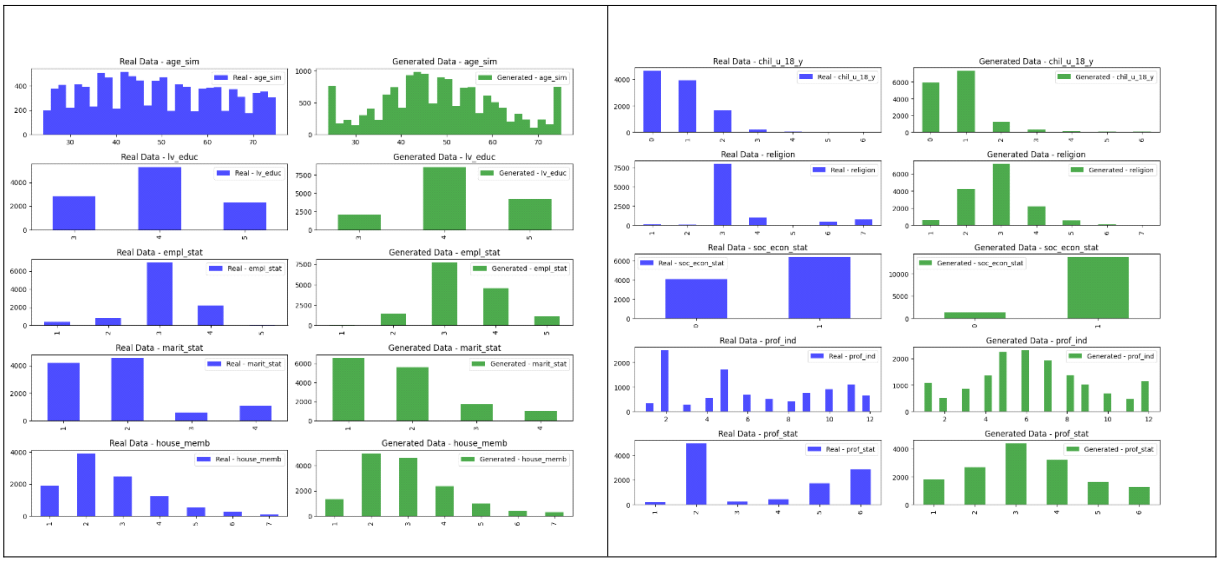
$$X = X_{scaled} \cdot (max(X) - min(X)) + min(X)$$

Тази формула връща мащабираните данни X_{scaled} обратно към първоначалния им мащаб, използвайки минималните ($min(X)$) и максималните ($max(X)$) стойности, съхранени по време на първоначалното мащабиране.

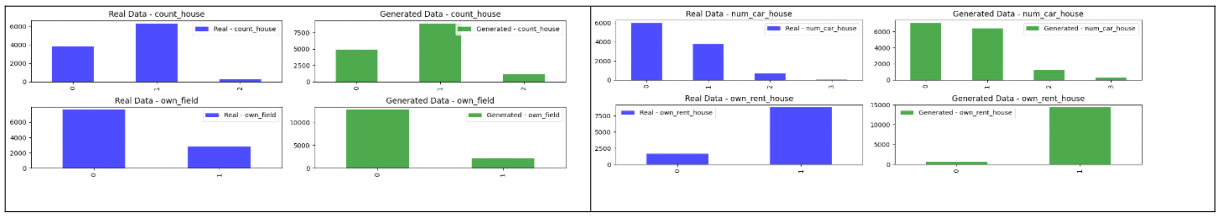
- Winsorization. За да се гарантира допълнително, че данните се придържат към оригиналните разпределения без екстремни отклонения, се прилага winsorization. Този метод ограничава екстремните стойности в набора от данни, за да намали влиянието на отклоненията;
- Обратно кодиране на етикети. В случай че категорийните характеристики са кодирани като цели числа, се прилага обратна трансформация, за да се преобразуват тези цели числа обратно в оригиналните категорийни данни;
- Закръгляване и преобразуване на типа на колоните. За да се гарантира, че генерираните данни отговарят на очакваните типове данни, особено за характеристики, които първоначално са имали целочислени стойности, в целия набор от данни се прилага персонализираща функция. Тази функция проверява всяка променлива спрямо списък с колони, които не трябва да се преобразуват в

цели числа (т.е. колони, които не са целочислени), като се обработва всяка колона по подходящ начин;

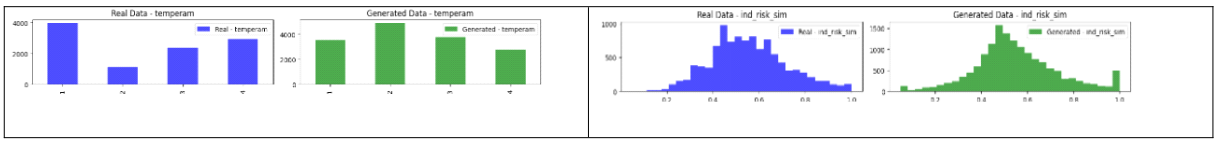
- Колони, съдържащи целочислени стойности. Тези колони се закръгляват до най-близкото цяло число и след това се преобразуват от тип данни с плаваща запетая, в целочислен тип данни. Тази стъпка е от решаващо значение, защото GAN обикновено генерират изходи като числа с плаваща запетая, което може да не е подходящо за оригинални набори от данни с целочислени ограничения.
- Нецелочислени колони: Колоните, които са идентифицирани като нецелочислени, остават непроменени, за да се запази целостта на данните им, като например непрекъснати променливи или колони, изискващи представяне с плаваща запетая.



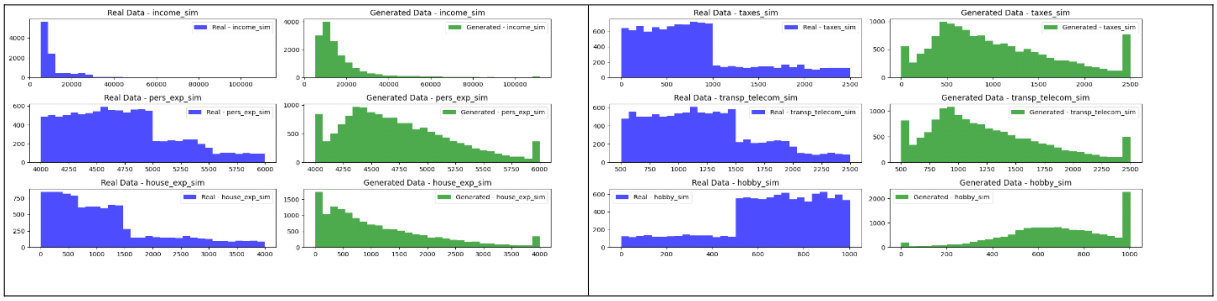
Фиг. 12.3.7-1 Резултати по демографски променливи – реални данни и генерирани данни за всяка променлива



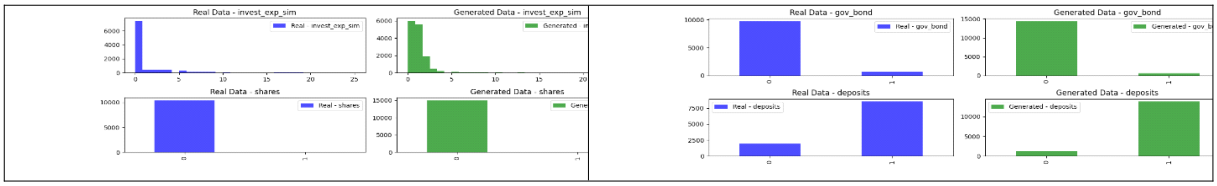
Фиг. 12.3.7-2 Резултати по променливи за икономически статус – реални данни и генерирани данни за всяка променлива



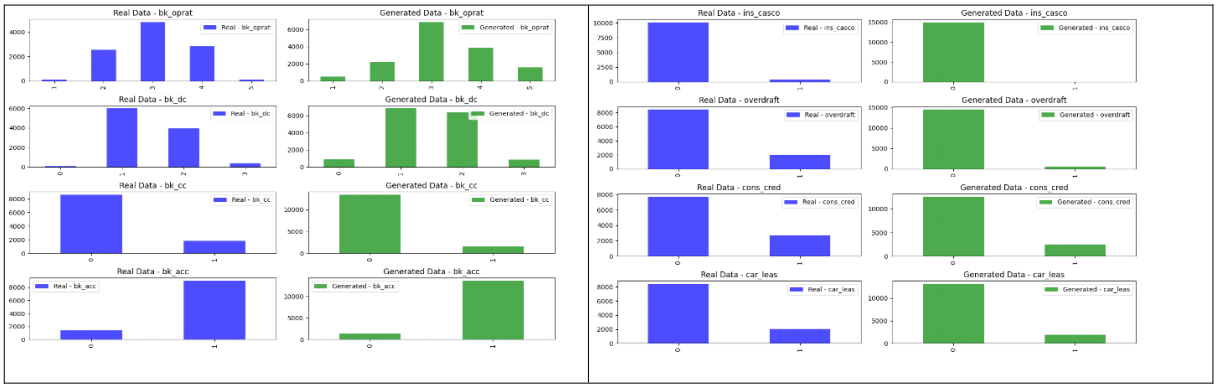
Фиг. 12.3.7-3 Резултати по психологически променливи – реални данни и генерирани данни за всяка променлива



Фиг. 12.3.7-4 Резултати по променливи за лични разходи – реални данни и генерирани данни за всяка променлива



Фиг. 12.3.7-5 Резултати по променливи за инвестиционни избори – реални данни и генерирани данни за всяка променлива



Фиг. 12.3.7-6 Резултати по променливи за лични разходи – реални данни и генерирани данни за всяка променлива

12.3.8 Оценка на косинусовото подобие на данни, генерирани от GAN

В изследването за приложение на генеративни състезателни мрежи (GAN) в процеса по анонимизиране на данни, критично използвана метрика за оценка е косинусовото подобие между средните вектори на реални и генерирани набори от данни. Тази мярка предоставя количествена оценка на сходството в пространството на характеристиките между два набора от данни, които по своята същност могат да притежават различен брой записи. Косинусовото сходство се изчислява, както следва:

Нека $(v_1)u(v_2)$ са средните вектори на реалния и генерирания набор от данни. Средният вектор на набор от данни се изчислява, като се вземе средноаритметичната стойност на всяка характеристика във всички записи:

$$v_k = \frac{1}{n_k} \sum x_i$$

Където (n_1) и (n_2) са съответно броят на записите в реалния и генерирания набор от данни $(k=2)$, а (x_i, y_i) представляват отделни записи в тези набори от данни. Въпреки разликата в броя на записите $(n_1 \neq n_2)$, векторите на средните стойности (v_1) и (v_2) стандартизират това несъответствие чрез осредняване на характеристиките, като по този начин позволяват смислено сравнение.

Косинусовото сходство (S_c) между тези средни вектори се изчислява по формулата:

$$S_c(v_1, v_2) = \frac{v_1 \cdot v_2}{\|v_1\| \|v_2\|}$$

Където $(v_1 \cdot v_2)$ означава скаларното произведение на векторите, а $(\|v_1\|)$ означава нормата на вектор (v_1). Това съотношение измерва косинуса на ъгъла между двата вектора, предоставяйки индикация за тяхното насочено подравняване, независимо от техните величини или несъответствията в размера на наборите от данни.

В настоящия анализ, коефициентът на сходство улеснява разбирането на това, до каква степен генерираните данни имитират общото разпределение на характеристиките на реалните данни, приемайки стойност между -1 и 1, където 1 показва перфектно уеднаквяване на данните, 0 показва ортогоналност, а -1 показва диаметрално противоположност.

Както за мащабирани, така и за немащабирани данни, тези мерки за сходство се изчисляват, за да се гарантира, че трансформациите на мащабиране не влияят прекомерно на сравнителния анализ. По този начин, този показател служи като надеждна, мащабно-устойчива мярка за сходство в многомерно пространство от характеристики, което е от решаващо значение за оценка на ефективността на техниките за анонимизиране на данни, внедрени посредством GAN.

Косинусово подобие между двата набора от данни:

- за мащабирани данни: $S_c(v_1, v_2) = 0.9730251885118525$
- за немащабирани данни: $S_c(v_1, v_2) = 0.9917135244748599$
- за демографски данни: $S_c(v_1, v_2) = 0.9999897188621633$
- за психологически данни: $S_c(v_1, v_2) = 1.0$
- за данни за лични предпочитания: $S_c(v_1, v_2) = 1.0$
- за данни за лични разходи: $S_c(v_1, v_2) = 0.9917141602816288$

Разгледаният пример обобщава същността на представеното методологично сравнение, като подчертава прилагането на косинусовото подобие като фундаментален показател при количествената оценка на качеството на синтетичните данни спрямо техния автентичен еквивалент.

12.4 Заключение и бъдеща работа

Разглежданият процес за генериране на данни, посредством генеративни състезателни мрежи, утвърждава, че всички характеристики в набора от данни отговарят на съответните им изисквания за тип данни. Чрез прилагане на персонализирани стъпки за последваща обработка, се осигурява не само статистическото подобие на синтетичните данни спрямо оригиналните данни, но и това, че съвпадат по отношение на формата на данните и използваемостта. По този начин се повишава надеждността и приложимостта на генерираните данни в реални сценарии. Този цялостен подход не само улеснява генерирането на реалистични и статистически подобни данни, но също така гарантира, че синтетичните данни са съвместими с форматите и разпределенията на оригиналния набор от данни. Това е от съществено значение за приложения, изискващи анонимизиране на данни, където поддържането на статистическата им цялост е от решаващо значение. Обсъжданите техники гарантират, че генерираните от GAN данни могат да бъдат директно използвани за аналитични цели или за допълнителни задачи за машинно обучение без допълнителна предварителна обработка.

В предишните радели авторите демонстрират набор от различни подходи за анонимизиране и създаване на данни, изпълнявайки общата цел на това изследване за намиране, чрез експериментално изследване, на най-добрата методология за синтез на данни.

Използвана литература

[1] Marchev, V., Marchev, A., Jr. (2024). Anonymizing Personal Information Using Distribution-based Data Synthesis XXII INTERNATIONAL SCIENTIFIC CONFERENCE “MANAGEMENT AND ENGINEERING’24”, Sozopol, Bulgaria.

[2] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs trained by a two-time-scale update rule converge to a local Nash equilibrium. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), Advances in Neural Information Processing Systems 30 (pp. 6626–6637). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

[3] Marchev, A., Jr, Marchev, V., Efremov, A. (2024). Methodology for Cold-Start Modeling Using Synthetic Data, AIP conference proceedings.
[4] Marchev, A., Jr, Marchev, V. (2022). SYNTHESIZING MULTI-DIMENSIONAL PERSONAL DATA SETS. AIP CONFERENCE PROCEEDINGS; 2505 (1): 020012
[5] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. (2014). Generative Adversarial Nets, Advances in Neural Information Processing Systems 27 (NIPS 2014).
[6] Ko, D., Ul Hassan, A., Suk, J., Choi, J. (2020). Korean Font Synthesis with GANs, International Journal of Computer Theory and Engineering vol. 12, no. 4, pp. 92-96.
[7] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X. (2016). Improved Techniques for Training GANs, Neural Information Processing Systems, proceedings.neurips.cc.
[8] Radford, A., Metz, L., Chintala, S., (2015). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv preprint arXiv:1511.06434.
[9] Ho, J., Stefano, E., (2016). "Generative Adversarial Imitation Learning". Advances in Neural Information Processing Systems. 29: 4565–4573. arXiv:1606.03476
[10] Goodfellow, I. (2017). "NIPS 2016 Tutorial: Generative Adversarial Networks". arXiv:1701.00160
[11] Zhang, H., Goodfellow, I., Metaxas, D., Odena, A. (2019). "Self-Attention Generative Adversarial Networks". International Conference on Machine Learning. PMLR: 7354–7363.
[12] Miyato, T., Koyama, M. (2018). cGANs with projection discriminator. In ICLR, 2018
[13] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A. (2017). Improved Training of Wasserstein GANs, n NIPS
[14] Kingma, D., Ba, J., L. (2015). Adam: A method for stochastic optimization, In ICLR, 2015
[15] Yoon, J., Drumright, L., van der Schaar, M. (2020). Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN), PMID: 32167919 DOI: 10.1109/JBHI.2020.2980262
[16] Venugopal, R., Shafqat, N., Venugopal, I., Tillbury, B., Stafford, H., Bourazeri, A. (2022). Privacy preserving Generative Adversarial Networks to model Electronic Health Records, PMID: 35779443 DOI: 10.1016/j.neunet.2022.06.022

[17] YiJaehyun, D., Ji, A. (2020). An Effective Optimization Method for Machine Learning Based on ADAM, Applied Sciences 10(3):1073, DOI: 10.3390/app10031073

[18] Astolfi, G. (2021). Generating tabular data using generative adversarial networks with differential privacy, CONFERENCE OF EUROPEAN STATISTICIANS, Expert Meeting on Statistical Data Confidentiality

13 Генериране на данни чрез съвместно разпределение

Автори: Ангел Марчев, мл., Димитър Любчев, Васил Марчев

Резюме. Тази глава разглежда метод за генериране на данни, основан на реални масиви с чувствителна информация, като изследва нов подход - обърната копула за многомерен синтез на данни. Използвайки копули, които са мощни инструменти за моделиране на зависимости между променливи, разглежданият метод генерира синтетични данни, които поддържат сложни взаимозависимости. В главата е демонстрирана ефективността на този подход посредством различни експерименти и казуси, показвайки висока прецизност при запазване на зависимостите и постигане на малки несъответствия в пределните разпределения. Устойчивостта на метода се валидира чрез сравнителен анализ и статистически проверки, включително теста на Колмогоров-Смирнов. Въвежда се гъвкав и ефикасен метод за генериране на синтетични данни, приложим за широк спектър от разпределения и практически приложения. Бъдещата работа ще проучи приложението на други видове копули и по-нататъшното усъвършенстване на метода, за да се подобри неговата гъвкавост и приложимост.

Ключови думи: синтетични данни, копула, обърнато генериране на извадки, съвместно разпределение

13.1 Въведение

В ерата на големите данни, търсенето на значими количества разнообразни и представителни набори от данни се е увеличава рязко в различни изследвани области: от здравеопазването и финансите, до изкуствения интелект и машинното обучение. Като обещаващ подход за задоволяване на това търсене се установява използването на синтетични данни, които се отнасят до масиви, които не съществуват в реалния свят, а се генерират чрез алгоритми, които включват статистически зависимости. Синтетичните данни предлагат няколко предимства, като например липсата на опасения за поверителност и регулаторни ограничения, като същевременно запазват взаимовръзките между променливите. Този аспект е особено важен в чувствителни области като финансовите услуги, където поверителността на данните е от първостепенно значение. Освен това, синтетичните данни са икономически ефективни и спестяват време, елиминирайки необходимостта от обширни усилия за събиране и обработка на данни. Към настоящия момент в литературата липсва цялостно проучване на сложни методи за генериране на

висококачествени синтетични данни, които точно представят демографски, финансови, психологически и данни за социален статус. Настоящата глава има за цел да запълни тази празнина, като проучи подхода за обърнато генериране на извадки посредством копула, които се използват като подход за многомерен синтез на данни.

За да се отговори на този изследователски въпрос е предложен метод, който използва копули – мощни статистически инструменти, моделиращи сложни зависимости между променливи. Чрез използване на метода за обратно копулно извличане на извадки може да се генерира набор от синтетични данни, които поддържат сложните взаимозависимости, присъстващи в многомерните информационни масиви. Разглежданият подход се възползва от усъвършенстван набор от данни и метод, който надхвърля традиционните техники за генериране на информация. Методът, базиран на копули, позволява гъвкаво и надеждно моделиране на зависимости, което го прави подходящ за широк спектър от приложения и разпределения на данни.

Разглежданите резултати показват, че подходът за подбор на извадка, основан на обратна копула, ефективно генерира висококачествени синтетични данни, които запазват зависимостите, наблюдавани в оригиналния набор. Методът е валидиран чрез различни експерименти и казуси, показвайки висока прецизност при поддържане на зависимостите между характеристиките, както и незначителни несъответствия в определени гранични разпределения. Тези резултати подкрепят издигнатата хипотеза, че генерирането на синтетични данни, базирани на копула, е мощен инструмент за създаване на представителни набори от данни.

Проведена е серия от проверки за устойчивост и разширения, с цел допълнителна валидация на описания подход. Сравнителният анализ с традиционните методи за генериране на данни показва висока производителност при поддържане на зависимостите на данните. Статистическите проверки, включително теста на Колмогоров-Смирнов, потвърждават високото качество на синтетичните данни. Проведени са изследвания по отношение на различните семейства копули, за да се подобри приложимостта на метода към различни разпределения на данни, осигурявайки пълно покритие и надеждност.

Изследването се основава на няколко направления в литературата в областта на генерирането на синтетични данни и теорията на копулите. Най-подходящите трудове включват изследвания върху вероятностни методи за синтез на данни, като тези на Роджър Нелсен [1] и Торстен Шмит [2]. В процеса на работа методът се позовава и на работата на Ян-Фредерик Май и Матиас Шерер [3] върху симулирането на копулите и последните постижения в

непараметричното моделиране на копулите от Субхадип Мукхопадхяй и Емануел Парзен [4], Хуан П. Рестрепо и колектив [5]. Приносът на настоящата работа надхвърля тези трудове, като прилага метода за обърнат подбор на извадки посредством копула, за генериране на синтетични многомерни данни, демонстрирайки неговите практически приложения и валидирайки неговата ефективност чрез строги тестове.

Настоящата разработка има няколко ключови приноса в областта на генерирането на синтетични данни:

1. Демонстрира ефективността на метода на обърната копула за подбор на извадки при запазване на сложни зависимости между променливите.

2. Валидира подхода чрез експерименти и проверки за устойчивост, включително сравнителен анализ и статистически тестове.

3. Сравнява целесъобразността на използването на няколко семейства копули, за предварително зададен набор от разпределения.

Частите на настоящата глава са структурирани както следва: част 1 предоставя по-задълбочено разбиране на синтеза на данни. Част 2 предоставя подробно обяснение на теоретичните основи на копулите и обратния процес на подбиране на извадки. Част 3 описва методологията и експерименталната настройка, използвани за валидиране на нашия подход. В част 4 се представя алгоритъма на предложения метод, докато в 5 са показани резултатите от направените експерименти и проверки за устойчивост. В последния дял се обсъждат последиците от направените открития и потенциалната бъдеща работа в тази област.

13.2 Мотивация

В ерата на големите данни, търсенето на значими количества разнообразни и представителни информационни масиви, в различни области, се увеличава с мащабни темпове: от здравеопазването и финансите, до изкуствения интелект и машинното обучение. Като обещаващ подход за посрещане на необходимостите, генерирани от това търсене, е използването на синтетични данни. Тези информационни масиви не съществуват в реалния свят, а биват генерирани посредством алгоритми, които включват статистически зависимости.

Съществуват няколко метода за генериране на синтетични данни [6, 7, 8, 9], които се разделят на вероятностни подходи и такива, които не са базирани на вероятности. Вероятностните методи разчитат на статистически модели за генериране на данни въз основа на вероятностните разпределения на реални данни. За разлика от тях, подходите, които не се основават на вероятности

използват техники като алгоритми, базирани на правила, или модели за машинно обучение, които не разчитат изрично на вероятностните разпределения. Изборът на метод зависи от специфичните изисквания, в зависимост от областта на приложение на данните. Всеки един от изброените методи има предимства и недостатъци.

В тази глава се изследва нов подход към генерирането на синтетични данни: обратно подбиране на извадки посредством копула, използвани за многомерен синтез на данни. Копулите представляват мощни статистически инструменти, които позволяват моделиране на сложни зависимости между променливи. Чрез използването на копулите, предложеният в настоящия труд метод има за цел да генерира висококачествени синтетични данни, които запазват сложните взаимозависимости, присъстващи в многомерните набори от данни. Следващите стъпки от изследването обхващат задълбочаване в теоретичните основи на копулите, подробно описание на процеса на обратно подбиране на извадки и нагледна демонстрация на ефективността на гореописания подход, посредством набор от експерименти и казуси.

13.3 Дефиниция на проблема

Генерирането на синтетични данни, които запазват първоначалните взаимозависимости между множество променливи е критично предизвикателство в науката за данни. Настоящата част разглежда проблемите с генерирането на синтетични многомерни данни, които поддържат сложните зависимости, наблюдавани в оригиналния набор от данни. По-конкретно, за постигане на посочената цел се разглежда възможността за използване на обратна копула.

Като се има предвид набор от данни с n наблюдения и d променливи, целта е да се генерират синтетични данни, които точно имитират съвместното разпределение на тези променливи. В разглеждания метод за моделиране на зависимостите между променливите се използва копулната функция, която описва съвместното разпределение на набор от случайни променливи. Тази глава представя един възможен подход за получаване на съвместно разпределение, използвайки копулата, като съществуват множество други методи за извличане на копулата от данните, които остават извън обхвата на настоящото изследване.

Копулата е функция $C:[0,1]^d \rightarrow [0,1]$, която съпоставя многомерни функции на разпределение с техните едномерни гранични функции на разпределение. В контекста на моделиране на зависимости между променливи,

копулите предоставят гъвкав начин за описание на съвместното разпределение на множество случайни променливи.

В случай че е налично многовариантното съвместно разпределение, представено от копулната функция, процесът за генериране на синтетични данни използва метода за обратно вземане на проби посредством копула. Ключов хиперпараметър на решението е видът копула, който ще се използва, въз основа на определени критерии, например информационен критерий на Акаике (AIC) или информационен критерий на Байес (BIC). Няколко семейства копули се използват често в статистическото моделиране и финансите, с цел въвеждане на зависимости между случайни променливи. Следващите параграфи разглеждат няколко широко използвани копули (списъкът е неизчерпателен):

- Гаусова копула [10, 11] е симетрична и улавя линейна зависимост между променливите. Полезна е за моделиране на широк спектър от зависимости поради своята гъвкавост. Кумулативна функция на разпределение (CDF):

$$C(u,v;\Sigma)=\Phi_{\Sigma}\left(\Phi^{-1}(u),\Phi^{-1}(v)\right)$$

Където Φ е CDF на стандартното нормално разпределение, Φ^{-1} е неговата обратна функция, а Φ_{Σ} е CDF на многовариантното нормално разпределение с корелационна матрица Σ.

- Клейтън Копула [12, 13, 14] представлява асиметрична копула и може да моделира зависимост от долната опашка. Полезна е за моделиране на сценарии, при които екстремните стойности в една променлива са свързани с екстремни стойности в друга променлива. Кумулативна функция на разпределение (CDF):

$$C(u,v;\theta)=\left(u^{-\theta}+v^{-\theta}-1\right)^{\frac{-1}{\theta}},\theta>0$$

- Копулата на Гъмбел [15, 16] е асиметрична и може да моделира зависимостта от горната опашка. Полезна е за сценарии, при които високите стойности в една променлива са свързани с високи стойности в друга променлива. Кумулативна функция на разпределение (CDF):

$$C(u,v;\theta)=\exp\left[-\left((-\log u)^{\theta}+(-\log v)^{\theta}\right)^{\frac{1}{\theta}}\right],\theta\geq 1$$

- Франк Копула [13, 14, 17] е симетрична и може да моделира както положителни, така и отрицателни зависимости. Не показва опашна зависимост. Кумулативна функция на разпределение (CDF):

$$C(u,v;\theta)=\frac{-1}{\theta}\log\left(1+\frac{\left(e^{-\theta u}-1\right)\left(e^{-\theta v}-1\right)}{e^{-\theta}-1}\right),\theta\neq 0$$

- Копула на Стюдънт, t-копула [13, 18]. Подобна на Гаусовата копула, но с по-тежки опашки, улавяща екстремни съпътстващи движения. Полезна за финансови приложения, където екстремни събития се случват по-често, отколкото предсказва нормалното разпределение. Кумулативна функция на разпределение (CDF):

$$C(u,v;\nu,\Sigma)=t_{\nu,\Sigma}\left(t_\nu^{-1}(u),t_\nu^{-1}(v)\right)$$

Където t_ν^{-1} е обратната функция на разпределението на Стюдънт с ν степени на свобода, а Σ е корелационната матрица.

Основното предимство на метода на обратна копула за семплиране пред традиционните методи, като например разлагането на Холецки, е че той не предполага нормално разпределение за характеристиките. Тази гъвкавост позволява методът да бъде приложим към по-широк диапазон от разпределения на данни, което го прави надежден инструмент за генериране на синтетични данни.

13.4 Алгоритъм за синтез на данни - обърната копула за подбор на извадки

Методът за подбор на извадки, посредством обърната копула, улавя зависимостите между променливите и позволява създаването на разпределения, които ефективно моделират корелирани многовариантни данни. Алгоритъмът конструира многовариантно разпределение чрез оптимално определяне на едновариантни разпределения. Използвайки копула, се установява корелационната структура между набор от променливи. Този подход поддържа използването на двувариантни разпределения, както и разпределения от по-високи измерения.

Ключово свойство на копулната функция, използвана в този алгоритъм, включва напасването ѝ към многовариантното разпределение и след това прилагане на обратната функция на копулата [1, 19, 20, 21]. Получените стойности се нормализират между нула и единица, което налага денормализация чрез мащабиране на параметри. Следващият етап включва

конкатениране на матрицата с денормализираните данни. Като последна стъпка в процеса се разглежда фазата на оценка.

Качеството на синтетичните данни се оценява с помощта на различни показатели за оценка, например статистически мерки за разстояние, като теста на Колмогоров-Смирнов [22, 23], визуално сравнение с помощта на точкови диаграми и сравнение на разпределенията с помощта на хистограми. В настоящата методология се използва едноизвадковият тест на Колмогоров-Смирнов, тъй като е най-строгият статистически тест за сравнение на разпределенията. Двуизвадковият KS тест сравнява ECDF на две независими извадки, за да определи дали те са извлечени от една и съща популация, като остава извън обхвата на настоящата глава.

1) Дадено е, че $\Phi = f_1, \ldots f_s(\omega_1 \ldots \omega_s)$ е многовариантно съвместно разпределение на m крайни случайни променливи от реални данни ω, обозначаващ m брой количествени характеристики на изследвана система S. В този случай кумулативната функция на разпределение се дефинира като $\Phi = P(\omega_1 \le \hat{\omega}_1, \omega_2 \le \hat{\omega}_2, \ldots, \omega_s \le \hat{\omega}_s)$, където P обозначава вероятността всички случайни променливи ω_i да са по-малки или равни на съответните стойности $\hat{\omega}_i$.

2) Напасване на копулата – оценка на параметрите за избраната копула, използвайки метода за оценка на максималното правдоподобие (MLE). Като се има предвид наборът от данни $X = \{x_1, x_2, \ldots, x_n\}$, където всяко x_i е d-мерно наблюдение, се напасва копула C, като се оценяват параметрите θ, които максимизират функцията на правдоподобие:

$$L(\theta; X) = \prod_{i=1}^{n} c_\theta\left(F_1(x_{i1}), \ldots, F_d(x_{id})\right)$$

Където c_θ е функцията на плътност на C, а F_j са граничните кумулативни разпределения на генерираните променливи.

3) Използва се копула C за генериране на симулирани нормализирани примерни вектори $(x_1, x_2, \ldots, x_m)$ с дължина n от Φ като:

$$C(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_s) = \Phi\left[\omega_1 \le f_1^{-1}(\bar{x}_1), \omega_2 \le f_2^{-1}(\bar{x}_2), \ldots, \omega_s \le f_s^{-1}(\bar{x}_s)\right]$$

Използвайки модела на напасната копула са генерирани синтетични данни чрез вземане на извадки от копулата и след това се прилага обратната функция на граничните кумулативни функции на разпределение. За всяко измерение j са генерирани извадки U_j от равномерно разпределение и след това

са трансформирани, като се използва обратната функция на граничните разпределения F_j^{-1}:

$$X_j = F_j^{-1}(U_j)$$

4) Преобразуват се нормализираните вектори \(u\) в мащаба на оригиналните данни, като се използва обратната функция на кумулативните функции на разпределение (CDF) на оригиналните променливи. Дадено за x_s. Нека \($u=(u_1,u_2,\ldots,u_d)$\) е симулираният вектор на извадката, където всеки \(u_i\) е в интервала \($[0,1]$\).

$$x_s = u_s . \gamma_{s_2} + \gamma_{s_1}$$

Където γ_{s_1}и γ_{s_2}са параметри на мащабиране

5) Конкатениране x_m в матрица $\widetilde{X}$: $\widetilde{X}_{n,s}=[x_1 \ x_2 \ \ldots \ x_s]$. Резултатът представлява синтетичен набор от данни \($X' = \{x_1', x_2', \ldots, x_m'\}$\), който запазва структурата на зависимостите на оригиналните данни.

6) За оценка се използва едноизвадковият тест на Колмогоров-Смирнов (КС). Това е статистически тест, използван за сравняване на разпределението на емпирична извадка с референтно теоретично разпределение. В едноизвадковия КС тест се сравнява емпиричната функция, на разпределение $F_n(x)$, на извадката с кумулативната функция на разпределение $F(x)$, на референтното разпределение.

Емпиричната функция на разпределение $F_n(x)$ се определя като:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{[X_i \le x]}$$

Където $I_{[X_i \le x]}$ е индикаторна функция, която е равна на 1, ако $X_i \le x$ и равна 0, в противен случай. Размерът на извадката е n.

Тестовата статистика за еднопробния KS тест е максималната абсолютна разлика между емпиричната функция на разпределение $F_n(x)$ и кумулативната функция на разпределение $F(x)$:

$$D_n = \text{¿} x \left| F_n(x) - F(x) \right|$$

Нулевата хипотеза (H_0) за KS теста е, че синтезираните данни са извлечени от едно и също теоретично разпределение. Стойността на α показва значимостта на вероятността за отхвърляне на нулевата хипотеза, когато тя е

вярна. Критичната стойност D_α за еднопробния KS тест при ниво на значимост α е дадена от:

$$D_\alpha = \sqrt{\frac{-1}{2n}\ln\left(\frac{\alpha}{2}\right)}$$

В случай че $D_n > D_\alpha$, тогава се отхвърля нулевата хипотеза H_0 на ниво на значимост α и се заключава, че извадката не произлиза от референтното разпределение (еднопробен KS тест)

13.5 Емпиричен пример

В числения пример по-долу се осъществява апробация и илюстрация на описания алгоритъм. За демонстрация на процеса на генериране на синтетични данни се следва процедура от стъпки, която гарантира, че генерираните данни запазват желаната структура на зависимост и следват специфични гранични разпределения. В настоящия дял използваното семейство копула е Гаусова копула.

Стъпка 1: Генериране на данни с помощта на Гаусова копула

Основната цел на тази стъпка е да се създаде многовариантен набор от данни, който показва специфична структура на зависимости или корелации между променливите. За постигането на тази цел се използва Гаусовата копула.

Генерираните данни следват многовариантно нормално разпределение, характеризиращо се със зададена средна стойност и ковариационна матрица. Това гарантира, че данните обхващат предвидените зависимости между променливите. Съвместната диаграма на разсейване на трите синтезирани променливи (както е показано на фиг. 13.5-3) визуално потвърждава зависимостите, наложени от копулата.

В този смисъл са генерирани и анализирани три различни вероятностни разпределения (нормално, логаритмично нормално и гама разпределения). За текущия пример, е приложена Гаусова копула с корелационна матрица (Σ) от:

$$\Sigma = \begin{matrix} 1.0 & 0.5 & 0.3 \\ 0.5 & 1.0 & 0.4 \\ 0.3 & 0.4 & 1.0 \end{matrix}$$

Първата фаза в процеса е свързана с дефиниране на параметрите за всяко разпределение, като впоследствие се изчисляват техните функции на плътност на вероятността (PDFs):

Нормално разпределение: Средна стойност (μ) от 0 и стандартно отклонение (σ) от 1.

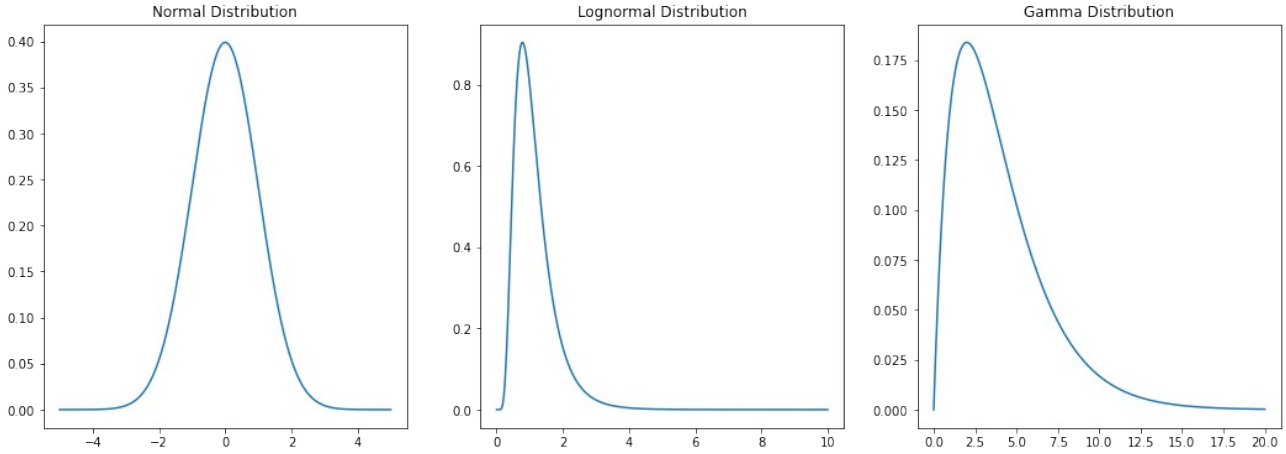
$$f_{\text{Normal}}(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

Логнормално разпределение: Средна стойност (μ) от 0 и стандартно отклонение (σ) от 0,5.

$$f_{\text{Lognormal}}(y)=\frac{1}{y\sigma\sqrt{2\pi}}\exp\left(\frac{-(\ln y-\mu)^2}{2\sigma^2}\right)$$

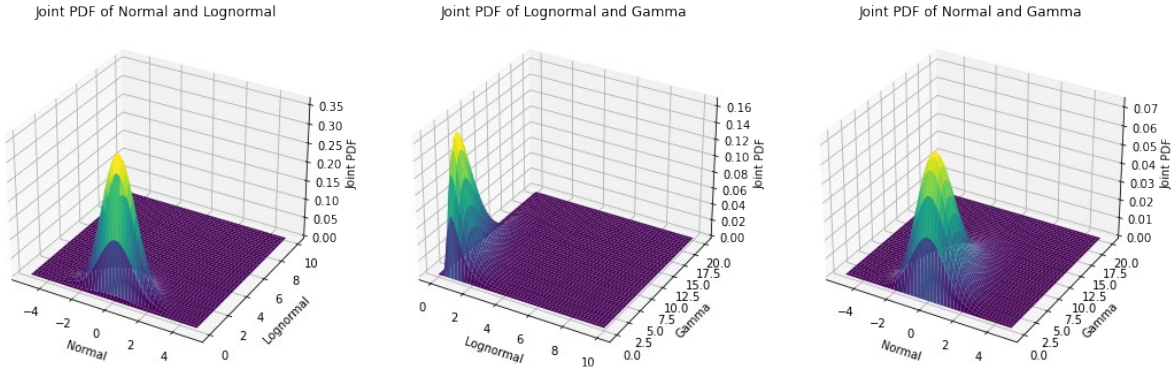
Гама-разпределение: параметър на формата (κ) от 2 и параметър на мащаба (θ) от 2.

$$f_{\text{Gamma}}(z)=\frac{z^{k-1}e^{-z/\theta}}{\theta^k\Gamma(k)}$$



Фиг. 13.5-1 Независими графични представяния на трите разпределения, използващи зададените стойности на параметрите

Изследването на съвместното поведение на двойки разпределения преминава през създаване на мрежови решетки за 3D изобразяване и изчисляване на съвместните PDF функции, за всяка двойка от разглежданите разпределения, при предположение за независимост.



Фиг. 13.5-2 PDF файлове, съединени по двойки, използващи 3D повърхност

Стъпка 2: Преобразуване на данните в равномерно разпределение

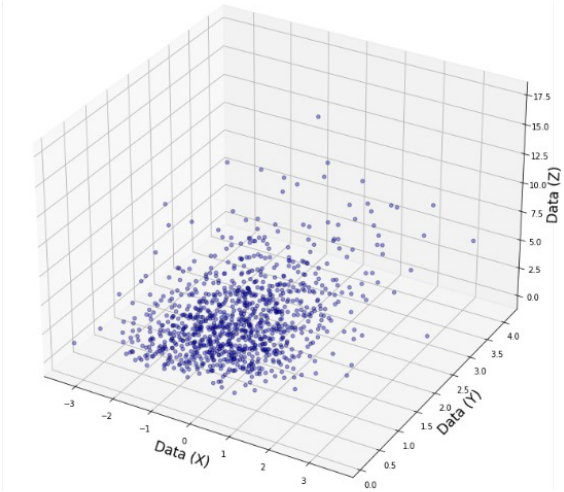
Следващата стъпка включва трансформиране на генерираните данни в равномерно разпределение. Това се постига чрез прилагане на кумулативната функция на разпределение (CDF), на нормалното разпределение, към всяка стойност в генерираните данни. CDF трансформацията съпоставя всяка стойност в диапазон между 0 и 1.

Чрез използване на свойствата на CDF, тази трансформация гарантира, че новите данни са равномерно разпределени. Въпреки това, трансформираните данни запазват структурата на зависимост, която първоначално е била наложена от Гаусовата копула.

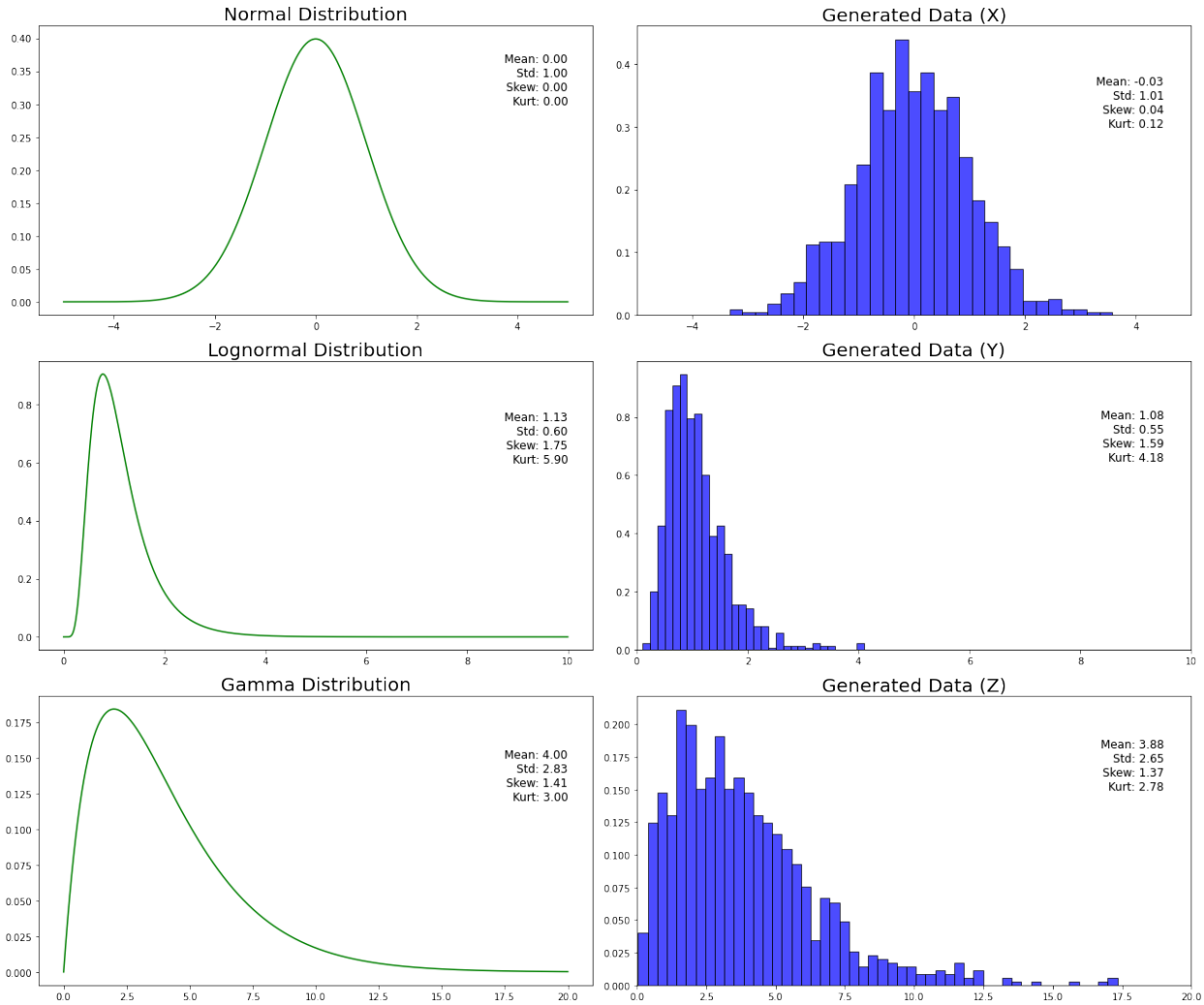
Стъпка 3: Трансформиране на унифицираните данни в желаните разпределения

Последната стъпка включва съпоставяне на равномерно разпределените данни с желаните гранични разпределения: нормално, логаритмично нормално и гама. Това се прави с помощта на обратната кумулативна функция на разпределение (PPF), за всяко целево разпределение.

Чрез картографиране на трансформираните данни, използвайки съответните обратни CDF, се получават три генерирани променливи (X, Y, Z). Тези набори от данни сега следват зададените гранични разпределения, като същевременно запазват структурата на зависимостите, получена от оригиналната копула. Фиг. 13.5-4 предоставя сравнение между получените емпирични параметри на разпределение и предварително зададените теоретични разпределения, демонстрирайки, че синтетичните данни успешно следват желаните разпределения и валидират ефективността на процеса на генериране на данни, базиран на копула.



Фиг. 13.5-3 Съвместна диаграма на разсейване на трите синтезирани променливи



Фиг. 13.5-4 Сравнения на получените емпирични параметри на разпределение с предварително зададените теоретични разпределения

13.6 Сравнителен анализ

Сравнението на оригиналните и синтетичните данни, посредством едностранния KS тест, се използва като средство за сравнение не само за статистическа значимост на съответствието на копулите, но и за сравняване на целесъобразността на използването на дадено семейство копули. В рамките на изследването са тествани петте семейства копули, описани в предишните части. Резултатите са показани в таблица 13.6-1.

Гаусовата копула показа висока прецизност при поддържане на зависимостите между характеристиките, като се наблюдават незначителни несъответствия в някои гранични разпределения. Едностранните тестове на

Колмогоров-Смирнов (КС) са проведени, с цел да се валидира еквивалентността на синтетичните набори от данни със съответните им теоретични разпределения.

Семейство Копула	**Променлива, разпределение на множества**	**Едностранен KS тест, p-стойност**
Гаусово	X, Нормално разпределение (μ = 0, σ = 1)	0,006, 0,93
Гаусово	Y, логнормално разпределение (μ = 0, σ = 0,5)	0,003, 0,98
Гаусово	Z. Гама разпределение (κ = 2, θ = 2)	0,012, 0,74
Клейтън	X, Нормално разпределение (μ = 0, σ = 1)	Не е изчислено*
Клейтън	Y, логнормално разпределение (μ = 0, σ = 0,5)	Не е изчислено*
Клейтън	Z. Гама разпределение (κ = 2, θ = 2)	Не е изчислено*
Гъмбел	X, Нормално разпределение (μ = 0, σ = 1)	Не е изчислено**
Гъмбел	Y, логнормално разпределение (μ = 0, σ = 0,5)	Не е изчислено**
Гъмбел	Z. Гама разпределение (κ = 2, θ = 2)	Не е изчислено**
Франк	X, Нормално разпределение (μ = 0, σ = 1)	0,019, 0,49
Франк	Y, логнормално разпределение (μ = 0, σ = 0,5)	0,020, 0,45
Франк	Z. Гама разпределение (κ = 2, θ = 2)	0,034, 0,09
Стюдънт	X, Нормално разпределение (μ = 0, σ = 1)	0,003, 0,15
Стюдънт	Y, логнормално разпределение (μ = 0, σ = 0,5)	0,012, 0,75
Стюдънт	Z. Гама разпределение (κ = 2, θ = 2)	0,005, 0,95

Таблица 13.6-1 Получени тестове на Колмогоров-Смирнов за трите променливи, синтезирани от различни семейства копули

* Изчислената θ стойност -0,069 е извън допустимите граници за дадената копула на Клейтън.

** Изчислената θ стойност 0.968 е извън допустимите граници за дадената копула на Клейтън.

И в трите случая KS тестовете показват високи p-стойности, което показва, че синтетичните набори от данни (X, Y, Z) са статистически неразличими от съответните им теоретични разпределения. Тези резултати потвърждават ефективността на описания метод за синтез на данни при запазване на зададената структура на зависимости и получаване на данни, които отговарят на желаните гранични разпределения.

Копулата на Клейтън не е изчислила съответствието. Параметърът θ, изчислен за копулата на Клейтън по време на процеса на напасване, е извън валидния диапазон - трябва да е положителен, тъй като не може да моделира отрицателна зависимост. Ако данните от извадката имат отрицателна зависимост, копулата на Клейтън, в общия случай, не е приложима.

По подобен начин, копулата на Гъмбел също не стига до числов резултат (не се изчислява за приложения пример), тъй като параметърът на напасване излиза извън валидния диапазон. За копулата на Гъмбел параметърът θ трябва да бъде по-голям или равен на 1, тъй като не може да моделира зависимост със стойности на θ по-малки от 1. Този проблем възниква, когато данните, използвани за напасване на копулата, не показват достатъчна зависимост за копулата на Гъмбел, което води до невалидна стойност на θ.

Резултатите от KS тестовете показват, че синтетичните данни, генерирани с помощта на Франк копула, съвпадат добре с определените теоретични разпределения. Високите p-стойности и в трите случая показват, че няма съществени доказателства за отхвърляне на нулевата хипотеза, потвърждавайки, че генерираните синтетични данни следват точно определените разпределения. Въпреки това, всички резултати в сравнение с Гаусовата копула са по-лоши, което удостоверява, че последната е по-подходяща за даденото разпределение.

Що се отнася до T-копулата на Стюдънт, синтетичните данни също съвпадат добре със зададените теоретични разпределения. Относително високите p-стойности и в трите случая показват, че няма съществени доказателства за отхвърляне на нулевата хипотеза, потвърждавайки, че генерираните синтетични данни следват точно зададените разпределения. Гама-разпределението показват най-добри резултати на тестовете от всички разглеждани разпределения, което показва, че използваната копула е най-подходяща за зададеното гама-разпределение.

13.7 Заключение

В настоящото изследване е разгледано предизвикателството за генериране на синтетични многомерни данни, които поддържат сложните зависимости, наблюдавани в оригиналния набор от данни. Чрез използване на силата на копулите, по-специално чрез метода на обърнато генериране на извадки, е демонстриран надежден подход за генериране на висококачествени синтетични данни.

Описаният подход включва три ключови стъпки: генериране на данни с помощта на различни семейства копули, чрез които се налага желаната

структура на зависимост, трансформиране на данните в равномерно разпределение и след това напасване на равномерните данни към специфични гранични разпределения (нормално, логаритмично нормално и гама). Ефективността на тези методи е валидирана посредством серия от тестове на Колмогоров-Смирнов (КС), които потвърждават, че синтетичните данни имитират коректно теоретичните разпределения на оригиналните данни.

Експерименталните резултати показват висока прецизност при поддържане на зависимостите между характеристиките и незначителни несъответствия в определени гранични разпределения, както се вижда от високите p-стойности от KS тестовете. Тези резултати потвърждават разглеждания в изследването метод за синтез на данни, базиран на копули, като мощен инструмент за генериране на синтетични набори от данни, които запазват сложни взаимозависимости, като същевременно се придържат към определени гранични разпределения.

По отношение на резултатите Гаусовата копула се представя най-добре, което показва, че тя поддържа предварително зададените зависимости между дадените променливи. KS тестовете за Гаусовата копула удостоверяват, че синтетичните набори от данни са статистически неразличими от съответните им теоретични разпределения.

T-копулата на Стюдънт също генерира синтетични данни, които съвпадат със зададените теоретични разпределения. Относително високите p-стойности и в трите случая свидетелстват, че няма значителни доказателства за отхвърляне на нулевата хипотеза, потвърждавайки, че генерираните синтетични данни следват точно зададените разпределения. Резултатите за гама разпределението представляват интерес, което предполага, че T-копулата на Стюдънт е добре пригодена за гама разпределението. Изводите от KS тестовете удостоверяват, че синтетичните данни, генерирани с помощта на копула на Франк, съвпадат добре със зададените теоретични разпределения. Резултатите, сравнени с Гаусовата копула, показват по-лошо представяне, което демонстрира, че Гаусовата копула е по-подходяща за дадения набор от разпределения.

Копулата на Клейтън и копулата на Гъмбел не стигат до числов резултат, тъй като параметърът, получен по време на процеса на напасване, остава извън валидния диапазон, което означава, че тези семейства копули не са подходящи за предварително зададените разпределения.

Методът за обратно извадково моделиране с копула предлага гъвкав и ефикасен подход за генериране на синтетични данни, което го прави привлекателна опция за различни приложения в науката за данни, включително обучение на модели за машинно обучение, симулации и аналитични задачи.

Бъдещата работа се съсредоточава върху проучването на приложението на други видове копули и допълнително да усъвършенства метода, за да подобри неговата приложимост към още по-широки набори от разпределения на данни.

Използвана литература

[1] Nelsen, R., 2006, An Introduction to Copulas (Second ed.) New York: Springer, ISBN 978-1-4419-2109-3.

[2] Schmidt, T., 2006, Coping with Copulas, Copulas-From theory to application in finance. ISBN 978-1-9043-3945-8.

[3] Mai, J., Scherer M., 2012, "Simulating Copulas (Stochastic Models Sampling Algorithms and Applications)". (“Copula (probability theory) - Infogalactic: the planetary knowledge core”) World Scientific. ISBN 978-1-84816-874-9.

[4] Mukhopadhyay, S., Parzen, E., 2019, Nonparametric Universal Copula Modeling https://doi.org/10.48550/arXiv.1912.05503.

[5] Restrepo, J., P., Rivera, J., C., Laniado, H., Osorio, P., Becerra, O., 2023, Nonparametric Generation of Synthetic Data Using Copulas, Electronics, 12(7), 1601; https://doi.org/10.3390/electronics12071601.

[6] Benali, F., Bodénès, D., Labroche, N., de Runz, C., 2021, MTCopula: Synthetic Complex Data Generation Using Copula, International Workshop on Data Warehousing and OLAP, (2021).

[7] Marchev, A., Jr., Marchev, V., 2022, Synthesizing multi-dimensional personal data sets, AIP Conference Proceedings, 2505(1):020012, DOI: 10.1063/5.0100615.

[8] Marchev, V., Marchev, A., Jr., Piryankova, M., Masarliev, D., Mitkov, V., 2023, Synthesizing an anonymized multidimensional dataset featuring financial, economic, demographic, and personal traits data, VANGUARD SCIENTIFIC INSTRUMENTS IN MANAGEMENT, vol. 19, no. 1, 2023, ISSN 1314-0582.

[9] Marchev, A., Jr., Marchev, V., 2024, Automated Algorithm for Multi-variate Data Synthesis with Cholesky Decomposition, ICACS '23: Proceedings of the 7th International Conference on Algorithms, Computing and Systems October 2023, Pages 1 – 6

[10] Sklar, A., 1996, Random variables distribution functions and copulas – a personal look backward and forward, ISBN 978-0-940600-40-9, DOI: 10.1214/lnms/1215452606.

[11] Wan, K., Kornhauser, A., 2023, On the Properties of Gaussian Copula Mixture Models, https://doi.org/10.48550/arXiv.2305.01479.

[12] Clayton, D., 1978, A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease

incidence, Biometrika, Volume 65, Issue 1, Pages 141–151, https://doi.org/10.1093/biomet/65.1.141.

[13] Kurowicka, D., van Horssen, W., T., 2015, On an interaction function for copulas, Journal of Multivariate Analysis Volume 138, June 2015, Pages 127-142, https://doi.org/10.1016/j.jmva.2014.12.012.

[14] Liebscher, E., 2008, Construction of asymmetric multivariate copulas, Journal of Multivariate Analysis Volume 99, Issue 10, November 2008, Pages 2234-2250, https://doi.org/10.1016/j.jmva.2008.02.025.

[15] Manner, H., 2007, Estimation and Model Selection of Copulas with an Application to Exchange Rates, RM/07/056.

[16] Fang, R., Li, X., Li, L., 2016, Generalized multivariate Gumbel distributions - Dependence, aging properties and applications. J Syst Sci Complex 29, 1752–1772 (2016). https://doi.org/10.1007/s11424-016-4272-8.

[17] Machler, M., 2011, Numerically Stable Frank Copula Functions via Multiprecision: R Package Rmpfr, DOI: 10.13140/RG.2.2.27067.92968 (2011).

[18] Lourme, A., Maurer, F., 2017, Testing the Gaussian and Student's t copulas in a risk management framework, Economic Modelling Volume 67, Pages 203-214, https://doi.org/10.1016/j.econmod.2016.12.014.

[19] Wilson, A., G., Ghahramani, Z., 2010, Copula Processes, arXiv:1006.1350.

[20] Jaworski, P., Durante, F., Härdle, W., K., Rychlik, T., 2010, Copula Theory and Its Applications, Lecture Notes in Statistics Springer, ISSN 2197-7186.

[21] Strelen, J., 2009, Tools for Dependent Simulation Input with Copulas, 2nd International ICST Conference on Simulation Tools and Techniques.

[22] Massey, F., J., Jr., 1951, The Kolmogorov-Smirnov Test for Goodness of Fit, Journal of the American Statistical Association. Vol. 46, No. 253, https://doi.org/10.2307/2280095.

[23] Drew, J., H., Glen, A., G., Leemis, L., M., 2000, Computing the cumulative distribution function of the Kolmogorov–Smirnov statistic, Computational Statistics & Data Analysis 34 (2000) 1–15. https://doi.org/10.1016/S0167-9473(99)00069-9.

[24]

14 Синтезиране на анонимизиран многомерен набор от данни

Автори: Васил Марчев, Ангел Марчев, мл., Милена Пирянкова, Даниел Масърлиев, Валентин Митков

Резюме: Настоящата глава представя детайлен разрез на подход за генериране на синтетични масиви от данни, който се справя с недостига на набори от данни, съдържащи чувствителна информация, поради ограничения, наложени от законодателство като GDPR и Закона за кредитните институции, в частта му уреждаща банковата тайна. Чрез интегриране на статистически методи, като: симулация Монте Карло и разлагане на Холецки, както и на предварително синтезирана бизнес логика, изследването очертава цялостна методология за създаване на многомерни синтетични набори от данни. Тези набори от данни включват демографски, личностни, финансови и банкови променливи, за да симулират профилите на финансово активните лица. Разглежданата алтернатива на традиционните методи за събиране на данни предлага решение на предизвикателствата, свързани с достъпа до чувствителни данни, като същевременно се спазват правните рамки. Използването на синтетични данни позволява запазването на взаимовръзките между променливите и осигурява сигурна среда за тестване, въпреки присъщите сложности при генерирането на висококачествени синтетични бази данни. Валидирането на синтезираните данни чрез теста на Колмогоров-Смирнов гарантира тяхната точност и релевантност. Този подход не само улеснява развитието на модели, базирани на данни, в области, в които достъпът до чувствителни данни е ограничен, но и насърчава етичното използване на данните чрез спазване на разпоредбите за поверителност. Главата демонстрира потенциала на синтетичните данни да служат като значим и релевантен ресурс за научни изследвания, предлагайки тестово проучване на процеса на генериране и последващите стъпки по отношение на бъдещите приложения при изследвания основани на чувствителни масиви от информация.

Ключови думи: Генериране на синтетични данни, Разлагане на Холецки, Тест на Колмогоров-Смирнов

14.1 Въведение

Научните изследвания, базирани на големи масиви от синтезирани данни, изискват набор от подготвителни действия за постигане на необходимото количество и качество на информацията. Особеностите на тези данни и

законовата рамка налагат различни подходи в процеса за тяхното генериране. В практиката се наблюдава недостиг на масиви, съдържащи чувствителна информация, като лични или банкови детайли. Подобен тип информационни набори са обект на европейското законодателство, GDPR и правилата за банковата тайна. Това ограничава възможностите на изследователите да осигурят нужните ресурси за своите модели

Разработен е алтернативен подход за осигуряване на необходимата синтетична информация. Разглежданият метод се основава на базата на статистическо моделиране, посредством което се извършва симулиране на информационни масиви въз основа на демографски, личностни и банкови променливи. Подходът създава предпоставки за генериране на многомерни набори от данни, които обединяват потребителски характеристики и предпочитания. Посредством набор от симулационни подходи, които се интегрират в система с предварително установени показатели, се цели да бъде синтезиран и установен цялостен профил на статистически значим брой лица. Конкретната цел на разглеждания подход е, чрез предварително установените критерии, да се изведе профил на финансово активния потребител - физическо лице.

14.2 Синтетични данни

Синтетичните данни представляват информационен набор, които не съществува в реалния свят. Това е генерирана информация, базирана на различни статистически зависимости, реализирани в алгоритми. Този тип данни са подходящи за тестови данни, за обучение, както и за валидиране на модели. Подробна информация за синтетичните информационни масиви може да бъде намерена в глава 3 – „Същност на синтетични данни. Приложение в управлението на бизнес структури“, в рамките на настоящата разработка.

Дадена е многомерна крайна случайна променлива от реални данни, обозначаваща S броя (от общо S) на количествените характеристики на изследваната система S и n броя наблюдения (редове). Използва се генеративен метод $g(A)$, за да бъдат симулирани информационни масиви, със статистическо сходство спрямо първоначалното разпределение.

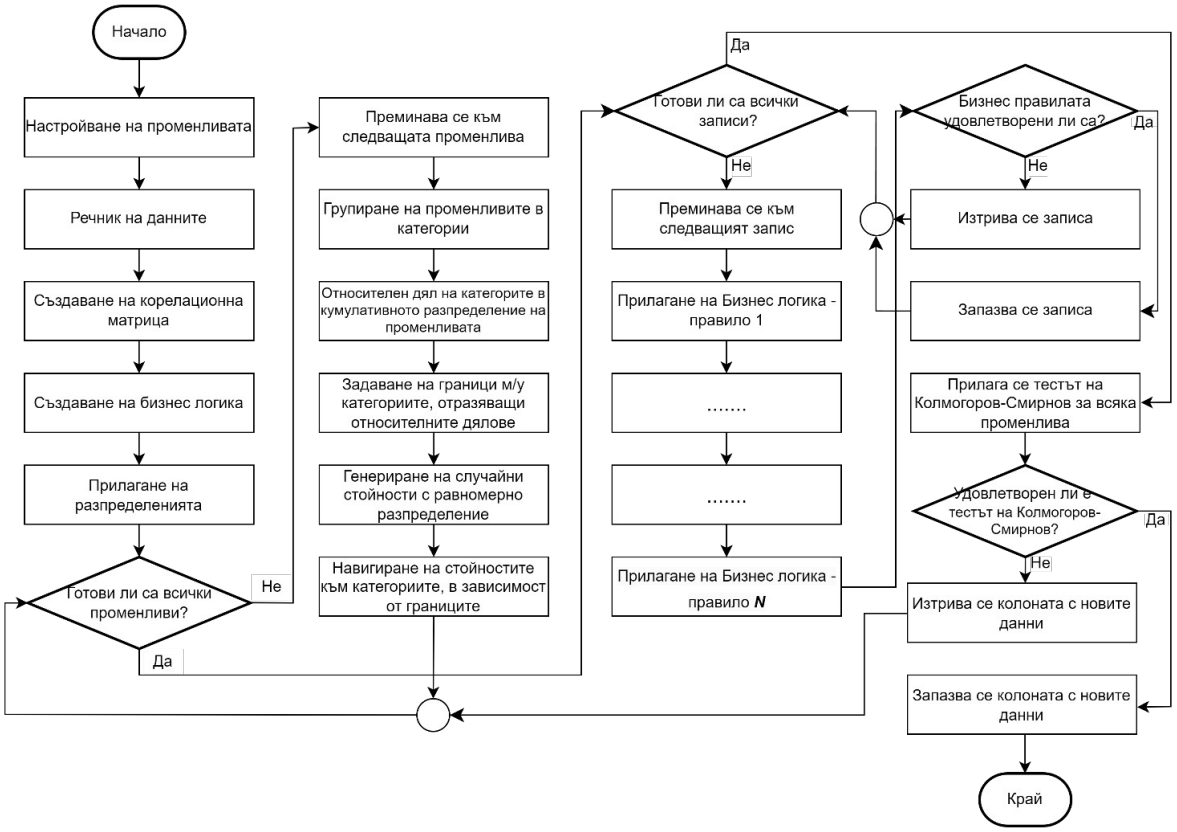
$$g(A)=\widetilde{X}_{n,s} \text{ където } \widetilde{X}_{n,s}\sim\omega$$

Тук A представлява набор от всички изисквания, свойства и ограничения за генериране на правилни, сходни и реалистични данни. Има много възможни подходи, както е показано в предишни публикации [1]. Информация за

различните подходи за симулация на данни използвани от авторите на настоящата разработка могат да бъдат намерени и в глава 7 - Синтезиране на многоизмерни набори от лични данни.

14.3 Симулиране на данни

Алгоритъмът (фиг.1) за симулация на данни преминава през няколко основни фази: задаване на променливите, създаване на речник от данни, задаване на статистически разпределения и създаване на корелационна матрица за числови променливи, както и генериране и дефиниране на бизнес логиката за елиминиране на невъзможни стойности, както и анализ на получените данни.



Фиг.14.3-1 Алгоритъм за симулация

14.3.1 Речник на данните

По време на тази фаза се създава речник, обясняващ наличните променливи. Той предоставя синтезирана информация за предварително дефинирани описателни индикатори по отношение на всяка избрана характеристика. Речникът служи като ръководство за разбиране и анализ на данните по време на процеса на синтез.

Разглежданият модул от модела включва няколко ключови индикатора за мониторинг, които подобряват разбираемостта на данните. Първият индикатор, „Коефициент“, представлява наблюдаваната характеристика. Колоната „Код“ предоставя компютърно четим код, свързан с всяка характеристика, улеснявайки автоматизираната обработка и анализ на наличните данни.

Информацията за „Тип променлива“ е от съществено значение. Тя помага за класифицирането на променливите в различни типове, като категорични, числени или ординални, като помага за избора на подходящи статистически техники и интерпретации. Индикаторът „Възможни стойности“ очертава потенциалните стойности, които една функция може да приеме. Тази информация е от решаващо значение за разбирането на обхвата и естеството на данните.

Колоната „Извеждане“ показва метода за синтез на данни, използван за конкретна характеристика. Това помага за разбирането на това, как данните са били получени или изведени, било то чрез директни измервания, преобразувания или изчисления.

Речникът на данните значително подобрява прозрачността и полезността на информационните масиви по време на процеса на комбиниране, което улеснява анализа и разбирането.

№	Фактор	Код	Тип на променливата	Възможни стойности	Получаване
1	Пол	sex	Бинарна	М / Ж	Симулация
2	Навършени години	age	Непрекъсната	20 - 85	Корелация
3	Ниво на образование	lv_educ	Ординална	-Незавършено начално образование -Начално образование -Основно образование -Висше образование -Университетска диплома	Симулация

4	Трудов статус	empl_stat	Номинална	-Работодатели -Самостоятелно наети -Служители в частни предприятия -Служители в държавни предприятия -Неплатени семейни работници -Безработен	Симулация
5	Семеен статус	marit_stat	Номинална	-Неженен/а -Женен/Омъжена -Разведен/а -Вдовец/а	Симулация
6	Брой членове на домакинство	house_memb	Интервална	1; 2; 3; 4; 5+	Симулация

Таблица 14.3.1-1 Речник на данните

14.3.2 Променливи

Променливите предоставят на модела необходимите входни данни за генериране на информационни масиви. В таблица 14.3.2-1 е показан общият списък от разглеждани променливи. За целите на настоящата работа са разграничени няколко основни групи променливи [2]:

Демографски променливи – За целите на описваното изследване се разглежда набор от демографски характеристики, с цел да се установят корелационни зависимости между демографските фактори и набор от други индивидуални характеристики. В резултат на споменатите фактори се цели установяване на по-пълни потребителски профили.

Индивидуални характеристики - чрез проследяване на изброените характеристики се установява по-пълна и адекватна диференциация на индивидуалните данни.

Психологически и лични променливи - в зависимост от редица лични качества, вродени и придобити, се проявяват различни реакции към заобикалящата ни среда, които имат съществена роля както за характера, така и за динамиката на действията на отделните индивиди в инвестиционния процес.

Финансови и банкови променливи - показват индивидуалното поведение на индивидите на пазарите на финансови услуги. Те предоставят възможност за установяване на профил на активния потребител на финансови услуги.

Променлива	Група	Променлива	Група
Пол	Демографски данни	Депозити на ФЛ	Банкови и финансови променливи
Възраст (навършени години)	Демографски данни	Доход	Банкови и финансови променливи
Степен на образование	Демографски данни	Лични разходи	Банкови и финансови променливи
Статус на заетост	Демографски данни	Жилищни разходи	Банкови и финансови променливи
Семейно положение	Демографски данни	Данъци и осигуровки	Банкови и финансови променливи
Брой членове на домакинството	Демографски данни	Транспорт и съобщения	Банкови и финансови променливи
Брой деца под 18 години	Демографски данни	Свободно време и хоби	Банкови и финансови променливи
Националност	Демографски данни	Предпочитан начин за банкиране	Банкови и финансови променливи
Вероизповедание	Демографски данни	Среден брой банкови операции	Банкови и финансови променливи
Социално-икономически статус	Демографски данни	Дебитна карта	Банкови и финансови променливи
Професия - индустрия	Демографски данни	Кредитна карта	Банкови и финансови променливи
Професионален статус	Демографски данни	Разплащателна сметка в лв	Банкови и финансови променливи
Образование - област	Индивидуални характеристики	Застраховка - имущество	Банкови и финансови променливи
Брой апартаменти/къщи	Индивидуални характеристики	Застраховка - живот	Банкови и финансови променливи
Брой парцели земя	Индивидуални характеристики	Застраховка – МПС (Каско)	Банкови и финансови променливи
Брой автомобили	Индивидуални характеристики	Допълнително здравно осигуряване	Банкови и финансови променливи
Жилището, в което живея е:	Индивидуални характеристики	Овърдрафт	Банкови и финансови променливи
Темперамент	Психологически и личностни	Потребителски кредит	Банкови и финансови променливи
Индивидуална склонност към риск	Психологически и личностни	Ипотечен кредит	Банкови и финансови променливи
Предишен инвестиционен опит в години	Индивидуални характеристики	Автомобилен лизинг/кредит	Банкови и финансови променливи
Сегмент акции	Банкови и финансови променливи	Допълнително пенсионно осигуряване	Банкови и финансови променливи
Сегмент за облигации	Банкови и финансови променливи	Овърдрафт - кандидатствали	Банкови и финансови променливи
Други	Банкови и финансови променливи	Потребителски кредит - одобрени	Банкови и финансови променливи
Инвестиционни фондове	Банкови и финансови променливи	Потребителски кредит - кандидатстващи	Банкови и финансови променливи
Пари във валута	Банкови и финансови променливи	Ипотечен кредит - кандидатстващи	Банкови и финансови променливи
Криптовалути	Банкови и финансови променливи	Кредитна карта - кандидатстващи	Банкови и финансови променливи
ДЦК	Банкови и финансови променливи		

Таблица 14.3.2-1 Променливи използвани в процеса на симулация

14.3.3 Анализ на разпределенията

За целите на това изследване съществуват няколко основни подхода за анализ на статистически разпределения: в случаите, когато имаме достатъчно информация, за да определим разпределение, в случаите, когато няма достатъчно информация, и в случаите, когато трябва да се направи правдоподобно предположение.

В случаите, при които се наблюдават достатъчно данни за генериране на разпределение и е налично предположение, че текущите условия и естеството на данните няма претърпи промени, първоначалните разпределения за всеки от факторите се запазват.

В хипотезата, при която не са налични достатъчно данни за създаване на разпределението, се използва специфична, аналитична информация за всяка от разглежданите променливи. Това могат да бъдат: средна стойност, минимум, максимум, среднопретеглена стойност и други. След определяне на наличната специфична информация за променливите, се прави частично предположение въз основа на наличните знания.

Правдоподобно допускане е приложимо в случаите, при които няма налична информация за разглежданите променливи. При разглежданата перспектива, въз основа на априорни знания, се прави правдоподобно предположение за разпределението.

Избраните променливи се подлагат на цялостен анализ, за да се определят разпределенията, които ще се използват.

За целите на настоящото изследване се използват два основни метода за оценка на разпределенията: Метод на моментите и Обобщен метод на моментите.

Методът на моментите [3] и обобщеният метод на моментите (GMM) [4] представляват статистически процедури, използвани за оценка на параметрите на разпределение, генерирано върху извадка от съществуващите данни.

При метода на моментите, моментите на извадката се приравняват към моментите на теоретичното разпределение и тези уравнения се решават за съответните параметри. Например, в случай че $X_1, X_2, \ldots, X_n$ е извадка от нормално разпределение, средната стойност на извадката $\overline{X}$ и нейната дисперсия S^2 могат да бъдат приравнени към теоретичната средна стойност μ и дисперсия σ, получавайки уравнението:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i =, S^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 = \sigma^2$$

GMM представлява процедура, при която се обобщава на Метода на моментите. Чрез посочената методология се постига подобрена гъвкавост и адаптивност на селекцията от настоящи условия. Използва се метод за оценка на параметрите, в случаите, при които са налични повече моментни състояния, отколкото са наличните параметри, или когато те не са точно разрешими. Оценяващата величина $\hat{\theta}$ предоставя възможно решение на следния оптимизационен проблем:

$$\hat{\theta} = argmin_{\theta \in \Theta} g(\theta)' W_g(\theta)$$

Където $g(\theta)$ е вектор на моментните условия на извадката, W е положително дефинирана претеглена матрица и Θ е параметричното пространство. Този оценител е практически приложим, в случаите, при които настоящите условия са получени от икономическата или поведенческата теория, какъвто е най-честия случаи в иконометрията [1].

14.3.4 Разпределения

В процеса на подготовка на симулацията, предварително избраните променливи се подлагат на внимателен анализ, за да се определят подходящите разпределения. Променливите са разделени на няколко основни групи, в зависимост от вида на разглеждания фактор.

14.3.4.1 Разпределения на демографски променливи

За целите на настоящата разработка са изследвани редица демографски характеристики, за да се установят възможните корелации между демографските фактори и други индивидуални характеристики. Като резултат от процеса на обособяване на факторите се цели да се установят по-големи и по-удобни подпопулации. Демографските характеристики, използвани в симулацията, са получени след анализ на проведеното Преброяването на населението и жилищата в Република България от 2021 г., както и информация от НСИ [5].

Разпределение по пол - данни към 28.04.2023 г. - Национален статистически институт [6].

Пол - 28.04.2023	
М	Ж
48.1%	51.9%

Таблица 14.3.4.1-1 Разпределение по пол

Разпределение по възраст – данни към 28.04.2023 г. източник НСИ [6].

Възраст - 28.04.2023						
20 - 24	25 - 34	35-44	45-54	55-64	65-74	75+
5.5%	14.5%	17.9%	18.7%	16.5%	15.6%	11.2%

Таблица 14.3.4.1-2 Разпределение по възраст

Степен на образование - данни към 28.04.2023 г. източник НСИ [6].

Завършена образователна степен - 28.04.2023				
Незавършено начално образование	Начално образование	Прогимназия	Средно образование	Висше образование
5.1%	5.8%	16.4%	47.1%	25.5%

Таблица 14.3.4.1-3 Степен на образование

Трудов статус - данни към 28.04.2023 г. източник НСИ [6].

Трудов статус - 28.04.2023				
Работодатели	Самонаети лица	Служители в частни предприятия	Служители в публични предприятия	Неплатени семейни работници
3.8%	6.7%	68.9%	20.0%	0.5%

Таблица 14.3.4.1-4 Трудов статус

Семеен статус - данни към 28.04.2023 г. източник НСИ [6].

Семеен статус - 28.04.2023			
Неженен/а	Женен/омъжена	Разведен/а	Вдовец/а
39.7%	44.3%	5.8%	10.2%

Таблица 14.3.4.1-5 Семеен статус

Брой членове на домакинство - данни към 07.09.2021 г. - Преброяване 2021 г. [5].

Брой членове на домакинство - 07.09.2021				
1	2	3	4	5+
35.8%	29.4%	17.8%	11.5%	5.5%

Таблица 14.3.4.1-6 Брой членове на домакинство

Брой деца под 18 години - данни към 07.09.2021 г. - Източник: Преброяване 2021 г. [5].

Брой деца под - 07.09.2021				
Няма деца под 18	Едно дете под 18	Две деца под 18	Три деца под 18	Четири или повече деца под 18
36.9%	41.1%	19.4%	2.1%	0.5%

Таблица 14.3.4.1-7 Брой деца под 18 години

Националност - данни към 07.09.2021г. – Източник: Преброяване 2021г. [5].

Националност - 07.09.2021		
България	ЕС	Друга
99.1%	0.2%	0.7%

Таблица 14.3.4.1-8 Националност

Разпределение по вероизповедание - данни към 07.09.2021 г. - Преброяване 2021 г. [5].

Вероизповедание - 07.09.2021						
Протестант	Католик	Православен	Мюсюлманин	Друго	Без религия	Не се самоопределям
1.2%	0.7%	69.7%	10.8%	0.1%	5.2%	12.4%

Таблица 14.3.4.1-9 Вероизповедание

Социално-икономически статус - данни към 2022 г., източник НСИ [6].

Социално-икономически статус - 2022	
Активни	Неактивни
73.6%	26.4%

Таблица 14.3.4.1-10 Социално-икономически статус

Професия - Индустрия - данни към 31.12.2022 г. – Национален статистически институт [7]

Професия - Индустрия 31.12.2022	
Селско, горско и рибно стопанство	6.4%
Минно дело и преработка	19.1%
Комунални услуги	2.3%
Строителство	8.7%
Търговия	17.0%
Транспорт, складови дейности и пощи	6.8%
Хотелиерство и ресторантьорство	4.7%
Създаване на информация и творчество	3.8%
Финанси и администрация	2.8%
Държавно управление	10.2%
Образование и научни изследвания	9.2%
Здравеопазване и социални грижи	5.6%
Други дейности	3.5%

Таблица 14.3.4.1-11 Професия-индустрия – разпределение

Професионален статус - данни към 31.12.2022 г. – източник Национален статистически институт (НСИ) [6].

Професионален статус - данни - 31.12.2022					
Договор за управление	Трудов договор	Граждански договор	Самонает	Безработен	Пенсионер
2.3%	50.3%	2.5%	4.0%	4.3%	36.7%

Таблица 14.3.4.1-12 Професионален статус – разпределение

Къщи/апартаменти – данни към 07.09.2021 г. - окончателни данни, източник: Преброяване 2021 г. [5] и правдоподобно допускане.

Къщи/апартаменти - данни 07.09.2021		
0	1	2+
34.5%	57.2%	8.3%

Таблица 14.3.4.1-13 Къщи/апартаменти – разпределение

Земеделска земя - данни към 25.06.2018 г. – Министерство на земеделието и Agri.bg [8].

Земеделска земя - данни 25.06.2018	
Собственици	Несобственици
28.9%	71.1%

Таблица 14.3.4.1-14 Земеделска земя – разпределение

Земеделска земя - данни към 25.06.2018 г. - източник за броя на лицата, притежаващи земеделска земя - Министерство на земеделието, храните и

горите - публикувано в Agri.bg. Разпределението е базирано на броя на лицата над 18 години. По данни на Националния статистически институт.

Автомобил на домакинство - данни към 18.04.2023 г. – източник Национален статистически институт (НСИ) [6] и правдоподобно допускане;

Автомобили на домакинство - данни 18.04.2023		
0	1	2+
46.0%	48.9%	5.1%

Таблица 14.3.4.1-15 Автомобили на домакинство – разпределение

Жилището, в която живея, е: - данни към 18.04.2023 г. – източник Trading economics [9].

Жилището, в което живея е: данни - 31.12.2022	
Собствено	Под наем
84.3%	15.7%

Таблица 14.3.4.1-16 Собственост на жилище - разпределение

14.3.4.2 Разпределения на индивидуални характеристики

Образование - средни данни за завършилите съответните професии за периода 2001 - 2016 г. - източник на данни – Инфостат (Инфостат) [10]. При разпределението е направено допускане за запазване на дела на завършилите в съответните професионални области;

Образование-данни 2001 - 2016	
Педагогическо образование	7.6%
Хуманитеарни науки	4.7%
Социални, икономически и правни науки	52.7%
Естествени науки, математика и информатика	4.6%
Технически науки	15.3%
Селскостопански науки и ветеринарна медицина	1.8%
Здравеопазване и спорт	9.3%
Изкуства	2.3%
Сигурност и отбрана	1.9%

Таблица 14.3.4.2-1 Образование (направление) – разпределение

Един от най-ярките примери за индивидуални характеристики може да се види в индивидуалните предпочитания за риск - проведено е проучване за индивидуалните предпочитания за риск с над 1000 респонденти. Проучването е проведено от доц. д-р Ангел Марчев за целите на друг проект, но резултатите са релевантни и за настоящото изследване;

Индивидуална склонност към риск - калкулация-данни 31.03.2021										
0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
0.0%	0.0%	0.5%	3.5%	12.1%	23.3%	25.2%	17.6%	9.9%	5.2%	2.7%

Таблица 14.3.4.2-2 Индивидуална склонност към риск – разпределение

Темперамент - данни към 19.05.2009 г. – източник b-t.com.ua [11]

Темперамент - данни 19.05.2009			
Холерик	Флегматик	Сангвиник	Меланхолик
38.0%	11.0%	23.0%	28.0%

Таблица 14.3.4.2-3 Темперамент – разпределение

Предишен инвестиционен опит - източник - правдоподобно допускане;

Предишен инвестиционен опит в години - 28.04.2023				
0	1-5	6-10	11-15	16-25
65.0%	22.0%	7.0%	5.0%	1.0%

Таблица 14.3.4.2-4 Предишен инвестиционен опит в години – разпределение

Инвестиционни предпочитания – финансови инструменти – източник на правдоподобно предположение;

Инвестиционни предпочитания - инвестиционни инструменти		
	Да	Не
Сегмент акции	0.3%	99.7%
Сегмент облигации	0.0%	100.0%
Други ИИ	0.1%	99.9%
Взаимни фондове	6.5%	93.5%
Валута	4.1%	95.9%
Криптовалути	0.3%	99.7%
ДЦК	6.8%	93.2%
Депозити	81.8%	18.2%

Таблица 14.3.4.2-5 Инвестиционни предпочитания – ИИ

14.3.4.3 Доход

Разпределението на личните доходи се основава на данни, събрани чрез проучване сред 491 души.

Годишен доход (EUR)				
До 9 817	**От 9 818 до 14 112**	**От 14 113 до 27 610**	**От 27 611 до 42 335**	**От 42 336**
19.0%	17.3%	34.2%	14.3%	15.2%

Таблица 14.3.4.3-1 Годишен доход – разпределение

14.3.4.4 Разходи

Разпределението на разходите на домакинствата се основава на установени допускания относно диференциацията на групите по реализирани общи разходи.

Разходи на домакинство - разпределение			
Група 1	**Група 2**	**Група 3**	**Група 4**
10.0%	44.0%	33.0%	13.0%

Таблица 14.3.4.4-1 Разходи на домакинство – разпределение

Разходите за съответните групи са както следва:

Годишни разходи на домакинство (EUR)- разпределение по групи			
Група 1	**Група 2**	**Група 3**	**Група 4**
3 862	6 437	9 012	11 587

Таблица 14.3.4.4-2 Годишни разходи на домакинство – разпределение по групи

По отношение на различните видове разходи, въз основа на априорни знания и предположения за видовете разходи в домакинствата, се установяват следните зависимости:

Разходи - разпределение				
	Група 1	**Група 2**	**Група 3**	**Група 4**
Лични разходи	75.0%	50.0%	39.3%	33.3%
Разхди за домакинство	8.3%	15.0%	21.4%	22.2%
Данъци и застраховки	8.3%	10.0%	14.3%	13.9%
Транспорт и комуникации	8.3%	10.0%	10.7%	13.9%
Свободно време и хоби	0.0%	15.0%	14.3%	16.7%

Таблица 14.3.4.4-3 Разходи – Разпределение по групи и по видове

14.3.4.5 Банкови и финансови характеристики

Предпочитан начин на банкиране - данни към 08.12.2023 г. - източник НСИ [7]

Тип банкиране	
Онлайн	**В офис**
23.4%	76.6%

Таблица 14.3.4.5-1 Предпочитан тип на банкиране – разпределение

Среден брой банкови трансакции. Разглежданият фактор предоставя информация за активния потребител на финансови услуги. Разпределенията се основават на предположения и априорни познания за банковата и финансовата система. Базовото предположение приема, че физическите лица извършват

средно по 14 банкови транзакции на месец. Чрез функцията LOGNORM.DIST в Excel се установява разпределението на фактора.

Среден брой банкови трансакции (месечно)				
До 10	**11 до 14**	**14 до 20**	**20 до 26**	**Повече от 26**
1.0%	24.0%	47.0%	26.0%	1.0%

Таблица 14.3.4.5-2 Среден брой банкови трансакции – разпределение

Посредством разглежданите банкови разпределения, се цели да се диференцират активните потребители на банкови услуги. Разпределението между посочените по-долу категории е направено въз основа на анализ на банковата система и априорни познания за нивото на банковите и финансовите услуги в България.

Дебитна карта			
Под една	**Една**	**Две**	**Три**
1%	57%	38%	4%

Таблица 14.3.4.5-3 Брой дебитни карти – разпределение

Като се отчетат нивата на финансовите и банкови услуги в България и се използва публично достъпна информация от БНБ [12] за броя на кредитните карти и общия брой възрастни, е направено следното допускане за разпределението на притежателите на кредитни карти, в съотношение:

Кредитна карта	
Да	**Не**
20.2%	79.8%

Заявление за КК	
Да	**Не**
21.0%	79.0%

Таблица 14.3.4.5-4 Разпределение на кредити карти и заявления за КК

Банкова сметка		
Деца	**Да**	**Не**
0.0%	86.0%	14.0%

Таблица 14.3.4.5-5 Разпределение – банкова сметка

14.3.4.6 Застрахователни услуги

Въз основа на публично достъпните данни от Комисията за финансов надзор относно застрахователните услуги в България и състоянието на финансовата сигурност на физическите лица, в тази разработка са взети предвид следните разпределения на различни фактори от категорията „Застрахователни услуги“ [13].

Имуществена застраховка	
Да	Не
10.6%	89.4%

Застраховка Живот	
Да	Не
21.2%	78.8%

Застраховка Каско	
Да	Не
3.0%	97.0%

Таблица 14.3.4.6-1 Застрахователни услуги – разпределение

14.3.4.7 Допълнително здравно осигуряване

Отчитайки публично достъпните данни в Комисията за финансов надзор за общата стойност на допълнителното здравно осигуряване към 31.08.2013 г., се приема, че максималната стойност, внесена от едно физическо лице, е около 225 EUR годишно.

Допълнително здравно осигуряване	
Да	Не
2.0%	98.0%

Таблица 14.3.4.7-1 Допълнително здравно осигуряване – разпределение

14.3.4.8 Кредитни услуги

В резултат на мониторинг, априорни знания и опит в банковата система, публично достъпни данни в БНБ и предположения се генерират разпределения на кредитните услуги по отношение на клиентите на банките [12].

Овърдрафт	
Да	Не
19.4%	80.6%

Овърдрафт - кандидатстване	
Да	Не
24.4%	75.6%

Таблица 14.3.4.8-1 Овърдрафт и подадени документи за овърдрафт – разпределение

Потребителски кредит	
Да	Не
20.9%	79.1%

Потр. кредит - документи	
Да	Не
30.5%	69.5%

Потр. кредит - одобрения	
Да	Не
82.6%	17.4%

Таблица 14.3.4.8-2 Потребителски кредит, кандидатствания и одобрени документи – разпределения

Ипотечен кредит	
Да	Не
2.9%	97.1%

Ипотечен кредит - документи	
Да	Не
57.0%	44.0%

Таблица 14.3.4.8-3 Ипотечен кредит и кандидатствания – разпределения

Лизинг на автомобил	
Да	Не
20.0%	80.0%

Таблица 14.3.4.8-4 Лизинг за автомобили - разпределение

Допълнително пенсионно осигуряване (ДПО) - трети стълб - данни към 30.09.2018 г. - източник Комисия за финансов надзор [13]. Като основа за текущото разпределение на осигурените лица и общите стойности на фондовете към 30.09.2018 г. са използвани публично достъпни данни от Националния статистически институт за общия брой. Изчисляването на съотношението се получава по следната формула:

$$API = \frac{Всички\ застраховани\ лица}{Общо\ население - деца}$$

Където:

API - допълнително пенсионно осигуряване

Допълнително пенсионно осигуряване	
Да	**Не**
11.0%	89.0%

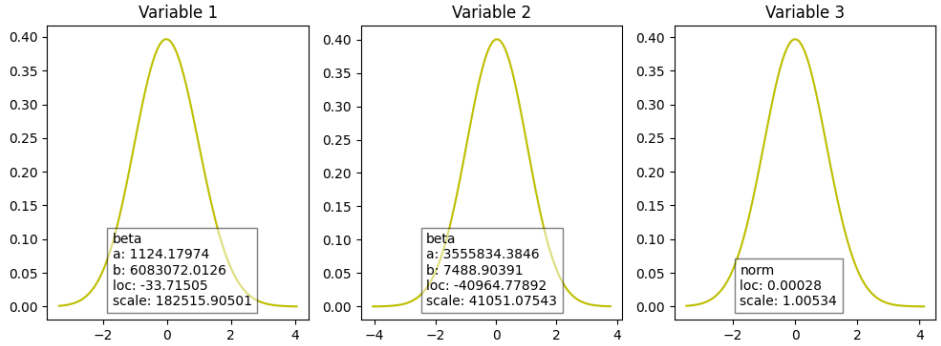
Таблица 14.3.4.8-5 Допълнително пенсионно осигуряване – разпределение

14.3.5 Симулация на данни

За целите на настоящото изследване симулацията се разглежда като процес, при който, използвайки статистически методи, интегрирани в алгоритми, се генерират голям брой записи, в зависимост от първоначално синтезиран набор от променливи, разделени в отделни групи.

Наблюдават се различни подходи за симулиране на данни, в зависимост от необходимостта, количеството и качеството на информацията [14]. Подходите се базират на разнообразни математически и статистически парадигми. За дадената задача, в настоящото изследване, се разглежда апробация на един от възможните подходи за генериране на масиви от данни.

Първата стъпка в процеса на симулация представлява последователност по извеждане на статистическо разпределение за всяка от избраните променливи. Това се прави чрез прилагане на метода на моментите. Като илюстрация на тази стъпка се разглеждат три случайни променливи и разпределенията, които са генерирани в предишната стъпка от процеса.



Фиг. 14.3.5-1 Изведени статистически разпределения и параметри

Следващата стъпка в процеса по синтезиране на информационни масиви е генерирането на записи за първата симулирана променлива. Това се прави с помощта на метода Монте Карло. Методът [15] предоставя възможност за генериране на голям брой синтетични записи, които улавят характеристиките и статистическите свойства на оригиналните данни. Тази техника изследва различни сценарии и оценява потенциалните резултати въз основа на наблюдаваните разпределения на оригиналните данни.

Методът Монте Карло може да бъде разширен, за да симулира данни за множество променливи. Освен това е необходимо да се интегрират зависимостите и корелациите между разглежданите разпределения. Той позволява генерирането на сложни синтетични набори от данни, които имитират основната структура и взаимовръзки на оригиналните данни.

Симулацията Монте Карло може да се използва за генериране на синтетични данни чрез извличане на случайни извадки от предварително зададени разпределения за всяка променлива. Процесът преминава през N броя итерации, за да се генерира набор от синтетични данни:

$$X_i \ \ D(\theta), i=1,2,\ldots,N$$

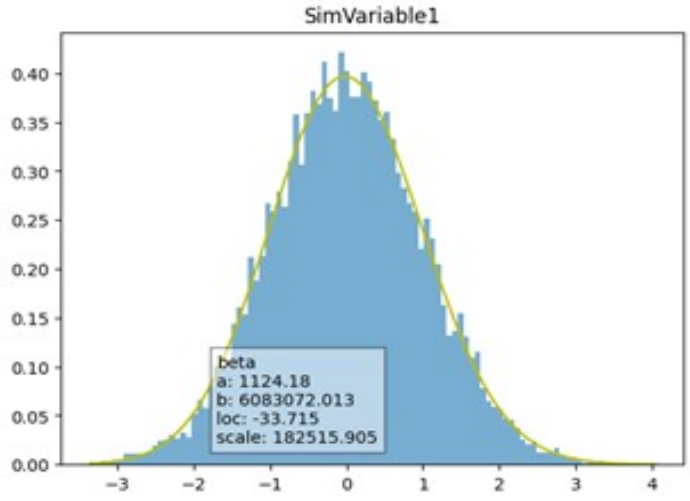
Където:

X - се генерира чрез вземане на проби от вероятностно разпределение

$D(\theta)$ - вероятностно разпределение

θ- обозначава параметрите на разпределението

N– брой итерации



Фиг. 14.3.5-2 Симулация Monte Carlo (базирана на реални статистически разпределения). Симулирани данни за разпределение на най-доброто съвпадение за променлива 1.

Като съществена част от процеса на симулация се разглежда генерирането на корелационна матрица на числените променливи. Тя представлява квадратна матрица с коефициенти на корелация за всяка двойка променливи. Коефициентът на корелация е мярка за линейната връзка между две променливи. Възможните диапазони са от -1 (перфектна отрицателна корелация) до 1 (перфектна положителна корелация). Корелационна матрица може да се използва за определяне на връзките между променливите.

$$R=\begin{vmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1m} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & \rho_{nm} \end{vmatrix}$$

Където ρ_{ij} е корелационен коефициент между X_i и X_j [1].

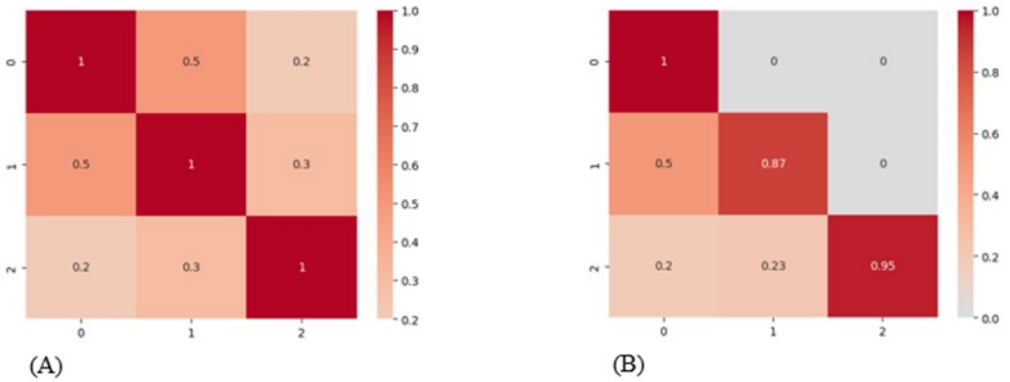
	age_sim	ind_risk_sim	income_sim	pers_exp_sim	house_exp_sim	taxes_sim	transp_telecom	hobby_sim
age_sim	1							
ind_risk_sim	-0.006659471	1						
income_sim	0.00291645	0.003991807	1					
pers_exp_sim	0.010777994	0.00806259	0.560949335	1				
house_exp_sim	0.006986746	0.004570236	0.588926663	0.928449924	1			
taxes_sim	0.007291537	0.006198534	0.581907425	0.929598635	0.930312793	1		
transp_telecom	0.009986651	0.007686998	0.56294651	0.934775948	0.927846735	0.929205101	1	
hobby_sim	0.009316307	-0.003323226	0.352350802	0.714298365	0.679286363	0.689442053	0.714114128	1

Таблица 14.3.5-1 Корелационна матрица – числови стойности

Следващата стъпка в процеса е да се изчисли разлагането на Холецки, което дава полуспрегната триъгълна матрица на корелационната матрица, като се генерира корелационна матрица като съществена част от алгоритъма.

Положително определена матрица, подобно на корелационна матрица, може да бъде разложена, като се използва подхода на разлагане на Холецки [16], за да се получи произведението на долна триъгълна матрица и нейната транспонирана матрица. Посредством посочената методология, набор от некорелирани променливи може да бъде трансформиран в набор от корелирани променливи.

Разлагането на Холецки обхваща по-голям диапазон от корелационните матрици на нормално разпределени променливи. Принципите на разлагането на Холецки могат да бъдат приложени към ненормални, некорелирани, случайни променливи [17]. Чрез прилагане на серия от трансформации става възможно да се получат корелирани ненормални променливи.



Фиг. 14.3.5-3 Получена корелационна матрица (A) и получена матрица на разлагане на Холецки (B)

14.3.6 Бизнес логика

За целите на настоящата глава, бизнес логиката се разглежда като синтезиран набор от правила и допускания, целящи да установят основните насоки за функциониране на системата. Това е последователен процес, при който, въз основа на априорни знания, свързани с наблюдаваните променливи, се установява набор от показатели, за които е възможно „да се получат невъзможни или много малко вероятни стойности“ [2].

Бизнес логиката се основава на поредица от потенциално взаимозависими фактори, техните ограничения, възможни комбинации и невъзможни такива. По този начин, разглежданата комбинация от правила има основната задача да елиминира невъзможните комбинации в новогенерирания набор от данни.

Основните стъпки, които могат да бъдат разграничени при създаването на бизнес логиката, са:

- Избор на потенциално взаимозависими фактори
- Установяване на възможни и невъзможни комбинации
- Елиминиране на невъзможните комбинации

№	Независима променлива	Стойност на независимата променлива	Зависима променлива	Филтър за стойност на зависимите характеристики	Забележка
1	Семеен статус	Женен	Брой членове на домакинство	>2	Броят членове на домакинство при семейните е по-вероятно да е по-голям от 2
2	Професия-индустрия	Финансови и административни дейности	Предишен инвестиционен опит (в години)	>0	В този случай е по-вероятно да притежават банкова сметка
3	Навършени	<25	Предишен	0	Под 24 години е

	години		инвестиционен опит (в години)		по-малко вероятно да имат инвестиционен опит. Между 35-44 и 45-54 години е по-вероятно да имат богат инвестиционен опит
4	Навършени години	<25	Степен на образование	< Висше образование	Под 24 години е по-малко вероятно да имат завършено висше образование
5	Навършени години	<25	Брой деца по 18 години	<2	От 20 до 24 години е по-малко вероятно да имат повече от едно под 18 години
6	Предишен инвестиционен опит	>0	Инвестиционен опит с банкови депозити	Да	По-вероятно да притежават банкова карта

Таблица 14.3.5-1: Бизнес логика – извадка

Пълният набор от бизнес правила е трицифрен (виж Приложение 1). Те се установяват в процеса на симулиране на данните, въз основа на априорната информация за разглежданата икономическа и банкова среда, демографските характеристики на населението и други.

След прилагане на описаната бизнес логика, новогенерираният набор от данни се намалява с около 40% от наличните записи.

14.3.7 Валидиране на данните

Процесът по валидация на синтезираните масиви потвърждава степента на съответствие на генерираната информация с реалните данни. Основната задача през този етап е верифицирането на високото качество и статистическата цялост на данните. Използва се детайлен сравнителен анализ, чрез който се установява дали първоначалните вероятностни разпределения са запазени в новия масив. За целите на настоящото изследване се използва статистическият тест на Колмогоров-Смирнов (KS). Този непараметричен метод измерва максималното разстояние между емпиричните функции на разпределение на двете извадки. Високата степен на припокриване служи като доказателство за прецизността на използвания модел и неговата способност да репликира сложни зависимости. По този начин се гарантира, че изкуствено създадените данни могат надеждно да заменят оригиналните масиви в рамките на прогнозиращите системи.

Таблица 14.3.7-1 показва всички характеристики, които не са удовлетворили приложения тест за валидност на данните. След това колоните с всички записи, които не удовлетворяват теста, се премахват и се преминават през процеса на симулация отново, докато всички променливи бъдат потвърдени от теста на Колмогоров-Смирнов.

Характеристика	D-Stat	P value	Заключение
Възраст (навършени години)	0.023283835	0.000000000	Съществена разлика
Степен на образование	0.047128882	0.000000000	Съществена разлика
Социално-икономически статус	0.096326690	0.000000000	Съществена разлика
Професия - индустрия	0.044784483	0.000000000	Съществена разлика
Брой парцели земя	0.473850462	0.000000000	Съществена разлика
Предишен инвестиционен опит в години	0.039663931	0.000000000	Съществена разлика
Инвестиционни фондове	0.042952976	0.000000000	Съществена разлика
Пари във валута	0.026182102	0.000000001	Съществена разлика
ДЦК	0.044255569	0.000000000	Съществена разлика
Доход	0.028233685	0.000000000	Съществена разлика
Лични разходи	0.479293018	0.000000000	Съществена разлика
Жилищни разходи	0.152993682	0.000000000	Съществена разлика
Данъци и осигуровки	0.153384460	0.000000000	Съществена разлика
Транспорт и съобщения	0.122513026	0.000000000	Съществена разлика
Свободно време и хоби	0.479390712	0.000000000	Съществена разлика
Разплащателна сметка	0.863423212	0.000000000	Съществена разлика

Таблица 14.3.7-1 Променливи, неудовлетворяващи теста на КС тест

14.4 Заключение

Разглежданият метод за симулация на информационни масиви предоставя възможност за генериране и синтезиране на необходимия набор от данни чрез интегриране на статистически разпределения и бизнес логика. Използваните методи за симулация, като Монте Карло и разлагане на Холецки, предоставят възможност за генериране на интегриран набор от синтетични данни. По този начин се получава информационен масив, който притежава всички характеристики на приложените към него статистически разпределения. Това от своя страна позволява използването на подхода в области, чувствителни към лични данни, както и приложението му в научните среди.

Настоящата глава предоставя изчерпателна информация за различните групи характеристики и начални статистически разпределения, интегрирани в алгоритмите на модела.

С помощта на бизнес логиката се постига филтриране и редуциране на новогенерираните данни, като се наблюдават потенциално взаимозависими фактори, техните ограничения, възможни комбинации и невъзможни такива.

Симулираните синтетични данни трябва да бъдат валидирани спрямо оригиналите разпределения, за да се гарантира тяхната релевантност и точност. В тази глава като метод за валидиране се използва тестът на Колмогоров-Смирнов (KS тест). Той е непараметричен тест за сравняване на извадково разпределение с теоретично разпределение или за сравняване на две извадкови разпределения едно с друго. Тестът предоставя информация дали разглежданият набор от данни следва първоначално зададените статистически разпределения.

Изследването предоставя информация за резултатите от първоначалното апробиране на метода, прилагането на бизнес логиката и запазването на новополучените данни, използвайки описания по-горе метод.

Процесът, обсъден в този текст, представлява практичен подход към синтетичните данни и възможностите за тяхното генериране. Методът разработва решение в случаите, когато данните, необходими за обучение на модели, са ограничени или попадат под закона за защита на личните данни.

Използването на синтетични данни предоставя практичен подход за избягване на проблема с личната информация на физическите лица. Изследването предоставя цялостна рамка за симулация на лични данни, използвайки статистически разпределения, корелационни матрици, алгоритми и бизнес логика.

Използвана литература

[1] Marchev, A., Marchev, V., 2023, Automated Algorithm for Multi-variate Data Synthesis with Cholesky Decomposition, ICACS 2023: the 7th International Conference on Algorithms, Computing and Systems, Larissa Greece, Association for Computing Machinery, New York, pp. 1 – 6, ISBN: 979-8-4007-0909-8.

[2] Marchev, V., Marchev, A., A., 2019, Simulation of a multi-criteria database for banking services. Algorithm and business logic, new information technologies, and big data: opportunities and perspectives in analyses and management decisions in business, economy, and social sphere, UNSS, pp. 179 – 190.

[3] Julier, S., Uhlmann, J., 1996, A general method for approximating nonlinear transformation of probability distributions

[4] Hansen, L., 1982, Large sample properties of generalized method of moments estimators, Econometrica, Vol. 50, No. 4 (July 1982).

[5] NSI, 2021, Census 2021, Sofia, https://census2021.bg

[6] NSI, Home, Demographic and social statistics

[7] NSI, 2023, Home, Business statistics R&D, Innovations and Information Society Information Society, Individuals Using the Internet by Purposes.

[8] Porozhanov, R., Broy na sobstvenitsite na zemedelski zemi i pritezhavanite ot tyah ploshti, agri.bg, 25.06.2018,
[9] Trading Economics, 2022, Home Ownership Rate in Bulgaria, https://tradingeconomics.com/bulgaria/home-ownership-rate
[10] Infostat.bg, Home, Demographic and social statistics, Population, Dec. 2023, infostat.nsi.bg/infostat/pages/module.jsf?x_2=80
[11] b-t.com.ua, 2009, Statistika po kommentariyam k testu Ayzenka, http://b-t.com.ua/test_ayzenk_rez_komment.html
[12] Bulgarian National Bank, 2023, Home, Statistics, Monetary and Interest Rate Statistics, Loans and Deposits by Amount Category and Economic Activity
[13] Financial Supervision Commission, 2022, Insurance Activity, Statistics
[14] Marchev, V, Marchev, A., 2021, "Methods for Simulating Multi-dimensional Data for Financial Services Recommendation", Bulgarian Economic Papers, Center for economic theories and policies, ISSN: 2367-7082, BEP 02-2021, Feb. 2021, http://www.bep.bg
[15] Moral, P, Doucet, A., Jasra, A., 2006. SEQUENTIAL MONTE CARLO SAMPLERS. J. R. STATIST. SOC. B (2006) 68, PART 3, PP. 411–436
[16] Dereniowsky, D., Kubale, M., 2003. Cholesky factorization of matrices in parallel and ranking of graphs, parallel processing and applied mathematics, 5TH INTERNATIONAL CONFERENCE, PPAM 2003, Czestochowa, Poland, Sep 7-10, 2003
[17] Qu, W., Liu, H. & Zhang, Z., 2020, A method of generating multivariate non-normal random numbers with desired multivariate skewness and kurtosis. Behav Res 52, 939 - 946

15 Методологични подходи за създаване на многомерни лични данни

Автори: Васил Марчев, Ангел Марчев мл., Калоян Харалампиев, Александър Ефремов, Боян Марков, Димитър Любчев, Милена Пирянкова, Богомил Филипов, Даниел Масърлиев, Валентин Митков

Резюме: Настоящата глава предоставя информация за описанието на метаданните при използване на алгоритъм за генериране на многоизмерен синтетичен набор от данни. Същевременно се разглеждат предизвикателствата, свързани със събирането и използването на обширни набори от данни за научни изследвания, особено в контекста на чувствителна информация, регулирана от правни рамки като GDPR и Закона за банковата тайна. Разглежданата методология използва симулационни техники за създаване на набор от данни, включващ 36 различни променливи, категоризирани в демографски, лични и банкови характеристики. Този синтетичен набор от данни е от съществено значение за емпирични изследвания, при които наличността на данни е ограничена поради законови рестрикции. В главата са разгледани разнообразни източници на данни, включително: Преброяването на населението на България през 2021 г., Националния статистически институт и Българската народна банка, осигурявайки цялостно покритие за извличане на разпределенията. В процеса на работа се подчертава важността от валидиране на генерираните данни, за да отговарят на стандартите за качество и да се подпомогне ефективното моделиране. Това проучване допринася за текущия дискурс за синтеза на информация в науката за данните, като подчертава набор от иновативни стратегии за справяне с недостига на информация. Същевременно се следват най-добрите практики на Евростат за описание на метаданни, като се прави подробна разбивка на всички променливи и се анализира необходимостта от тяхното включване в обобщения набор от информация, с оглед на целите на изследването.

Ключови думи: Синтетични данни, генериране на данни, разпределения, бизнес логика, симулация

15.1 Въведение

Научните изследвания базирани на големи информационни масиви изискват строга методология за набиране на данни, но стриктните правни рамки като GDPR и Закона за кредитните институции, в частта му уреждаща

банковата тайна силно ограничават достъпа до чувствителна информация. За преодоляване на тези бариери е приложена алтернативна стратегия за генериране на многомерни синтетични бази данни. Този подход позволява конструирането на необходимите количествени и качествени ресурси, интегриращи демографски, персонални и финансови променливи за нуждите на прогнозиращите модели.

Авторите на настоящата разработка [1] подчертават необходимостта от наличието на разглеждания алтернативен подход поради гореспоменатите предизвикателства при конвенционалните методи за събиране на данни. Предвид специфичните изисквания на емпиричното изследване е от съществено значение да се създаде набор от данни, който да поддържа модела, използвайки комбинация от симулационни техники. Отделните компоненти на данните се интегрират чрез разработения вероятностен метод на конкатенация.

Предвид възникващата нужда от синтезиране на набори от данни, е генериран набор от данни със социално-икономически, демографски, лични и финансови характеристики, които са описани в тази глава. Информационният масив е публично достъпен за бизнес, образователни и изследователски цели под лиценз на MIT (валидиран набор от данни https://doi.org/10.57967/hf/3701) [16].

Наборът от данни включва 36 различни променливи, организирани в различни категории. Тези категории обхващат демографски, индивидуални, личностни и банково-финансови характеристики. Наборът от данни се използва в случаите, когато необходимата ни информация липсва поради правни причини (данните изискват лиценз, попадат под закона за банковата тайна, GDPR и др.)[2].

15.2 Основни фази на процеса

15.2.1 Покритие

15.2.1.1 Географско покритие

Данните за демографските характеристики се базират на Преброяването на населението на България от 2021 г. [3] и Националния статистически институт, докато банковата информация се основава на публично достъпна информация от Българската народна банка и банковата система (търговски банки). Използвана е информация, получена от Министерството на финансите и Министерството на земеделието. Проведени са няколко независими проучвания и въпросници относно инвестиционните предпочитания на индивидите, както и външни изследвания върху личността и темперамента,

базирани на предварително проведено проучване с повече от 15 000 международни участници [4]

15.2.1.2 Времево покритие

Процесът на симулация на данни включва еднократно генериране на данни, базирани на временна снимка на разглежданите променливи. В случай че е необходимо данните да бъдат актуализирани, от съществено значение е да се осъвременят и разпределенията. Генерираните данни трябва да бъдат валидирани, за да се гарантира, че отговарят на установените критерии за количеството и качеството на получената информация.

15.2.1.3 Демографско покритие

Синтезираните данни представят определения обект на изследване като български индивид, непрофесионален инвеститор, с ограничен инвестиционен опит, който разполага с налични финансови средства за инвестиране.

15.2.2 Източник на данни

15.2.2.1 Първични източници на данни

Необходимият набор от данни се синтезира, използвайки метод, избран за целта. Данните се симулират въз основа на предварително събрана информация и генерирани разпределения. Разпределенията се извеждат въз основа на следните подходи:

При наличие на данни за формиране на разпределения, се прави предположение, че няма да има промяна в условията, по време на използване на този подход. А именно, запазване на по-ранните разпределения за всеки фактор, за който имаме достатъчно данни [3].

При липса на достатъчно данни – за изготвяне на съответно разпределение се използва информацията, която е налична за конкретната променлива, като например средна стойност на данните, минимум, максимум, среднопретеглена стойност и др. След установяване на наличната информация се извършва частична симулация, базирана на известните ни данни и на априорна информация.

Предположения – В случаите, когато нямаме налична информация за разпределенията на разглежданите променливи, се осъществява правдоподобно допускане. Извършват се симулации въз основа на априорни знания за разглежданите променливи, анализ на групата, към която принадлежи конкретният индикатор и др.

15.2.2.2 Доставчици на данни

За настоящото изследване, при генерирането на информация, е необходимо да се извлекат разпределения, базирани на публично достъпна информация с високо ниво на точност. Поради тази причина са използвани следните източници:

Национален статистически институт (НСИ) – основната държавна институция, отговорна за събирането и разпространението на статистически данни относно населението, икономиката и околната среда на България [3].

Преброяване 2021 г. – предоставя подробна демографска информация за населението на България. То включва данни за числеността на населението, разпределението, възрастовата структура, нивата на образование, социално-икономическите характеристики и др. [3].

Българска народна банка (БНБ) – Централна банка на Република България, която контролира паричната политика, финансовата стабилност и банковата система. Тя предоставя важни данни за паричните агрегати, лихвените проценти, валутните курсове и банковата статистика [5].

Банковата система в България – предоставя финансови услуги на физически лица и фирми, като същевременно генерира широк набор от информация относно практиките за кредитиране, депозитите, финансовите трансакции и др.

Комисия за финансов надзор (КФН) – отговаря за регулирането и надзора на небанковия финансов сектор в България. Това включва застрахователни компании, пенсионни фондове и инвестиционни посредници [6].

Министерство на финансите – контролира фискалната политика на страната, управлението на публичните финанси и бюджетните процеси [7].

Министерство на земеделието и храните – фокусира се върху селскостопанските политики и развитието на селските райони в България. То събира данни за селскостопанското производство, ползването на земеделските земи, добивите от култури и статистиката за животновъдството [8].

Априорна информация и предположения – отнасят се до знания или предположения, направени въз основа на знания, опит и предварителни данни за разглежданите събития и средата, в която те се развиват. В контекста на статистическите изследвания в България това може да включва исторически тенденции в демографските или икономическите показатели, които ръководят разработването на хипотези или модели.

15.2.3 Методология

В областта на изкуствения интелект и анализа на големи данни, обширните набори от данни са все по-често търсени. Разрастването на тези големи масиви от информация води до засилен фокус върху защитата на личните данни. Специфичните изисквания за определени видове данни, заедно с присъщата им сложност и правни съображения, представляват различни предизвикателства при получаването на необходимото количество и качество на информацията, които са съществени за ефективно моделиране [9].

Търсенето на специализирани информационни масиви често кара изследователите да проучват решения, базирани на синтетични данни. Синтетичните данни се отнасят до информация, която не съответства на действителни записи, а се генерира алгоритмично. Този тип данни се създават чрез статистически модели, а не се събират от наблюдения в реалния свят. Съществуват множество методологии за създаване на многомерни синтетични информационни масиви и много учени в областта на изкуствения интелект са разглеждали тази тема [10].

От методологична гледна точка, различни сценарии подчертават необходимостта от структуриран подход за генериране на многомерни набори от данни. Тази подглава разглежда ключови аспекти, при които генерирането на данни е съществен компонент от процеса на анализ на информацията, организиран съгласно методологични принципи.

Принцип на необходимото разнообразие

Тази концепция подчертава необходимостта от разнообразен и изчерпателен набор от входни данни, за да се подобри ефективността и точността на анализа и моделирането на данни. Критична подфаза в този контекст представлява инженерингът на характеристики, което включва избиране, трансформиране и създаване на нови характеристики от сурови данни, за да се оптимизира производителността на моделите за машинно обучение.

Случайни липсващи данни

В случаите, когато се установява липса на записи на случаен принцип, процесът включва специфични стъпки, като, например, попълване на данни от типа „Пропуснати на случаен принцип“ (MAR). Тази фаза е жизненоважна за поддържане на точността, особено при работа с набори от данни, които съдържат произволно липсващи стойности.

Липсващи данни за даден клас

Когато липсват изходни данни за даден клас, се наблюдава необходимост от включване на допълнителен модул в модела. Този модул служи за

балансиране на набора от данни, тъй като небалансираните набори от данни могат да представляват значителни предизвикателства за моделите за машинно обучение.

Генериране на данни

В определени ситуации уникалните характеристики на данните, заедно с тяхната сложност и правни съображения, могат да възпрепятстват способността за осигуряване на моделите с необходимото количество и качество на информацията. Когато възникне недостиг на данни, се установява ясна необходимост от приложение на процес за генериране на данни.

Синтез на данни

Процесът на синтезиране на многомерен масив от синтетични данни има няколко основни фази: избор на променливи, анализ на разпределенията, извличане на бизнес логиката и приложение на генерираните данни. Последната фаза е свързана с валидирането на новополучения набор от синтетични данни.

Избор на променливи

Идентифицирането на ключовите индивидуални характеристики на потребителите на финансови услуги изисква задълбочен анализ на различни подмножества от отличителни характеристики. Всяка внимателно подбрана група променливи обогатява интерпретацията на ситуацията, в която се разглежда масива от информация, позволявайки по-цялостен и точен профил на активните потребители. Този 360-градусов подход разкрива ключови аспекти от същността на факторите, които формират финансовото поведение на индивидите.

Демографски характеристики

Разбирането на демографските характеристики е жизненоважно, тъй като хвърля светлина върху основните фактори, които влияят върху финансовото поведение в обществото. Чрез изследване на сложните причинно-следствени връзки между демографските елементи и индивидуалните черти се разкриват основните модели, които проследяват последователността в изработването на финансови решения.

Индивидуални характеристики

Индивидуалните характеристики, които се влияят както от вродени, така и от придобити качества, играят важна роля във формирането на начина, по който индивидите взаимодействат с околната среда. Чрез целенасочени задачи и оценки се предоставя възможност за ефективно идентифициране на основните черти, които имат съществено влияние на финансовото поведение на субектите.

Социално-икономически статус

Личните характеристики служат като перспектива към социално-икономическия статус на индивида. Анализът на тези отличителни черти предоставя цялостен поглед върху текущите постижения на финансово активните лица, което създава предпоставки за по-добро разбиране на техните нужди и мотивации във финансовия пейзаж.

Банкови и финансови характеристики

Разглежданите банкови и финансови характеристики разкриват поведението на индивидите на пазарите на финансови услуги. Тези характеристики позволяват профилирането на активните потребители на финансови услуги.

15.2.4 Обработка на данни

По време на процеса на симулация на данни се извършва първоначална оценка на извлечените разпределения. Всяко статистическо разпределение се определя от специфични параметри, които описват неговите характеристики, включително форма, централна тенденция, променливост и асиметрия.

№	Фактор	Код	Тип променлива	Възможни стойности	Получаване
1	Пол	sex	Номинална	М; Ж	Симулация
2	Възраст – навършени години	age	Непрекъсната	20 - 85	Корелация
3	Ниво на образование	lv_educ	Ординална	Незавършено; Начално; Основно; Средно; Висше	Симулация
4	Трудов статус	empl_stat	Номинална	Работодатели; Самонаети лица; Заети в частния сектор; Заети в публичния сектор; Неплатени семейни работници; Безработни	Симулация
5	Семейно положение	marit_stat	Номинална	Неженен/ Неомъжена; Женен/Омъжена; Разведен/Разведена; Вдовец/Вдовица	Симулация
6	Брой членове на домакинството	house_memb	Интервална	1; 2; 3; 4; 5; 6; 7+	Симулация
7	Брой деца под 18 години	chil_u_18_y	Интервална	Няма деца под 18 години; Едно дете	Симулация

№	Фактор	Код	Тип променлива	Възможни стойности	Получаване
				под 18 години; Две деца под 18 години; Три деца под 18 години; Четири деца под 18 години; Пет деца под 18 години; Шест или повече деца под 18 години	
8	Националност	nation	Номинална	България; ЕС; Други	Симулация
9	Религия	religion	Номинална	Протестант; Католик; Православен; Мюсюлманин; Друг; Без религия; Не се самоопределям	Симулация
10	Професия – индустрия	prof_ind	Номинална	Селско стопанство, горско стопанство и рибарство; Минно дело и преработвателна промишленост; Комунални услуги (електроразпределение и водоснабдяване); Строителство; Търговия, ремонт на автомобили и мотоциклети; Транспорт, складиране и пощи; Хотелиерство и ресторантьорство; Създаване и разпространение на информация и творчески продукти; Телекомуникации; Финансови и административни дейности; Държавна администрация; Образование и изследвания; Хуманно здравеопазване и социална работа; Други дейности	Симулация

№	Фактор	Код	Тип променлива	Възможни стойности	Получаване
11	Професионален статус	prof_stat	Номинална	Договор за управление; Трудов договор; Граждански договор; Самостоятелно зает; Безработен; Пенсионер	Симулация
12	Брой притежавани апартаменти/къщи	count_house	Интервална	0; 1; 2 и повече	Симулация
13	Собственост върху земя	own_field	Двоична	ДА/НЕ	Симулация
14	Автомобили на домакинство	num_car_house	Интервална	0; 1; 2; 3+	Симулация
15	Образование	edu	Номинална	Педагогически науки; Хуманитарни науки; Социални, икономически и правни науки; Природни науки, математика и информатика; Технически науки; Селскостопански науки и ветеринарна медицина; Здравеопазване и спорт; Изкуства; Сигурност и отбрана	Симулация
16	Темперамент	temp	Номинална	Холерик; Флегматик; Сангвиник; Меланхолик	Симулация
17	Индивидуални предпочитания за риск	ind_risk	Непрекъсната	0 - 1	Корелация
18	Предишен инвестиционен опит в години	invest_exp	Ординална	0; 1-5; 6-10; 11-15; 16-25	Симулация
19	Инвестиционен опит с акции	shares	Двоична	ДА/НЕ	Симулация
20	Инвестиционен опит с облигации	corp_oblig	Двоична	ДА/НЕ	Симулация
21	Инвестиционен	oth	Двоична	ДА/НЕ	Симулация

№	Фактор	Код	Тип променлива	Възможни стойности	Получаване
	опит с други активи				
22	Инвестиционен опит с инвестиционни фондове	inv_fund	Двоична	ДА/НЕ	Симулация
23	Инвестиционен опит с валути	cash	Двоична	ДА/НЕ	Симулация
24	Инвестиционен опит с криптовалути	crypto	Двоична	ДА/НЕ	Симулация
25	Инвестиционен опит с държавни ценни книжа	gov_bond	Двоична	ДА/НЕ	Симулация
26	Инвестиционен опит с банкови депозити	deposits	Двоична	ДА/НЕ	Симулация
27	Доход	income	Ординална	До 6121; До 12001; До 27601; До 43201; До 58801; До 74401; Над 90001+	Корелация
28	Лични разходи	pers_exp	Ординална	до 4500; до 5000; до 5500; до 6000	Корелация
29	Разходи за жилище	house_exp	Ординална	до 500; до 1500; до 3000; до 4000	Корелация
30	Данъци и застраховки	taxes	Ординална	до 500; до 1000; до 2000; до 2500	Корелация
31	Транспорт и комуникации	transp_telecom	Ординална	до 500; до 1000; до 1500; до 2500	Корелация
32	Свободно време и хоби	hobby	Ординална	0; до 1500; до 2000; до 3000	Корелация
33	Предпочитан метод за банкиране	banking	Номинална	Онлайн/Офлайн	Симулация
34	Среден брой банкови трансакции	bk_oprat	Ординална	До 7; От 8 до 10; От 11 до 13; От 14 до 18; От 19 до повече	Симулация
35	Дебитна карта	bk_dc	Интервална	Не е налична; Една; Две; Три	Симулация
36	Банкова сметка	bk_acc	Двоична	ДА/НЕ	Симулация

Таблица 15.2.4-1: Речник на данните - пълен списък с променливи

За да се оценят тези параметри, от дадена извадка, обикновено се използват два статистически метода: методът на моментите [11] и обобщеният метод на моментите (GMM) [12].

15.2.4.1 Метод на моментите

При метода на моментите, извадковите моменти се съпоставят с теоретичните моменти на разпределението. Този подход включва решаване на набор от уравнения за извеждане на параметрите на разпределението.

15.2.4.2 Обобщен метод на моментите (GMM)

GMM разширява метода на моментите, като предлага по-голяма гъвкавост при избора на моментни условия. Тази техника е особено полезна, когато има повече моментни условия, отколкото параметри, или когато моментните условия не могат да бъдат решени директно.

15.2.5 Достъпност

Спецификата, комплексността и регулаторната рамка, обграждаща данните, представляват значителни предизвикателства при получаването на необходимото количество и качество на информацията, с цел разработване на модели. Информацията, която е потребна представлява интерес за европейските регламенти, като например GDPR и Закона за банковата тайна, наред с други.

Събирането на данни чрез конвенционални методи се разглежда като процес с висока степен на сложност. Предвид тези предизвикателства, е идентифицирана алтернативна стратегия за получаване на необходимия обем данни. Този подход включва генериране на многовариантен набор от данни, който съдържа различни демографски, лични, индивидуални и банкови променливи и показатели [8].

15.2.6 Формат на данните

Новополученият набор от данни е достъпен в широк набор от различни формати. Възможностите, предлагани от разглеждания модел, са както следва: Excel, CSV, Pandas, Croissant, Polars и Parquet. Достъпът до множество различни формати предполага по-голям набор от опции за изследователите, които се нуждаят от многомерен синтетичен набор от данни.

15.2.7 Осигуряване на качеството

Бизнес логика при генериране на данни

Бизнес логиката, прилагана в процеса на генериране на данни, се определя посредством идентифициране на потенциални взаимозависими фактори, техните ограничения, възможните и невъзможни комбинации от тези фактори. При комбиниране на разпределения съществува риск от получаване на недостижими или много слабо вероятни стойности. Следователно, включването на бизнес логика в модела е от съществено значение за филтриране на нежелани, малко вероятни и невъзможни комбинации. Интегрирането на бизнес логиката в модела протича през три ключови фази:

Избор на потенциални взаимозависими фактори: Идентифициране на съответните променливи, които си влияят взаимно.

Установяване на възможни и невъзможни комбинации: Определяне кои комбинации, от тези фактори, са осъществими и кои не.

Елиминиране на невъзможни комбинации: Премахване на всички комбинации, които не отговарят на установените критерии за поддържане на целостта на данните.

№	Независима х-ка	Стойност независима х-ка	Зависима х-ка	Филтър зависима х-ка	Забележка
1	Семейно положение	Женен/а	Брой членове на домакинството	>2	Броят на членовете в семейните домакинства е по-вероятно да бъде по-голям от 2
2	Професия – индустрия	Финансови и административни дейности	Банкова сметка	>0	По-вероятно е да притежават банкова сметка
3	Възраст – навършени години	<25	Предишен инвестиционен опит в години	0	Под 24 години е по-малко вероятно да имат инвестиционен опит. Между 35-44 и 45-54 години е по-вероятно да имат богат инвестиционен опит.
4	Възраст – навършени години	<21	Ниво на образование	<По-високо	Под 21 години е по-малко вероятно да имат университетска диплома
5	Възраст – навършени години	<25	Брой деца под 18 години	<2	Предвид определеното демографско покритие, от 20 до 24 години, е по-малко вероятно да има повече от 1 дете под 18 години.

6	Предишен инвестиционен опит в години	>0	Инвестиционен опит с банкови депозити	Y	По-вероятно е да притежават банкова сметка
7	Инвестиране в акции	Y	Предишен инвестиционен опит в години	>0	Ако инвестицията в акции = да, тогава предишният инвестиционен опит в години е >0.
8	Инвестиции в облигации	Y	Предишен инвестиционен опит в години	>0	Ако инвестицията в облигации = да, тогава предишният инвестиционен опит в години е >0.
9	Други инвестиции	Y	Предишен инвестиционен опит в години	>0	Ако инвестицията в други инвестиции = да, тогава предишният инвестиционен опит в години е >0.
10	Инвестиции във фонд	Y	Предишен инвестиционен опит в години	>0	Ако инвестицията във фондове = да, тогава предишният инвестиционен опит в години е >0.
11	Валутни инвестиции	Y	Предишен инвестиционен опит в години	>0	Ако инвестицията във валута = да, тогава предишният инвестиционен опит в години е >0.
12	Инвестиции в криптовалути	Y	Предишен инвестиционен опит в години	>0	Ако инвестицията в криптовалути = да, тогава предишният инвестиционен опит в години е >0.
13	Инвестиции в държавни ценни книжа	Y	Предишен инвестиционен опит в години	>0	Ако инвестицията в държавни ценни книжа = да, тогава предишният инвестиционен опит в години е >0.
14	Възраст – навършени години	<25	Банкова сметка	H	Под 24 години е по-малко вероятно да имат разплащателна сметка
15	Възраст – навършени години	<18	Банкова сметка	H	Лица под 18 години не могат да имат разплащателна сметка
16	Ниво на образование	Висше	Доход	>27601	По-високото ниво на образование предполага доходи в горния диапазон
17	Брой деца под	>1	Брой членове на	>3	Броят на членовете на

	18 години		домакинството		домакинството е пряко зависим от броя на децата под 18 години
18	Доход	>27601	Данъци и застраховки	>2500	Доходите в горния диапазон съответстват на по-високи данъци и осигуровки

Таблица 15.2.7-1 Пример за бизнес логика

15.2.8 Надеждност

Процес на валидиране. Валидирането на предложения модел за симулиране на големи набори от данни включва няколко ключови етапа, които изискват повишено внимание [13].

Анализ на данните. Този етап обхваща задълбочено проучване на синтезираните данни, заедно със сравнение на различни източници на информация, за да се гарантира точност и надеждност.

Валидиране на данните. По време на тази фаза е изключително важно да се провери дали генерираният набор от данни съответства на оригиналните статистически разпределения, интегрирани в симулационния модел, или са получени посредством проучване.

Съседни честоти. За новогенерираните данни е необходимо да се обърне специално внимание на съседните честоти. Плавният преход между тези стойности е жизненоважен за валидирането на модела, тъй като помага да се избегнат резки колебания. Подобни флуктуации свидетелстват за наличието на проблеми в процеса на симулация.

Оценка на качеството. Като последна стъпка в процеса, от съществена важност е да се оцени качеството на получената информация. Извършва се анализ на новите разпределения, за да се определи дали те адекватно отразяват характеристиките на оригиналните разпределения.

15.2.9 Точност

Тестът на Колмогоров-Смирнов (КС) се използва като основен метод за валидиране на новополучените данни. КС тестът е непараметричен статистически тест, който изпълнява две основни функции: първо, той оценява съвместимостта между разпределение на извадката и теоретично разпределение - в КС теста с една извадка. Втората цел, за която се използва, е да се изследва сходството между две разпределения на извадката - в КС теста с две извадки. КС тестът използва емпиричната кумулативна функция на разпределение (ECDF), за да определи дали даден набор от данни съответства на определено теоретично разпределение.

Едноизвадков KS тест

KS тестът за една извадка сравнява ECDF на извадка с кумулативната функция на разпределение (CDF) на теоретично разпределение. ECDF представлява дела на наблюденията в извадката, които са по-малки или равни на определена стойност, докато CDF показва теоретичната вероятност за получаване на случайно наблюдение от това разпределение, което също е по-малко или равно на тази стойност. Този тест е приложим, когато основното разпределение е хипотетично или е изведено от предходни теории. Тестовата статистика D се изчислява като максималната абсолютна разлика между ECDF на извадката и теоретичния CDF [14].

Двуизвадков KS тест

Двуизвадковият KS тест оценява дали има значимо съответствие между две едномерни вероятностни разпределения. Този метод е особено чувствителен към разликите в разпределенията, от които са извлечени двете извадки. Тестовата статистика D за този тест се определя като максималната абсолютна разлика между двете ECDF [15].

Хипотези

Както при еднопробния, така и при двупробния KS тест, нулевата хипотеза (H0) постулира, че извадката(ите) съответства(т) на определеното разпределение (за една извадка) или че и двете извадки произлизат от едно и също разпределение (за две извадки). Обратно, алтернативната хипотеза (H1) твърди, че извадката(ите) не съответства(т) на определеното разпределение (за една извадка) или че двете извадки произлизат от различни разпределения (за две извадки).

15.2.10 Честота на актуализиране

Данните трябва да бъдат получени еднократно, като се използва метод за генериране на големи количества информация. В случай на промяна в общите условия за основните групи от разглеждани променливи, може да се предвиди повторно генериране на набора от данни. Пример за такъв случай може да бъде провеждането на следващото преброяване (през 2030/2031 г.).

15.3 Метаданни за статистически характеристики

Що се отнася до разглеждания подход за симулиране на многомерен масив от синтетични данни, е необходимо да се подготви подробно описание на всяка от разглежданите характеристики. Метаданните предоставят основен

контекст и документация за статистическите данни. Те обхващат структурирана информация, която описва данните, техните процеси и методологии, което помага за ефективното разбиране, интерпретиране и използване на статистическата информация. Пълен списък на подробните променливи, съдържащи се в генерирания набор от данни, е наличен в Приложение 2.

15.4 Заключение

Настоящата глава акцентира върху ролята на метаданните в контекста на генерирането на многоизмерен синтетичен набор от данни с помощта на алгоритми. Тя допринася за продължаващата дискусия относно синтеза на информационни масиви в областта на науката за данните, като представя иновативни методи за смекчаване на недостига на данни, като същевременно се придържа към най-добрите практики на Евростат за описание на метаданни. Предоставен е цялостен анализ на всички използвани променливи, като са подчертани обосновката за включването им в синтезирания набор от данни, в съответствие с целите на изследването.

Проучването подчертава предизвикателствата, свързани с получаването на висококачествени набори от данни за модели, базирани на данни, свързани със сложността на характеристиките на данните и строгите законови разпоредби, като GDPR и Закона за банковата тайна. Традиционните методи за събиране на информация често се оказват неадекватни, което води до необходимостта от алтернативни стратегии. Това изследване въвежда нов подход, който генерира многовариантна синтетична информационна база, обхващаща различни демографски, лични и банкови променливи. Чрез използване на симулационни техники и вероятностни методи за конкатенация е синтезиран наборът от данни, включващ 36 различни променливи, категоризирани в съответните групи. Констатациите подчертават важността на този процес по отношение на преодоляването на законовите бариери пред достъпа до чувствителна информация, като по този начин се позволява ефективно моделиране и анализ във финансов и демографски контекст.

Използвана литература

[1] Marchev, V., Marchev, A., 2021. “Methods for Simulating Multi-dimensional Data for Financial Services Recommendation,” Bulgarian economic paper, ISSN: 2367-7082

[1] Marchev, V., Marchev, A., Piryankova, M., Masarliev, D., & Mitkov, V., 2023. Synthesizing an anonymized multidimensional dataset featuring financial, economic, demographic, and personal traits data. Vanguard Scientific Instruments in Management, vol. 19, no. 1, 2023, ISSN 1314-0582, 79-99.
[2] NSI, 2021, Census 2021, Sofia, https://census2021.bg
[3] Tipatov, N., 2009, "Statistika po kommentariyam k testu Ayzenka", Biznes Trener, http://b-t.com.ua/test_ayzenk_opis.html (available in Russian language)
[4] BNB, 2024, Statistical Database, Selection of statistics, https://www.bnb.bg/statistics/index.htm?toLang=_EN
[5] FSC, 2024, Insurance Activity, Statistics, https://www.fsc.bg/en/insurance-activity/statistics/
[6] Ministry of Finance, 2024, Economic Policy, Analyses and Research, https://www.minfin.bg/en/865
[7] Ministry of Agriculture and Food, 2024, Statistics and Analyses, https://www.mzh.government.bg/en/statistics-and-analyses/
[8] Marchev, V., Marchev, A. (2025). Anonymizing Personal Information Using Distribution-Based Data Synthesis. In: El Khatib, M., Alzoubi, H. M., Angelova, Y. (eds) Sustainable Innovation for Engineering Management (SIEM). ISCME 2024. Advances in Science, Technology & Innovation. Springer, Cham. https://doi.org/10.1007/978-3-031-89889-1_38
[9] Nikhil. 2024. Google AI Introduces CodecLM: A Machine Learning Framework for Generating High-Quality Synthetic Data for LLM Alignment], [How to Use Synthetic and Simulated Data Effectively; https://towardsdatascience.com/how-to-use-synthetic-and-simulated-data-effectively-04d8582b6f88
[10] Pearson, K., 1936. Method of Moments and Method of Maximum Likelihood, Biometrika 28(1/2), 35–59.
[11] Hansen, L., P., 1982. Large Sample Properties of Generalized Method of Moments Estimators, Econometrica, Vol. 50, No. 4 (July 1982)
[12] Marchev, A., Marchev, V., 2022. Synthesizing multi-dimensional personal data sets, AIP Conference Proceedings, 2505 (1): 020012. https://doi.org/10.1063/5.0100615, 2022
[13] Brown, J., Harvey, M., 2008, "Rational Arithmetic Mathematica Functions to Evaluate the Two-Sided One Sample K-S Cumulative Sampling Distribution", Journal of Statistical Software, Volume 26, Issue 2.
[14] Simard, R., L'Ecuyer, P., 2010, "Computing the Two-Sided Kolmogorov-Smirnov Distribution", Journal of Statistical Software.
[15] Infostat, Tertiary education graduates by educational-qualification degree, sex and narrow field of education (foet), 2001 – 2016, Report result

[16] Marchev JR, A., Marchev, V., Haralampiev, K., Efremov, A., Markov, B., Lyubchev, D., Piryankova, M., Filipov, B., Masarliev, D., & Mitkov, V. (2025), Synthesized Economic Agents Dataset, HuggingFace.co, DOI address https://doi.org/10.57967/hf/3701

V. ЗАКЛЮЧЕНИЕ

В монографията е направен подробен преглед на съвкупност от подходи за генериране на синтетични информационни масиви, с цел синтезиране на многомерен набор от социално-финансови и демографски данни с фокус към преодоляване на строгите регулаторни ограничения, в следствие на GDPR и Закона за кредитните институции (в частта му уреждаща банковата тайна).

Големите данни по своята същност представляват сложен и многоаспектен информационен актив, изискващ специализиран статистически инструментариум. Ефективното извличане на скрити зависимости от структури с подобен характер изисква специализиран статистически софтуер и алгоритми за машинно обучение. От своя страна, анализът на чувствителна информация е изправен пред строги регулаторни ограничения, произтичащи от GDPR и националното законодателство. Критичен аспект в настоящия труд е балансът между поверителността на данните и тяхното трансформиране в стратегически ресурс. Разработването на прогностични модели върху исторически масиви трябва строго да съблюдава установената правна рамка за финансова сигурност.

Работната дефиниция на големите данни в настоящия труд се фокусира върху многомерни масиви, произтичащи от исторически събития или алгоритмични процеси, които носят ценна вторична информация. Анализът на тези структури изисква прилагането на специализиран статистически софтуер и мащабируеми изчислителни архитектури за паралелна обработка. Чрез използването на усъвършенствани алгоритми се идентифицират скрити зависимости, които остават незабележими при традиционните хомогенни извадки, с фокус към ефективното извличане на знания от масиви с висока размерност на характеристиките.

При възникване на необходимост от обучаване на модели основани на големи информационни масиви, традиционните методи за информационно обезпечаване, включващи анкети, интервюта и системно наблюдение, осигуряващи детайлни качествени данни стават все трудно приложими. Въпреки своята прецизност, тези подходи са времеемки и финансово неизгодни. Освен това трудно генерират мащабни масиви за обучение на съвременни модели. Допълнително предизвикателство представляват строгите регулаторни рамки на GDPR и защитата на банковата тайна, които ограничават достъпа до реална чувствителна информация. Следователно класическите инструменти често се оказват недостатъчни за нуждите на високотехнологичния прогнозен анализ.

Като ефективна алтернатива се установява прилагането на статистически подходи за симулация на многомерни синтетични масиви. Тези модели

интегрират йерархична структура от демографски, социално-икономически и финансови променливи, базирани върху строга бизнес логика. Чрез възпроизвеждане на реални корелационни зависимости се постига висока автентичност на потребителските профили без наличие на правни рискове. Този подход елиминира субективността и осигурява надеждна информационна база за изследване на пазарни тенденции в контролирана среда.

Методите за генериране на синтетични данни предлагат устойчиво решение базирано на статистически принципи интегрирани в алгоритми за създаване на многомерни набори от информация. Тези синтетични масиви запазват статистическите зависимости между променливите, без да идентифицират реални лица или да нарушават поверителността. Технологията позволява ефективна обработка на небалансирани данни и значително намаляване на разходите за събиране на ресурси. Генерираната синтетична информация служи като икономически изгодна и безопасна алтернатива на скъпите процеси по съхранение на реални записи.

Синтетичните данни намират широко приложение в процесите по тестване, валидиране и трениране на алгоритми за машинно обучение. Основно предимство на тези данни е възможността за мащабиране, осигурявайки необходимия информационен масив в сфери, за които е характерно наличието на дефицит на реални записи (или реалната информация не може да бъде използвана поради регулаторни причини). Генерираният набор от данни създава предпоставки за прецизна персонализация според конкретните изследователски нужди и технически изисквания. Посредством въвеждане на контролирани вариации се постига оптимално балансирано представяне на различните класове в извадката.

Във финансовия сектор защитата на чувствителна информация е критичен приоритет, наложен от строги законови разпоредби за банковата тайна. Синтетичните набори от данни позволяват на институциите да провеждат сложни анализи за борба с прането на пари, без да компрометират поверителността. Посредством набор от усъвършенствани статистически техники се изграждат реалистични модели, които преодоляват регулаторните ограничения и повишават надеждността на прогнозите.

Синтетичните масиви представляват стратегическо решение за елиминиране на феномена „студен старт“ при внедряването на нови аналитични системи. Организациите могат да обучават персонализирани препоръчителни алгоритми, симулирайки потребителско поведение без необходимост от реална историческа информация. Този метод значително ускорява иновационните процеси и гарантира статистическа правдоподобност за бизнес анализите. Процесите по синтез на данни минимизират риска от

генериране на погрешни стратегии в началната фаза на експлоатация на нови услуги.

Генерирането на синтетични данни представлява комплексна задача, изискваща сложни алгоритми и задълбочено разбиране на статистическите взаимовръзки. Критичен момент в процеса представлява осигуряването на фактология и прецизност, тъй като неверните или халюцинирани масиви могат да доведат до погрешни изводи при приложение в реални сценарии. Процесът изисква внимателно проектиране и валидиране, за да се избегне усилването на съществуващи отклонения или въвеждането на нови такива. Запазването на оригиналните зависимости в данните е от фундаментално значение за успешното прилагане на моделите.

Процесът по валидиране на синтетични данни е критичен момент в последователността по синтезиране на данни. Тази фаза обезпечава статистическата автентичност и аналитична стойност на новите масиви. Изготвя се оценка на общата полезност чрез сравнение на разпределенията и специфично съответствие при конкретни модели. Основната цел е запазването на съвместното разпределение на променливите и тяхната корелационна структура спрямо оригиналния масив.

Валидирането функционира като систематичен механизъм за контрол на качеството, който балансира между структурната цялост и поверителността. Този етап включва детайлен анализ на честотите за избягване на резки колебания и аномалии в новополучените данни. Чрез количествено измерване на прецизността се минимизира рискът от алгоритмични отклонения и систематични грешки.

Друг критичен момент в процеса по синтез на данни представлява обезпечаване на модела с необходимата бизнес логика. Бизнес логиката функционира като основен архитектурен слой, който трансформира генерираните масиви в съгласувани стратегически ресурси. Тя осигурява семантична цялост и контекстуална валидност на синтезираните данни, като преодолява структурните различия чрез формални логически съответствия. Чрез дефиниране на експертни ограничения и реалистични зависимости, алгоритмите гарантират надеждността на изводите. Прилагането на тези бизнес правила позволява деривация на нови показатели и автоматизирано генериране на масиви, отговарящи на оперативната реалност. Правилното структуриране на бизнес логиката минимизира рисковете при интеграция и улеснява адаптацията към динамичната пазарна среда.

По-нататък се разглеждат основните методически концепции свързани с процесите по генериране на многомерни синтетични набори от данни.

Процес за генериране на многомерни информационни масиви се пречупва през призмата на методологично дърво с три основни функционални направления: обогатяване, интерполация и синтез. Обогатяването на данни се ръководи от принципа на необходимото разнообразие, докато интерполацията е критична за коректното справяне със случайно липсващи стойности в масивите. Синтезът на данни се представлява съществен фактор в случаите, при които високата сложност и регулаторните закономерности възпрепятстват осигуряването на нужното количество и качество на реална информация. Прилагането на тази систематизирана рамка гарантира аналитичната мощ на моделите чрез преодоляване на информационния дефицит.

В процеса на работа се установява необходимостта от провеждане на предварителни одити с цел минимизиране на обема данни за синтез и оптимизиране на управлението на ресурсите. Този подход намалява изчислителните разходи и времето за обработка, като същевременно създава предпоставки за спазването на регулаторните принципи за защита на информацията. Рационализирането на масивите опростява последващата валидация и позволява по-прецизно идентифициране на статистически несъответствия. Чрез ограничаване на размерността се минимизира експоненциалният риск от грешки при възпроизвеждането на сложните взаимовръзки между променливите.

В рамките на методическата част от монографията е обособен **набор от подходи за генериране на голям масив от синтетична информация**. Съвкупността обхваща разнообразни методи включващи генериране на данни посредством: общи разпределения бизнес логика, общи разпределения и корелации, генеративни състезателни мрежи. Също така, синтез на данни при наличие на частични информационни набори, при наличие на части от масиви и посредством съвместно разпределение.

Процесът по изграждане на методическите аспекти обхваща **създаването на концепция за изработване на модел за генериране многомерен информационен масив от синтетични данни** за използване в банковата и финансовата системи. Използваните алгоритми предоставят набор от възможности за генериране на голям масив от данни в зависимост от специфични характеристики (демографски данни, финансови и банкови променливи, лични характеристики). Предоставя се подробно описание на различните методи като е налична и достъпна информация за стъпките, през които е необходимо да се премине за осъществяване на различните подходи.

Осъществяването на посочените модели се базира на създаването на широк набор от комплексни алгоритми, обединени посредством специализирани софтуерни системи. За целта е разработено програмно

осигуряване за компютърна симулационна среда, която създава предпоставките за висока степен на автоматизация при генерирането и моделирането на синтетични масиви от данни. Процесът се обезпечава чрез комбинация на множество различни софтуерни продукти. Интеграцията на софтуерите предоставя възможност за изграждане на завършена система, удовлетворяваща разнообразната необходимост от данни.

В рамките на IV глава е направен **подробен разрез на всички подходи за симулиране на големи бази данни**, като са описани проведените симулационни експерименти с моделираните процеси. В рамките на изследването е **генериран многомерен информационен масив от синтетични данни**, като е предоставена и детайлна разбивка на **метаданните**, както и публично достъпен масив от данни за бизнес, образователни и изследователски цели под лиценз на MIT (валидиран набор от данни https://doi.org/10.57967/hf/3701).

Разгледаните и анализирани подходи са обобщени в следващите редове.

Анализирана е методология за алгоритмично генериране на синтетични данни, предназначени за валидиране и обучение на модели за машинно обучение посредством методи, които не са основани върху математически вероятности. включващи обратна копула и разлагане на Холецки. Тези подходи се базират върху съхранение на структурните зависимости и ковариациите. Разработеният алгоритъм цели създаването на синтетични масиви, които репродуцират както маргиналните разпределения, така и корелационните връзки. Валидацията се извършва чрез теста на Колмогоров-Смирнов (KS), а представеният емпиричен модел потвърждава ефикасността на подхода спрямо оригиналните параметри.

Систематизирана е методологична рамка за генериране на синтетични информационни структури, целяща преодоляване на дефицита от чувствителни данни при статистическото моделиране. Разработеният подход включва последователност от процеси, насочени към разширяване на ограничен първоначален масив, посредством интегриране на зависимости, произтичащи от конкретната бизнес логика. Внедреният алгоритъм дефинира взаимовръзките между променливите чрез експертни ограничения и тегловни коефициенти за мащабиране. Валидацията обхваща серия от тестове за качество и съгласуваност, базирани на хоризонтална синхронизация и вертикална верификация.

Въведена е унифицирана рамка за валидиране на синтетични данни, разграничавайки статистическата правдоподобност от бизнес допустимостта. Методологията формализира три независими валидатора: експертен одит

(HECA) за съответствие с политиките, програмен LLM анализ (LLM-PA) за мащабируема проверка и автоматизирано LOFO калибриране за техническа прецизност. Сравнителният анализ показва, че всеки инструмент адресира специфични рискове, свързани с управлението, производителността и вътрешната съгласуваност на моделите. Изследването дефинира критерии за приемане и предоставя практически насоки за регуларизация и избор на извадки при генерирането.

Следващият подход за генериране на синтетични масиви е авторски метод, който анализира перспективите за обединяване на множествени информационни масиви въз основа на идентифицирани повтарящи се характеристики. Фокусът е поставен върху реализирането на хибриден модел за вероятностно конкатениране на записи, който минимизира грешките при сливане на хетерогенни структури от данни. Акцентът е поставен върху бинарната трансформация на ключови параметри, която гарантира алгоритмична стабилност и унификация на променливите. Представения алгоритъм създава предпоставки за поетапно надграждане на масива чрез включване на разнородни характеристики. Добавен е алгоритъм за запазване на анонимността и филтриране на данните.

Реализиран и апробиран е подход за анонимизиране на информационни масиви посредством приложение на генеративни състезателни мрежи (GAN). Процесът обхваща фази на трансформиране на реални масиви в статистически достоверни синтетични аналози. Използва се итеративно обучение между генератор и дискриминатор за постигане на високо сходство с оригиналните данни, като същевременно се използват мерки за анонимизация. Разгледани са архитектурата на невронните мрежи и процедурите по обучение, които осигуряват техническата прецизност на анонимизацията. Внедряването на този подход позволява извличане на ценни аналитични изводи при пълно спазване на принципите за конфиденциалност.

Разгледан е апробиран иновативен подход, базиран на обърната копула за синтез на информационни масиви, който ефективно моделира сложни многомерни зависимости между променливите. Експерименталните резултати демонстрират висока прецизност при репродуциране на маргиналните разпределения и структурните връзки в данните. Процесът по валидация е осъществена чрез теста на Колмогоров-Смирнов и сравнителен анализ потвърждава статистическата устойчивост и гъвкавост на предложената методология.

Разгледана е и подробно анализирана методология за генериране на синтетични информационни масиви, чрез интегриране на статистически методи, като: симулация Монте Карло и разлагане на Холецки, както и на

предварително синтезирана бизнес логика. Структуриран е цялостен подход за създаване на многомерни набори със синтетични променливи. Разработен е модел, който симулира профили на финансово активни лица, обхващайки демографски, персонални и банкови характеристики в съответствие с правните изисквания за конфиденциалност. Алгоритъмът се основава на статистически разпределения извлечени от реални данни или посредством комбинации с правдоподобни допускания, основани на априорна информация. Прилагането на теста на Колмогоров-Смирнов за валидация гарантира статистическата точност на синтезираната среда и запазването на структурните взаимовръзки между атрибутите.

В края на изследването е добавена глава, детайлизираща описанието на метаданните в рамките на приложения алгоритъм за генериране на многомерни синтетични набори от данни. Анализирани са предизвикателствата при използването на обширни масиви от чувствителна информация, регулирана от GDPR и Закона за кредитните институции (в частта му уреждаща банковата тайна). Разгледани са симулационните техники за създаване на многомерен синтетичен набор от данни в размер от 36 променливи, обхващащи демографски, персонални и банкови показатели, базирани на официални източници като НСИ, БНБ и Преброяването от 2021 г. Процесът акцентира върху валидирането на генерираните данни съгласно стандартите за качество и най-добрите практики на Евростат за описание на метаданните. Изследването предлага набор от иновативни стратегии за справяне с информационния дефицит в работата с големи информационни масиви, осигурявайки статистическа обоснованост на обобщения набор от информация.

В процеса на работа са установени **редица изводи**. Ефективността на процеса по синтез на данни се определя от прилагането на последователен методологичен контрол обхващащ редица етапи.

Първоначалният анализ и подбор на минимален брой необходими променливи се установява като критичен фактор в процеса по оптимизация на изчислителните ресурси. Чрез този подход се създават предпоставки за намаляване на процедурите по валидация на данните, като съществено се минимизира риска от неоснователно елиминиране на данни, запазвайки информативната стойност на масива. Генерирането на информацията се основава на целеви подбор на методи, чиято неизчерпателна класификация в монографията предоставя инструментариум за адаптация към различни икономически сценарии.

Съществен аспект в разработката представлява необходимостта от хоризонтална синхронизация. Чрез прилагане на специализирана методология

(бизнес правила, мащабиране и съвместни разпределени) се гарантира, че всеки отделен запис съответства реалните информационни масиви, което подсигурява логическата цялост на информацията.

Друг извод, който може да се направи е свързан с вертикалното валидиране на характеристиките посредством статистически тестове (като теста на Колмогоров-Смирнов). Посредством тази фаза се потвърждава, че всяка променлива стриктно следва очакваното теоретично разпределение.

Обобщената оценка на синтезирания масив, осъществена чрез метрики за косинусово подобие, дава възможност за обективно потвърждаване на качеството на целия набор от данни.

В процеса на работа е се установява, че качество на генерирания масив от синтетични данни е в пряка корелация с уместността на избрания генеративен модел спрямо спецификата на моделираната бизнес среда.

Като следствие от апробацията на модела се откриват множество **възможности за задълбочаване на изследването** в разнообразни посоки. Като най-съществените и значими задачи за бъдещи изследвания могат да бъдат определени: автоматизиране на процеса по извеждане на разпределения, интеграция на изкуствен интелект при синтезиране на теглата на променливите при подхода за генериране на данни посредством корелация и бизнес логика и интеграция на изкуствен интелект в процеса по подбор на променливи, в зависимост от разглежданата бизнес среда.

VI. ОБОБЩЕНА БИБЛИОГРАФИЯ

Astolfi, G. (2021). Generating tabular data using generative adversarial networks with differential privacy, CONFERENCE OF EUROPEAN STATISTICIANS, Expert Meeting on Statistical Data Confidentiality

Barbierato, E., Vedova, M. L. D., Tessera, D., Toti, D., and Vanoli, N., (2022). A methodology for controlling bias and fairness in synthetic data generation. Applied Sciences, 12(9):4619

Benali, F., Bodénès, D., Labroche, N., de Runz, C., 2021, MTCopula: Synthetic Complex Data Generation Using Copula, International Workshop on Data Warehousing and OLAP, (2021).

Berman, F., Fox, G., Hey, T., (2003) Grid Computing: Making the Global Infrastructure a Reality. John Wiley and Sons. P747-771.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. A New Form of Web Content That Is Meaningful to Computers Will Unleash a Revolution of New Possibilities. Scientific American, 284, 1-5.

Bernstein, P. A., & Melnik, S. (2007). Model management 2.0: Manipulating richer objects. In Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data (pp. 1–12). Association for Computing Machinery. https://doi.org/10.1145/1247480.1247482

Boyce, C., & Neale, P. (2006). Conducting in-depth interviews: A guide for designing and conducting in-depth interviews for evaluation input. Pathfinder International.

Boyd, D., & Crawford, K. (2012). Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication & Society, 15(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Brown, J., Harvey, M., 2008, “Rational Arithmetic Mathematica Functions to Evaluate the Two-Sided One Sample K-S Cumulative Sampling Distribution”, Journal of Statistical Software, Volume 26, Issue 2.

b-t.com.ua, 2009, Statistika po kommentariyam k testu Ayzenka, http://b-t.com.ua/test_ayzenk_rez_komment.html

Bulgarian National Bank, 2023, Home, Statistics, Monetary and Interest Rate Statistics, Loans and Deposits by Amount Category and Economic Activity

Bulgarian National Bank, 2024, Statistical Database, Selection of statistics, https://www.bnb.bg/statistics/index.htm?toLang=_EN

Ceri, S., Bozzon, A., Brambilla, M., Della Valle, E., Fraternali, P., & Quarteroni, S. (2013). Web Information Retrieval. Springer Science & Business Media

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357

Chen, H., Chiang, R. H. and Storey, V. C. (2012) Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly, 36, 1165-1188

Christen, P. (2012) Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer Science and Business Media, Berlin.

Claeskens, G., & Hjort, N. L. (2008). Model selection and model averaging. Cambridge University Press.

Clayton, D., 1978, A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence, Biometrika, Volume 65, Issue 1, Pages 141–151, https://doi.org/10.1093/biomet/65.1.141.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37–46. https://doi.org/10.1177/001316446002000104

Conover, W. J. (1999). Practical nonparametric statistics (3rd ed.). John Wiley & Sons.

Couper, M. P. (2000). Web surveys: A review of issues and approaches. Public Opinion Quarterly.

Darling, D. A. (1957). The Kolmogorov-Smirnov, Anderson-Darling tests of goodness of fit. The Annals of Mathematical Statistics, 28(4), 823–838

Davenport, T. H., & Harris, J. (2017). Competing on Analytics: Updated, with a New Introduction: The New Science of Winning. Harvard Business Review Press.

De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features.

Dean, J. and Ghemawat, S. (2008) MapReduce: Simplified Data Processing on Large Clusters. Communications of the ACM, 51, 107-113. https://doi.org/10.1145/1327452.1327492

Dehghan, P., Alashwal, H., & Moustafa, A. A. (2022). Applications of machine learning to behavioral sciences: Focus on categorical data. Discovery Psychology, 2, Article 22. https://doi.org/10.1007/s44202-022-00027-5

Del Moral, P., Doucet, A., & Jasra, A. (2006). Sequential Monte Carlo samplers. Journal of the Royal Statistical Society Series B: Statistical Methodology, 68(3), 411–436. https://doi.org/10.1111/j.1467-9868.2006.00553.x

Dereniowsky, D., & Kubale, M. (2004). Cholesky factorization of matrices in parallel and ranking of graphs. In R. Wyrzykowski, J. Dongarra, M. Paprzycki, & J. Waśniewski (Eds.), Parallel Processing and Applied Mathematics (pp. 985–992). Springer.

Dong, X. L., & Srivastava, D. (2013). Big data integration. In Proceedings of the 2013 IEEE 29th International Conference on Data Engineering (ICDE), 1245-1248. DOI: 10.1109/ICDE.2013.6544914

Drechsler, J. (2011). Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation. Springer Series in Statistics

Drew, J., H., Glen, A., G., Leemis, L., M., 2000, Computing the cumulative distribution function of the Kolmogorov–Smirnov statistic, Computational Statistics & Data Analysis 34 (2000) 1–15. https://doi.org/10.1016/S0167-9473(99)00069-9.

Du, B., Qian, K., Claudel, C., & Sun, D. (2022). Jacobi-style iteration for distributed submodular maximization. IEEE Transactions on Automatic Control, 67(9), 4687–4702.

Dwork, C. (2008). Differential privacy: A survey of results. In M. Agrawal, D. Du, Z. Duan, & A. Li (Eds.), Theory and Applications of Models of Computation (pp. 1–19). Springer. https://doi.org/10.1007/978-3-540-79228-4_1

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science, 9(3–4), 211–407. https://doi.org/10.1561/0400000042

El Emam, K., Mosquera, L., & Hoptroff, R. (2020). Practical Synthetic Data Generation: Balancing Privacy and Utility in Data Sharing. O'Reilly Media.

Erevelles, S., Fukawa, N., & Swayne, L. E. (2016). Big Data consumer analytics and the transformation of marketing. Journal of Business Research, 69(2), 897–904. https://doi.org/10.1016/j.jbusres.2015.07.001

Erl, T. (2005) Service-Oriented Architecture: Concepts, Technology, and Design. Prentice Hall, Upper Saddle River.

European Data Protection Board. (2020, January 30). Guidelines 3/2019 on processing of personal data through video devices (Version 2.0). https://www.edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-32019-processing-personal-data-through-video_en

Fang, R., Li, X., Li, L., 2016, Generalized multivariate Gumbel distributions - Dependence, aging properties and applications. J Syst Sci Complex 29, 1752–1772 (2016). https://doi.org/10.1007/s11424-016-4272-8.

Figueira, A., & Vaz, B. (2022). Survey on synthetic data generation, evaluation methods and GANs. Mathematics, 10(15), 2733. https://doi.org/10.3390/math10152733

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin, 76(5), 378–382. https://doi.org/10.1037/h0031619

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. Harvard Data Science Review, 1(1). https://doi.org/10.1162/99608f92.8cd550d1

Fowler, M. (2002). Patterns of enterprise application architecture. Addison-Wesley Professional

FSC, 2024, Insurance Activity, Statistics, https://www.fsc.bg/en/insurance-activity/statistics/

Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4). https://doi.org/10.1145/2523813

Gandomi, A. and Haider, M. (2015) Beyond the Hype: Big Data Concepts, Methods, and Analytics. International Journal of Information Management, 35, 137-144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Gantz, J., & Reinsel, D. (2011). Extracting Value from Chaos.

Goldberger, A. S. (1964). Econometric theory. John Wiley & Sons.

Golfarelli, M., & Rizzi, S. (2009). Data Warehouse Design: Modern Principles and Methodologies. McGraw-Hill Education

Goodfellow, I. (2017). "NIPS 2016 Tutorial: Generative Adversarial Networks". arXiv:1701.00160

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014) Generative Adversarial Nets. Proceedings of the 27th International Conference on Neural Information Processing Systems, 2, 2672-2680.

Grolinger, K., Higashino, W. A., Tiwari, A., & Capretz, M. A. M. (2013). Data management in cloud environments: NoSQL and NewSQL data stores. Journal of Cloud Computing: Advances, Systems and Applications, 2(1), 22. https://doi.org/10.1186/2192-113X-2-22

Gruber, M. H. J. (1998). Improving efficiency by shrinkage: The James–Stein and ridge regression estimators (pp. 7–15). CRC Press.

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., Wang, Y., Gao, W., Ni, L., & Guo, J. (2024), A survey on LLM-as-a-judge. arXiv. https://doi.org/10.48550/arXiv.2411.15594

Guarnera, L., Giudice, O., and Battiato, O., (2020). Deepfake detection by analyzing convolutional traces. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pages 666–667.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A. (2017). Improved Training of Wasserstein GANs, n NIPS

Gupta, A., Bhatt, D., and Pandey, A. (2021). Transitioning from real to synthetic data: Quantifying the bias in model. ArXiv preprint, abs/2105.04144, 2021. URL https://arxiv.org/abs/2105.04144.

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. Morgan Kaufmann.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. Econometrica, 50(4), 1029–1054. https://doi.org/10.2307/1912775

Harron, K., Goldstein, H., & Dibben, C. (Eds.). (2015). Methodological developments in data linkage. John Wiley & Sons. DOI:10.1002/9781119072454

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Ullah Khan, S. (2015). The rise of "big data" on cloud computing: Review and open research issues. Information Systems, 47, 98–115. https://doi.org/10.1016/j.is.2014.07.006

He, H., and Garcia, E. A. (2009) Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, 21, 1263-1284.http://dx.doi.org/10.1109/TKDE.2008.239

Heath, M. T. (c. 2013-2015). Iterative methods for linear systems [Lecture notes]. Department of Computer Science, University of Illinois. https://courses.grainger.illinois.edu/cs554/fa2015/notes/10_iterative.pdf

Hernández, M., Epelde, G., Alberdi, A., Cilla, R., & Rankin, D. (2022). Synthetic data generation for tabular health records: A systematic review. Neurocomputing, 493, 28–45. https://doi.org/10.1016/j.neucom.2022.04.053

Herschel, M., Diestelkämper, R. & Ben Lahmar, H. A survey on provenance: What for? What form? What from? The VLDB Journal 26, 881–906 (2017). https://doi.org/10.1007/s00778-017-0486-1

Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). Data Quality and Record Linkage Techniques, New York: Springer Science+Business Media. 234 pp. US$44.95. ISBN: 978-0-387-69502-0. DOI: 10.1007/s11336-007-9040-1

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S., (2017) Gans trained by a two-time-scale update rule converge to a local Nash equilibrium. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, Advances in Neural Information

Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 6626–6637, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Hilt, D. E., & Seegrist, D. W. (1977). Ridge, a computer program for calculating ridge regression estimates (General Technical Report NE-28). U. S. Department of Agriculture, Forest Service, Northeastern Forest Experiment Station. https://doi.org/10.5962/bhl.title.68934

Ho, J., Stefano, E., (2016). "Generative Adversarial Imitation Learning". Advances in Neural Information Processing Systems. 29: 4565–4573. arXiv:1606.03476

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Hoyer, W. D., Chandy, R., Dorotic, M., Krafft, M., & Singh, S. S. (2010). Consumer Cocreation in New Product Development. Journal of Service Research, 13(3), 283-296. https://doi.org/10.1177/1094670510375604

Infostat, Tertiary education graduates by educational-qualification degree, sex and narrow field of education (foet), 2001 – 2016, Report result

Infostat.bg, Home, Demographic and social statistics, Population, Dec. 2023, infostat.nsi.bg/infostat/pages/module.jsf?x_2=80

International Business Machines LLC. (2019), IBM SPSS Modeler 18.1.1 Python Scripting and Automation Guide, Simgennode properties pp. 93

Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2014, July). Big Data and Its Technical Challenges. Communications of the ACM, 57. https://doi.org/10.1145/2611567

Jaworski, P., Durante, F., Härdle, W., K., Rychlik, T., 2010, Copula Theory and Its Applications, Lecture Notes in Statistics Springer, ISSN 2197-7186.

Jordan, M.I. and Mitchell, T. M. (2015) Machine Learning: Trends, Perspectives, and Prospects. Science (New York, N. Y.), 349, 255-260. https://doi.org/10.1126/science.aaa8415

Jordon, J., Yoon, J., & van der Schaar, M. (2018). PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees. International Conference on Learning Representations (ICLR).

Julier, S., Uhlmann, J., 1996, A general method for approximating nonlinear transformation of probability distributions

Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J. and Melançon, G. (2008) Visual Analytics: Definition, Process, and Challenges. In: Kerren, A., Stasko, J. T., Fekete, J.-D. and North, C., Eds., Information Visualization, Springer, Berlin, 154-175. https://doi.org/10.1007/978-3-540-70956-5_7

Kingma, D., Ba, J., L. (2015). Adam: A method for stochastic optimization, In ICLR, 2015

Kitchin, R. (2014). The data revolution: Big data, open data, data infrastructures & their consequences. SAGE Publications Ltd, https://doi.org/10.4135/9781473909472

Ko, D. H., Hassan, A. U., Suk, J., & Choi, J. (2020). Korean font synthesis with GANs. International Journal of Computer Theory and Engineering, 12(4), 92–96.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2 (pp. 1137–1143). Morgan Kaufmann Publishers Inc.

Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione [On the empirical determination of a distribution law]. Giornale dell'Istituto Italiano degli Attuari, 4, 83–91

Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: A review of classification and combining techniques. Artificial Intelligence Review, 26(3), 159–190. doi.org

Kurowicka, D., van Horssen, W., T., 2015, On an interaction function for copulas, Journal of Multivariate Analysis Volume 138, June 2015, Pages 127-142, https://doi.org/10.1016/j.jmva.2014.12.012.

Leamer, E. E. (1973). Multicollinearity: A Bayesian interpretation. The Review of Economics and Statistics, 55(3), 371–380. https://doi.org/10.2307/1927962

Legard, R., Keegan, J., & Ward, K. (2003). In-depth interviews. In J. Ritchie & J. Lewis (Eds.), Qualitative research practice: A guide for social science students and researchers (pp. 138–169). Sage Publications.

Lemon, K. N. and Verhoef, P. C. (2016) Understanding Customer Experience throughout the Customer Journey. Journal of Marketing, 80, 69-96. https://doi.org/10.1509/jm.15.0420

Lenzerini, M. (2002). Data integration: A theoretical framework. Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, 233–246.

Liebscher, E., 2008, Construction of asymmetric multivariate copulas, Journal of Multivariate Analysis Volume 99, Issue 10, November 2008, Pages 2234-2250, https://doi.org/10.1016/j.jmva.2008.02.025.

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. Journal of the American Statistical Association, 62(318), 399–402.

Little, R. J., Rubin, D. B. (2019) Statistical Analysis with Missing Data. Vol. 793, John Wiley & Sons, Hoboken. https://doi.org/10.1002/9781119482260

Little, R. J. A. (1993) Statistical Analysis of Masked Data. Journal of Official Statistics 9, 407–426

Liu, R., Wei, J., Liu, F., Si, C., Zhang, Y., Rao, J., Zheng, S., Peng, D., Yang, D., Zhou, D., and Dai., A., M., (2024). Best Practices and Lessons Learned on Synthetic Data for Language Models

Liu-Thompkins, Y., & Malthouse, E. C. (2017). A primer on using behavioral data for testing theories in advertising research. Journal of Advertising, 46(1), 213–225.

Lourme, A., Maurer, F., 2017, Testing the Gaussian and Student's t copulas in a risk management framework, Economic Modelling Volume 67, Pages 203-214, https://doi.org/10.1016/j.econmod.2016.12.014.

Löwdin, P. (1970). On the nonorthogonality problem. Advances in Quantum Chemistry, 5, 185–199.

Machler, M., (2011), Numerically Stable Frank Copula Functions via Multiprecision: R Package Rmpfr, DOI: 10.13140/RG.2.2.27067.92968

Mai, J., Scherer M., (2012), "Simulating Copulas (Stochastic Models Sampling Algorithms and Applications)". (“Copula (probability theory) - Infogalactic: the planetary knowledge core”) World Scientific. ISBN 978-1-84816-874-9.

Manner, H., (2007), Estimation and Model Selection of Copulas with an Application to Exchange Rates, RM/07/056.

Manyika, J., et al. (2011) Big Data: The Next Frontier for Innovation, Competition, and Productivity. San Francisco, McKinsey Global Institute, CA, USA.

Massey, F., J., Jr., (1951), The Kolmogorov-Smirnov Test for Goodness of Fit, Journal of the American Statistical Association. Vol. 46, 68-78, No. 253, https://doi.org/10.2307/2280095

MathWorks, Inc. (2024). MATLAB (Version R2022b) [Computer software]. https://www.mathworks.com

McKechnie, L. E. F. (2008). Observational Research. In L. M. Given (Ed.), The SAGE Encyclopedia of Qualitative Research Methods (pp. 573-577). Sage.

Ministry of Agriculture and Food, (2024), Statistics and Analyses, https://www.mzh.government.bg/en/statistics-and-analyses/

Ministry of Finance, (2024), Economic Policy, Analyses and Research, https://www.minfin.bg/en/865

Miyato, T., Koyama, M. (2018). cGANs with projection discriminator. In ICLR, 2018

Motiwalla, L., Deokar, A. V., Sarnikar, S., & Tao, L. (2019). Leveraging data analytics for behavioral research. Information Systems Frontiers, 21, 735–742. https://doi.org/10.1007/s10796-019-09928-8

Mukhopadhyay, S., Parzen, E., 2019, Nonparametric Universal Copula Modeling https://doi.org/10.48550/arXiv.1912.05503.

Nelsen, R. B. (2006). An Introduction to Copulas (Second ed.). New York: Springer. ISBN 978-1-4419-2109-3

Newcombe, H. B., Kennedy, J. M., Axford, S. J., & James, A. P. (1959). Automatic linkage of vital records. Science, 130(3381), 954-959. https://doi.org/10.1126/science.130.3381.954

Nikhil, (2024). Google AI Introduces CodecLM: A Machine Learning Framework for Generating High-Quality Synthetic Data for LLM Alignment], [How to Use Synthetic and Simulated Data Effectively, https://towardsdatascience.com/how-to-use-synthetic-and-simulated-data-effectively-04d8582b6f88

NSI, (2021), Census 2021, Sofia, https://census2021.bg

NSI, (2023), Home, Business statistics R&D, Innovations and Information Society Information Society, Individuals Using the Internet by Purposes.

NSI, Home, Demographic and social statistics

O'Neil, C., (2016) Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York: Crown Publishers, 2016. 272p. Hardcover, ISBN 978-0553418811

Papazoglou, M. P. (2003). Service-oriented computing: Concepts, characteristics, and directions. In Fourth International Conference on Web Information Systems Engineering pp. 3-12. IEEE

Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The Synthetic Data Vault. In 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA) (pp. 399-410). IEEE. https://doi.org/10.1109/DSAA.2016.49

Pearson, K. (1936). Method of moments and method of maximum likelihood. Biometrika, 28(1/2), 34–59.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. Annual Review of Psychology.

Porozhanov, R. (2018), Broy na sobstvenitsite na zemedelski zemi i pritezhavanite ot tyah ploshti, agri.bg, 25.06.2018,

Pradhan, A., Ortan, A., Verma, A., & Seshadri, M. (2024). LLM-as-a-judge: Rapid evaluation of legal document recommendation for retrieval-augmented generation. arXiv. https://doi.org/10.48550/arXiv.2509.12382

Qu, W., Liu, H., & Zhang, Z. (2020). A method of generating multivariate non-normal random numbers with desired multivariate skewness and kurtosis. Behavior Research Methods, 52(3), 939–946. https://doi.org/10.3758/s13428-019-01291-5

Raab, G. M., Nowok, B., & Dibben, C. (2016). synthpop: Bespoke Creation of Synthetic Data in R. Journal of Statistical Software, 74(11), 1-26.

Radford, A., Metz, L., Chintala, S. (2015). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv preprint arXiv:1511.06434.

Raghunathan, T. E., Reiter, J. P., & Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. Journal of Official Statistics, 19(1), 1-16.

Rahm, E. and Do, H. H. (2003) Data Cleaning Problems and Current Approaches. IEEE Bulletin on Data Engineering, 7, 182-185.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. Survey Methodology, 29(2), 181–188.

Reiter, J. P. (2005). Significance tests for multi-component estimands from multiply imputed, synthetic microdata. Journal of Statistical Planning and Inference, Volume 131, Issue 2, 1 May 2005, Pages 365-377

Restrepo, J., P., Rivera, J., C., Laniado, H., Osorio, P., Becerra, O., (2023), Nonparametric Generation of Synthetic Data Using Copulas, Electronics, 12(7), 1601; https://doi.org/10.3390/electronics12071601

Ridder, G., & Moffitt, R. (2007). Chapter 75 The Econometrics of Data Combination. In J. J. Heckman & E. E. Leamer (Eds.), Handbook of Econometrics (Vol. 6B, pp. 5469–5547). Elsevier, DOI: 10.1016/S1573-4412(07)06075-8

Rosemann, M., vom Brocke, J. (2010). The Six Core Elements of Business Process Management. In: Brocke, J. v., Rosemann, M. (eds) Handbook on Business Process Management 1. International Handbooks on Information Systems. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-00416-2_5

Rubin, D.B. (1993) Discussion: Statistical Disclosure Limitation. Journal of Official Statistics, 9, 461-468.

Russom, P. (2011) Big Data Analytics. TDWI Best Practices Report, Fourth Quarter, 19, 1-34

Salganik, M. J. (2017). Bit by Bit: Social Research in the Digital Age. Princeton University Press.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., (2016). Improved Techniques for Training GANs

Santangelo, G., Nicora, G., Bellazzi, R., & Kalra, D. (2025). How good is your synthetic data? SynthRO, a dashboard to evaluate and benchmark synthetic tabular data. BMC Medical Informatics and Decision Making, 25, Article 89. https://doi.org/10.1186/s12911-024-02731-9

Saris, W. E., & Gallhofer, I. N. (2014). Design, Evaluation, and Analysis of Questionnaires for Survey Research. John Wiley & Sons.

Sarmin, F. J., et al. (2025). Synthetic data: Revisiting the privacy-utility trade-off. International Journal of Information Security, 24(4), Article 156. https://doi.org/10.1007/s10207-025-01072-6

Schmidt, T., 2006, Coping with Copulas, Copulas-From theory to application in finance. ISBN 978-1-9043-3945-8.

Seufert, E. B. (2014). Quantitative methods for product management. In Freemium economics: Leveraging analytics and user segmentation to drive revenue (pp. 47–82). Elsevier. https://doi.org/10.1016/B978-0-12-416690-5.00003-8

Shneiderman, B. (2003), The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations, The Craft of Information Visualization Readings and Reflections Interactive Technologies 2003, Pages 364-371, https://doi.org/10.1016/B978-155860915-0/50046-9

Simard, R., L'Ecuyer, P., 2010, "Computing the Two-Sided Kolmogorov-Smirnov Distribution", Journal of Statistical Software.

Simmhan, Y. L., Plale, B., & Gannon, D. (2005). A survey of data provenance in e-science. ACM SIGMOD Record, 34(3), 31-36. https://doi.org/10.1145/1084805.1084812

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. Journal of Business Research, 70, 263-286. https://doi.org/10.1016/j.jbusres.2016.08.001

Skabar, A., 2024. Evaluating Synthetic Data — The Million Dollar Question, https://towardsdatascience.com/

Sklar, A., 1996, Random variables distribution functions and copulas – a personal look backward and forward, ISBN 978-0-940600-40-9, DOI: 10.1214/lnms/1215452606.

Smirnov, N. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. Bulletin Mathématique de l'Université de Moscou, 2(2), 3–14.

Smirnov, N. (1948) Table for Estimating the Goodness of Fit of Empirical Distributions. The Annals of Mathematical Statistics, 19, 279-281. https://doi.org/10.1214/aoms/1177730256

Snoke, J., Raab, G. M., Nowok, B., Dibben, C., & Slavkovic, A. (2018). General and specific utility measures for synthetic data. Journal of the Royal Statistical Society: Series A (Statistics in Society), 181(3), 663-688.

Spiekermann, S. (2012). The challenges of privacy by design. Communications of the ACM, 55(7), 33–35. https://cacm.acm.org/opinion/the-challenges-of-privacy-by-design/

Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. Journal of the American Statistical Association, 69(347), 730–737

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society: Series B (Methodological), 36(2), 111–133. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x

Strelen, J. C. (2009). Tools for dependent simulation input with copulas. Proceedings of the 2nd International ICST Conference on Simulation Tools and Techniques. ICST. https://doi.org/10.4108/ICST.SIMUTOOLS2009.5630

Tipatov, N., 2009, "Statistika po kommentariyam k testu Ayzenka", Biznes Trener, http://b-t.com.ua/test_ayzenk_opis.html (available in Russian language)

Trading Economics, (2022), Home Ownership Rate in Bulgaria, https://tradingeconomics.com/bulgaria/home-ownership-rate

Van Breugel, B., Qian, Z. and Van Der Schaar, M., (2023). Synthetic data, real errors: how (not) to publish and use synthetic data. In International Conference on Machine Learning, pages 34793–34808. PMLR.

Van der Aalst, W. M., Adriansyah, A., de Medeiros, A. K. A., Arcieri, F., Baier, T., Blickle, T., & Wynn, M. T. (2011). Process mining manifesto. In Business Process Management Workshops (pp. 169-194). Springer.

Van Der Aalst, W. M. P. (2013). Business Process Management: A Comprehensive Survey. ISRN Software Engineering, 2013, 1-37

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9(11), 2579-2605.

Vargo, S. L., Lusch, R. F. (2016). Institutions and axioms: an extension and update of service-dominant logic. J. of the Acad. Mark. Sci. 44, 5–23. https://doi.org/10.1007/s11747-015-0456-3

Vassiliadis, P. (2009) A survey of Extract-transform-Load technology. International Journal of Data Warehousing and Mining, 5, 1-27. https://doi.org/10.4018/jdwm.2009070101

Vatsalan, D., Christen, P., & Verykios, V. S. (2013). A taxonomy of privacy-preserving record linkage techniques. Information Systems, 38(6), 846-869. https://doi.org/10.1016/j.is.2012.11.005

Venugopal, R., Shafqat, N., Venugopal, I., Tillbury, B., Stafford, H., Bourazeri, A. (2022). Privacy preserving Generative Adversarial Networks to model Electronic Health Records, PMID: 35779443 DOI: 10.1016/j.neunet.2022.06.022

Waller, M. A., & Fawcett, S. E. (2013). Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management. Journal of Business Logistics, 34(2), 77-84. https://doi.org/10.1111/jbl.12010

Wan, K., Kornhauser, A., 2023, On the Properties of Gaussian Copula Mixture Models, https://doi.org/10.48550/arXiv.2305.01479

Wang, Y., Kung, L., and Byrd, T. A. (2018) Big Data Analytics: Understanding Its Capabilities and Potential Benefits for Healthcare Organizations. Technological Forecasting and Social Change, 126, 3-13. https://doi.org/10.1016/j.techfore.2015.12.019

Wilson, A. G., Z. Ghahramani (2010). Copula Processes, arXiv:1006.1350

Winkler, W. E. (2006). Overview of record linkage and current research directions. Bureau of the Census, 25(4), 603-623.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems, 58(2), 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1

Wood, E., Baltrusaitis, T., Hewitt, C., Dziadzio, S., Cashman, T. J., and Shotton, J., (2021). Fake it till you make it: face analysis in the wild using synthetic data alone. In 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, pages 3661–3671. IEEE, 2021. DOI: 10.1109/ICCV48922.2021.00366. URL https://doi.org/10.1109/ICCV48922.2021.00366.

Wu, X., Zhu, X., Wu, G.-Q. and Ding, W. (2014) Data Mining with Big Data. IEEE Transactions on Knowledge and Data Engineering, 26, 97-107. https://doi.org/10.1109/TKDE.2013.109

Xie, L., Lin, K., Wang, S., Wang, F., & Zhou, J. (2018). Differentially Private Generative Adversarial Networks for Time Series, Continuous, and Discrete Data. arXiv:1802.07831.

Yan, C, Yan, Y., Wan, Z., Zhang, Z., Omberg, L., Guinney, J., Mooney, S. D., & Malin, B. A. (2022). A multifaceted benchmarking of synthetic electronic health record generation models. Nature Communications, 13, Article 7609. https://doi.org/10.1038/s41467-022-35295-1

YiJaehyun, D., Ji, A. (2020). An Effective Optimization Method for Machine Learning Based on ADAM, Applied Sciences 10(3):1073, DOI: 10.3390/app10031073

Yoon, J., Drumright, L., van der Schaar, M. (2020). Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN), PMID: 32167919 DOI: 10.1109/JBHI.2020.2980262

Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., Franklin, M. J., Ghodsi, A., Gonzalez, J., Shenker, S., & Stoica, I. (2016). Apache Spark: A unified engine for big data processing. Communications of the ACM, 59(11), 56–65. https://doi.org/10.1145/2934664

Zamzmi, G., Subbaswamy, A., Sizikova, E., Margerrison, E., Delfino, J. G., & Badano, A. (2025). Scorecard for synthetic medical data evaluation. Communications Engineering, 4, Article 130. https://doi.org/10.1038/s44172-025-00450-1

Zarsky, Tal, (2017), Incompatible: The GDPR in the Age of Big Data (August 8, 2017). Seton Hall Law Review, Vol. 47, No. 4(2), 2017, Available at SSRN: https://ssrn.com/abstract=3022646

Zhang, H., Goodfellow, I., Metaxas, D., Odena, A. (2019). "Self-Attention Generative Adversarial Networks". International Conference on Machine Learning. PMLR: 7354–7363.

Zheng, L., Chiang, W.-L., Sheng, Y., et al. (2023). Judging LLM-as-a-judge with MT-bench and chatbot arena. Advances in Neural Information Processing

Systems 36 (NeurIPS 2023, Datasets and Benchmarks track). https://doi.org/10.48550/arXiv.2306.05685

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Закон за кредитните институции, В сила от 01.01.2007 г. Обн. ДВ. бр. 59 от 21 Юли 2006г., изм. ДВ. бр. 105 от 22 Декември 2006г., изм. ДВ. бр. 52 от 29 Юни 2007г., изм. ДВ. бр. 59 от 20 Юли 2007г., изм. ДВ. бр.109 от 20 Декември 2007г., изм. ДВ. бр. 69 от 5 Август 2008г., изм. ДВ. бр. 23 от 27 Март 2009г., изм. ДВ. бр. 24 от 31 Март 2009г

Авторски публикации

Marchev JR, A., Marchev, V., Haralampiev, K., Efremov, A., Markov, B., Lyubchev, D., Piryankova, M., Filipov, B., Masarliev, D., & Mitkov, V. (2025), Synthesized Economic Agents Dataset, HuggingFace.co, DOI address https://doi.org/10.57967/hf/3701

Marchev JR, A., Marchev, V., Haralampiev, K., Efremov, A., Markov, B., Lyubchev, D., Piryankova, M., Filipov, B., Masarliev, D., & Mitkov, V. (2026), Software tools for synthetic-data generation, GitHub.com, https://github.com/Marchev-Science/synthetic-data.git

Marchev, A. A., Jr., & Marchev, V. S. (2026). Methodological findings from generating multidimensional synthetic datasets. ARPHA Proceedings, 10, 301–314. https://doi.org/10.3897/ap.10.e0301

Marchev, A. Marchev, V. (2024). - Methodological Considerations for Anonymizing Tabular Data Using Generative Adversarial Networks, ICACS '24: Proceedings of the 8th International Conference on Algorithms, Computing and Systems

Marchev, A., Jr, Marchev, V., (2023). Automated Algorithm for Multi-variate Data Synthesis with Cholesky Decomposition, ICACS '23: Proceedings of the 7th International Conference on Algorithms, Computing and Systems, https://doi.org/10.1145/3631908.3631909

Marchev, A., Jr., & Marchev, V. (2022). Synthesizing multi-dimensional personal data sets. In Proceedings of the 47th International Conference “Applications of Mathematics in Engineering and Economics” (AIP Conference Proceedings, Vol. 2505, No. 1, 020012). AIP Publishing. https://doi.org/10.1063/5.0100615

Marchev, A., Jr., Marchev, V., & Efremov, A. (2025). Data expansion by business-logic scaling - an applied approach for data synthesis. In Proceedings of the

2025 ICARAI Conference (pp. 1–7). IEEE. https://ieeexplore.ieee.org/document/11137819/

Marchev, A., Lyubchev, D., Marchev, V., (2024). Inverse Copula Sampling for Multi-Dimensional Data Synthesis, 50th International Conference Applications of Mathematics in Engineering and Economics, Sozopol, June 07-13, 2024, AIP Conference proceedings (in publishing)

Marchev, A., Marchev, V., & Efremov, A. (2025). Methodology for cold-start modeling using synthetic data. AIP Conference Proceedings, 3182(1), 020002. https://doi.org/10.1063/5.0246285

Marchev, V, Marchev, A., 2021, Methods for Simulating Multi-dimensional Data for Financial Services Recommendation, Bulgarian Economic Papers, Center for economic theories and policies, ISSN: 2367-7082, BEP 02-2021, Feb. 2021, http://www.bep.bg

Marchev, V., & Marchev, A., Jr. (2019). Simulation of a multi-criteria database for banking services: Algorithm and business logic. In New information technologies and big data: Opportunities and perspectives in analyzes and management decisions in business, economy, and social sphere (pp. 179–190). UNSS Publishing House.

Marchev, V., Marchev JR, A., Haralampiev, K., Efremov, A., Markov, B., Lyubchev, D., Piryankova, M., Filipov, B., Masarliev, D., & Mitkov, V. (2025). Methodological Approaches for Multidimensional Personal Data Creation. Vanguard Scientific Instruments in Management, 20, 108-131.

Marchev, V., Marchev, A. (2025). Anonymizing Personal Information Using Distribution-Based Data Synthesis. In: El Khatib, M., Alzoubi, H. M., Angelova, Y. (eds) Sustainable Innovation for Engineering Management (SIEM). ISCME 2024. Advances in Science, Technology & Innovation. Springer, Cham. https://doi.org/10.1007/978-3-031-89889-1_38

Marchev, V., Marchev, A., Piryankova, M., Masarliev, D., Mitkov, V., (2023), Synthesizing an anonymized multidimensional dataset featuring financial, economic, demographic, and personal traits data, VSIM, vol. 19, no. 1, 2023, ISSN 1314-0582

VII. ПРИЛОЖЕНИЯ

Приложение 1 - Бизнес логиката на модела

№	Фактор	Комбиниращ фактор	Забележка
1	Възраст - навършени години	Степен на образование	Под 24 г. е по-малко вероятно да има висше образование
2	Възраст - навършени години	Семейно положение	Под 24 г. е по-малко вероятно да е женен/омъжена, вдовец/вдовица. Над 64г. съществува по-голяма вероятност да е вдовец/вдовица
3	Възраст - навършени години	Брой деца под 18 години	От 20-24 е по-малко вероятно да има повече от 1 дете под 18 години, над 65г. е малко вероятно да има деца под 18 г.
4	Възраст - навършени години	Социално-икономически статус	От 20-24 е по-малко вероятно да има повече от 1 дете под 18 години
5	Възраст - навършени години	Бр. апартамент/къща	Под 24 г. е по-малко вероятно да притежава собствен апартамент/къща
6	Възраст - навършени години	Парцели земя	Под 24 г. е по-малко вероятно да притежава собствен парцели земя
7	Възраст - навършени години	Предишен инвестиционен опит в години	Под 24 г. е по-малко вероятно да притежава инвестиционен опит. Между 35-44 и 45-54 е по-вероятно да притежават богат инвестиционен опит
8	Възраст - навършени години	Доход	Под 24 г. е по-малко вероятно да имат висок доход, след 65г. е по-малко вероятно да имат висок доход
9	Възраст - навършени години	Жилищни разходи	Под 24 г. е по-малко вероятно да имат високи жилищни разходи
10	Възраст - навършени години	Данъци и осигуровки	Под 24 г. е по-малко вероятно да имат високи разходи за данъци и осигуровки
11	Възраст - навършени години	Транспорт и съобщения	Под 24 г. е по-малко вероятно да имат високи разходи за транспорт и съобщения
12	Възраст - навършени години	Кредитна карта	Под 24 г. е по-малко вероятно да имат кредитна карта, над 65г. е по-малко вероятно да имат кредитна карта
13	Възраст - навършени години	Разплащателна сметка	Под 24 г. е по-малко вероятно да имат разплащателна сметка
14	Възраст - навършени години	Застраховка - имущество	Под 24 г. е по-малко вероятно да имат имуществена застраховка
15	Възраст - навършени години	Застраховка - живот	Под 24 г. е по-малко вероятно да имат застраховка живот
16	Възраст - навършени години	Застраховка – МПС (Каско)	Под 24 г. е по-малко вероятно да имат застраховка каско
17	Възраст - навършени години	Допълнително здравно осигуряване	Под 24 г. е по-малко вероятно да имат ДЗО
18	Възраст - навършени години	Овърдрафт	Под 24 г. е по-малко вероятно да имат Овърдрафт
19	Възраст - навършени години	Потребителски кредит	Под 24 г. е по-малко вероятно да имат потребителски кредит
20	Възраст - навършени години	Ипотечен кредит	Под 24 г. е по-малко вероятно да имат ипотечен кредит
21	Възраст - навършени години	Автомобилен лизинг/кредит	Под 24 г. е по-малко вероятно да имат лизинг
22	Възраст - навършени години	Допълнително пенсионно осигуряване - 3-ти стълб	Под 24 г. е по-малко вероятно да имат Допълнително пенс. осигуряване
23	Степен на образование	Статус на заетост	По-високата степен на образование, предполага статус на заетост - работодатели, постоянно заети в частни и обществени предприятия
24	Степен на образование	Брой деца под 18 години	По-високата степен на образование предполага по-малък брой деца. Макс 2 до 3

№	Фактор	Комбиниращ фактор	Забележка
25	Степен на образование	Социално-икономически статус	По-високата степен на образование предполага - статус икономически активен
26	Степен на образование	Доход	По-високата степен на образование предполага доходи в горния диапазон
27	Степен на образование	Данъци и осигуровки	По-високата степен на образование предполага доходи в горния диапазон, съответно по-високи данъци и осигуровки
28	Статус на заетост	Професия - Индустрия	Работещите в администрацията са - постоянно заети лица в обществения сектор
29	Статус на заетост	Професионален статус	Постоянно наетите лица са на трудов или на договор за управление
30	Статус на заетост	Доход	Постоянно наетите и работодателите е по-вероятно да имат по-висок доход
31	Статус на заетост	Данъци и осигуровки	Постоянно наетите и работодателите е по-вероятно да имат по-висок доход => да платят повече данъци и осигуровки
32	Статус на заетост	Среден брой банкови операции	Постоянно наетите и работодателите е по-вероятно да имат по-висок доход => да извършат по-голям брой банкови операции
33	Статус на заетост	Застраховка - имущество	Постоянно наетите и работодателите е по-вероятно да имат по-висок доход => биха платили за застраховка на имуществото
34	Статус на заетост	Застраховка - живот	Постоянно наетите и работодателите е по-вероятно да имат по-висок доход => биха платили за застраховка живот
35	Статус на заетост	Овърдрафт	Постоянно наетите и работодателите е по вероятно да имат овърдрафт
36	Статус на заетост	Потребителски кредит	Постоянно наетите и работодателите е по вероятно да имат потребителски кредит
37	Статус на заетост	Ипотечен кредит	Постоянно наетите и работодателите е по вероятно да имат ипотечен
38	Статус на заетост	Автомобилен лизинг/кредит	Постоянно наетите и работодателите е по вероятно да имат автомобилен лизинг
39	Статус на заетост	Допълнително пенсионно осигуряване - 3-ти стълб	Постоянно наетите и работодателите е по вероятно да имат ДПО
40	Семейно положение	Брой членове на домакинството	Броят на членовете на домакинствата при семейните е по-вероятно да е по-голям от 2
41	Семейно положение	Брой деца под 18 години	По-вероятно е семейните да имат поне едно дете под 18г.
42	Семейно положение	Социално-икономически статус	По-вероятно е семейните да са икономически активни
43	Семейно положение	Доход	По-вероятно е семейните да имат по-висок доход
44	Семейно положение	Лични разходи	По-вероятно е семейните да имат по-високи лични разходи
45	Семейно положение	Жилищни разходи	По-вероятно е семейните да имат по-високи жилищни разходи
46	Семейно положение	Транспорт и съобщения	По-вероятно е семейните да имат по-високи разходи за транспорт и съобщения
47	Семейно положение	Свободно време и хоби	По-вероятно е семейните да имат по-ниски разходи за свободно време и хоби
48	Брой членове на домакинството	Брой деца под 18 години	Броят членове на домакинството е пряко зависим от броя деца под 18г. възраст
49	Брой членове на домакинството	Жилищни разходи	По-вероятно е да има по-високи жилищни разходи при домакинства по-големи от 3 чл.
50	Брой членове на домакинството	Лични разходи	По-вероятно е да има по-високи лични разходи при домакинства по-големи от 3 чл.

№	Фактор	Комбиниращ фактор	Забележка
51	Брой членове на домакинството	Разходи за транспорт и съобщения	По-вероятно е да има по-високи разходи за транспорт при домакинства по-големи от 3 чл.
52	Брой членове на домакинството	Разходи за свободно време и хоби	По-вероятно е да има по-ниски разходи за транспорт и свободно време при домакинства по-големи от 3 чл.
53	Брой деца под 18 години	Лични разходи	При повече от 1 дете под 18 г. е по-вероятно личните разходи да се повишат
54	Брой деца под 18 години	Жилищни разходи	При повече от 1 дете под 18 г. е по-вероятно жилищните разходи да се повишат
55	Брой деца под 18 години	Транспорт и съобщения	При повече от 1 дете под 18 г. е по-вероятно разходите за транспорт да се повишат
56	Социално-икономически статус	Професионален статус	По-вероятно е икономически неактивните да са безработни или пенсионери
57	Социално-икономически статус	Бр. апартамент/къща	По-вероятно е икономически активните да имат поне 1 апартамент/къща
58	Социално-икономически статус	Парцели земя	По-вероятно е икономически активните да притежават парцели
59	Социално-икономически статус	Автомобили на домакинство	По-вероятно е икономически активните да имат поне 1 автомобил
60	Социално-икономически статус	Доход	По-вероятно е икономически активните да имат по-високи доходи
61	Социално-икономически статус	Лични разходи	По-вероятно е икономически активните да имат по-високи лични разходи
62	Социално-икономически статус	Жилищни разходи	По-вероятно е икономически активните да имат по-високи жилищни разходи
63	Социално-икономически статус	Данъци и осигуровки	По-вероятно е икономически активните да имат по-високи данъци и осигуровки
64	Социално-икономически статус	Транспорт и съобщения	По-вероятно е икономически активните да имат по-високи разходи за транспорт и съобщения
65	Социално-икономически статус	Разходи за свободно време и хоби	По-вероятно е икономически активните да имат по-високи разходи за свободно време и ходи
66	Социално-икономически статус	Овърдрафт	По-вероятно е икономически активните да имат овърдрафт
67	Социално-икономически статус	Потребителски кредит	По-вероятно е икономически активните да имат потребителски кредит
68	Социално-икономически статус	Ипотечен кредит	По-вероятно е икономически активните да имат ипотечен кредит
69	Социално-икономически статус	Автомобилен лизинг/кредит	По-вероятно е икономически активните да имат автомобилен лизинг
70	Професия - Индустрия	Бр. апартамент/къща	Финансовата сфера, Търговията, добивната промишленост, телекомуникациите - предполагат по-висок доход => по-вероятно е да притежават повече от 1 апартамент/къща
71	Професия - Индустрия	Автомобили на домакинство	Финансовата сфера, Търговията, добивната промишленост, телекомуникациите - предполагат по-висок доход => по-вероятно е да притежават повече от 1 автомобил
72	Професия - Индустрия	Степен на образование	Финансовата сфера, Търговията, добивната промишленост, телекомуникациите, администрацията - предполагат по-висок доход => по-вероятно е да притежават по-висока степен на образование
73	Професия - Индустрия	Предишен инвестиционен опит в години	Финансовата сфера по-вероятно е да притежават предишен инвестиционен опит
74	Професия - Индустрия	Доход	Финансовата сфера, Търговията, добивната промишленост, телекомуникациите - предполагат по-висок доход

№	Фактор	Комбиниращ фактор	Забележка
75	Професия - Индустрия	Данъци и осигуровки	Финансовата сфера, Търговията, добивната промишленост, телекомуникациите - предполагат по-висок доход => по-високи данъци и осигуровки
76	Професия - Индустрия	Транспорт и съобщения	Финансовата сфера, Търговията, добивната промишленост, телекомуникациите - предполагат по-висок доход => по-високи разходи за транспорт
77	Професия - Индустрия	Допълнително пенсионно осигуряване - 3-ти стълб	Финансовата сфера, Търговията, добивната промишленост, телекомуникациите - предполагат по-висок доход => по-вероятно е да инвестират в ДПО
78	Професионален статус	Доход	трудовия договор, договорът за управление, самонаетото лице - предполагат по-високи доходи
79	Професионален статус	Данъци и осигуровки	трудовия договор, договорът за управление, самонаетото лице - предполагат по-високи данъци и осигуровки
80	Професионален статус	Допълнително пенсионно осигуряване - 3-ти стълб	трудовия договор, договорът за управление, самонаетото лице - предполагат по-голяма вероятност за ДПО
81	Бр. апартамент/къща	Жилището, в което живея е:	Наличието на поне 1 жилище, предполага, че обитаваното жилище е собствено
82	Бр. апартамент/къща	Данъци и осигуровки	Наличието на повече от 1 жилище предполага по-високи разходи за данъци
83	Бр. апартамент/къща	Жилищни разходи	Наличието на повече от 1 жилище предполага по-високи жилищни разходи
84	Бр. апартамент/къща	Ипотечен кредит	Наличието на повече от 1 жилище - има по-голяма вероятност за наличие на ипотечен кредит
85	Бр. апартамент/къща	Застраховка - имущество	Наличието на повече от 1 жилище предполага по-високи разходи за данъци
86	Автомобили на домакинство	Доход	По-големият броя автомобили предполага по-високи доходи
87	Автомобили на домакинство	Разходи за транспорт и съобщения	По-големият броя автомобили предполага по-високи разходи за транспорт
88	Автомобили на домакинство	Автомобилен лизинг/кредит	По-големият броя автомобили предполага по-високи разходи за лизинг
89	Автомобили на домакинство	Застраховка – МПС (Каско)	По-големият броя автомобили предполага по-високи разходи за застраховка каско
90	Жилището, в което живея е:	Доход	Собственото жилище предполага по-висок доход
91	Жилището, в което живея е:	Жилищни разходи	Собственото жилище предполага по-високи жилищни разходи
92	Жилището, в което живея е:	Данъци и осигуровки	Собственото жилище предполага по-високи данъци
93	Жилището, в което живея е:	Ипотечен кредит	Собственото жилище предполага закупуване чрез ипотечен кредит
94	Образование	Доход	Образование- средно и по-високо, предполага по-висок доход
95	Доход	Жилищни разходи	При по-високи доходи се повишават разходите за жилище
96	Доход	Данъци и осигуровки	При по-високи доходи се повишават разходите за данъци и осигуровки
97	Доход	Разходи за транспорт и съобщения	При по-високи доходи се повишават разходите за транспорт
98	Доход	Разходи за свободно време и хоби	При по-високи доходи се повишават разходите за свободно време и ходи
99	Доход	Застраховка - имущество	При по-високи доходи е по-вероятно да се направи разход за имуществена застраховка
100	Доход	Застраховка - живот	При по-високи доходи е по-вероятно да се направи

№	Фактор	Комбиниращ фактор	Забележка
			разход за застраховка живот
101	Доход	Застраховка – МПС (Каско)	При по-високи доходи е по-вероятно да се направи разход за и застраховка каско
102	Доход	Допълнително здравно осигуряване	При по-високи доходи е по-вероятно да се направи разход за ДЗО
103	Доход	Потребителски кредит	При по-високи доходи е по-вероятно да има потребителски кредит
104	Доход	Ипотечен кредит	При по-високи доходи е по-вероятно да има ипотечен кредит
105	Жилищни разходи	Застраховка - имущество	Повишава жилищните разходи
106	Жилищни разходи	Ипотечен кредит	Повишава жилищните разходи
107	Жилищни разходи	Застраховка - имущество	Повишава жилищните разходи
108	Данъци и осигуровки	Допълнително здравно осигуряване	ДЗО - повишаване на разходите за данъци и осигуровки
109	Данъци и осигуровки	Допълнително пенсионно осигуряване - 3-ти стълб	ДПО - повишаване на разходите за данъци и осигуровки
110	Разходи за транспорт и съобщения	Застраховка – МПС (Каско)	Застраховката каско повишава косвените разходи за транспорт
111	Разходи за транспорт и съобщения	Автомобилен лизинг/кредит	Наличието на лизинг предполага по-високи разходи за транспорт
112	Дебитна карта	Разплащателна сметка	Наличието на дебитна карта изисква разплащателна сметка
113	Разплащателна сметка	Овърдрафт	Наличието на овърдрафт изисква разплащателна сметка
114	Разплащателна сметка	Потребителски кредит	Наличието на потребителски кредит изисква разплащателна сметка
115	Разплащателна сметка	Ипотечен кредит	Наличието на ипотечен кредит изисква разплащателна сметка
116	Разплащателна сметка	Автомобилен лизинг/кредит	Наличието на лизинг предполага разплащателна сметка
117	Застраховка – МПС (Каско)	Автомобилен лизинг/кредит	Наличието на лизинг предполага застраховка каско
118	Инвестиционен опит	Предишен инвестиционен опит в години	По-вероятно е да притежават банкова сметка
119	Инвестиционен опит	Инвестиция в акции	Предишен инвестиционен опит в години
120	Инвестиционен опит	Инвестиция в облигации	Предишен инвестиционен опит в години
121	Инвестиционен опит	Други инвестиции	Предишен инвестиционен опит в години
122	Инвестиционен опит	Инвестиция във фонд	Предишен инвестиционен опит в години
123	Инвестиционен опит	Кешови инвестиции	Предишен инвестиционен опит в години
124	Инвестиционен опит	Инвестиции в крипто валути	Предишен инвестиционен опит в години
125	Инвестиционен опит	Инвестиция в ДЦК	Предишен инвестиционен опит в години
126	Инвестиционен опит	Инвестиции в депозити	Предишен инвестиционен опит в години

Приложение 2 - Списък с описани променливи

Име на функция	Описание	Метод на изчисление	Метод на изчисление. Източници на данни	Мерна единица	Уместност	Грешка при извадката	Грешка, свързана с липса на семплиране	Географско	Временна дезагрегация	Сравним ост	Региони за
Пол/Друг	Показва полова идентичност	Синтезирана променлива*	Разпределението е получено от преброяването на населението в България през 2021 г.	**Номинална:** М/Ж	Целта е да се установят възможни корелации между пол и други индивидуални характеристики.	Официални данни – Преброяване 2021 г.	Потенциалните грешки включват неправилно записване	Данни за България	Тези данни са статични и отразяват пола към момента на преброяването през 2021 г.	Не е приложимо	Не е приложимо
Възраст	Показва брой навършени години.	Синтезирана променлива	Разпределението е извлечено от последното национално преброяване на населението в България, проведено през 2021 г.	**Непрекъсната:** 20 - 85	Възрастта е основен демографски фактор, който е от съществено значение за анализа на различни социални динамики.	Официални данни – Преброяване 2021 г.	Потенциалните грешки включват неправилно записване или неправилна дата на раждане в административните записи.	Данни за България	Данните са статични и отразяват възрастта на населението към датата на преброяването (2021 г.)	Не е приложимо	Не е приложимо

Име на функция	Описание	Метод на изчисление	Метод на изчисление. Източници на данни	Мерна единица	Уместност	Грешка при извадката	Грешка, свързана с липса на семплиране	Географско	Временна дезагрегация	Сравнимост	Регионално
Ниво на образование	Завършено ниво на образование.	Синтезирана променлива	Разпределението е получено от преброяването на населението в България през 2021 г.	**Ординална:** -Незавършено начално образование -Начално образование -Основно образование -Висше образование -Университетска диплома	Нивото на образование е ключова демографска характеристика, използвана за анализ на индивидуалните и общностните резултати.	Официални данни – Преброяване 2021 г.	Потенциалните грешки могат да произтичат от неправилно самоотчитане или класификация по време на събирането на данни.	Данни за България	Тези данни са статични и отразяват нивата на образование, отчетени по време на преброяването през 2021 г.	Не е приложимо	Не е приложимо
Трудов статус	Показва текущото участие на дадено лице в работната сила.	Синтезирана променлива	Разпределението е получено от преброяването на населението в България през 2021 г. и проучването на работната сила.	**Номинално:** -Работодатели -Самостоятелно наети -Служители в частни предприятия -Служители в държавни предприятия -Неплатени семейни работници -Безработен	Трудовият статус е критична демографска характеристика, използвана за оценка на динамиката на пазара на труда.	Официални данни – Преброяване 2021 г.	Грешки могат да възникнат поради неправилна класификация или липса на отговор.	Данни за България	Тези данни са статични и отразяват статуса на заетост през референтния период на преброяването от 2021 г.	Не е приложимо	Не е приложимо

Име на функция	Описание	Метод на изчисление	Метод на изчисление. Източници на данни	Мерна единица	Уместност	Грешка при извадката	Грешка, свързана с липса на семплиране	Географско	Времева дезагрегация	Сравнимост	Регионално
Семейно положение	Оценява правния статус на физическото лице.	Синтезирана променлива	Разпределението е получено от преброяването на населението в България през 2021 г.	**Номинална:** -Неженен/а -Женен/Омъжена -Разведен/а -Вдовец/а	Демографски характеристики обясняващи състава на домакинствата и социалната динамика.	Официални данни – Преброяване 2021 г.	Потенциални грешки могат да възникнат поради неправилно отчитане.	Данни за България	Тези данни са статични и отразяват семейното положение към момента на преброяването през 2021 г.	Не е приложимо	Не е приложимо
Брой членове на домакинството	Показва общия брой на лицата, живеещи в домакинство.	Синтезирана променлива	Разпределението е извлечено от преброяването на населението в България през 2021 г.	**Интервална:** 1; 2; 3; 4; 5+	Използва се за анализ на условията на живот, тенденциите в размера на домакинствата и социално-икономическите прогнози	Официални данни – Преброяване 2021 г.	Потенциалните грешки включват неправилно отчитане на състава на домакинството.	Данни за България	Тези данни са статични и отразяват броя на членовете на домакинствата към момента на преброяването през 2021 г.	Не е приложимо	Не е приложимо
Брой деца под 18 години	Наблюдава общия брой лица под 18 години.	Синтезирана променлива	Разпределението е извлечено от преброяването на населението в България през 2021 г., като данните са събрани от въпросници за домакинствата.	**Интервална:** 1; 2; 3; 4+	Броят на децата под 18 години оценява коефициентите на зависимост, образователната структура и подпомага разбирането семейните структури.	Официални данни – Преброяване 2021 г.	Потенциалните грешки включват неправилно класифициране на възрастта или пропускане на членове на домакинството.	Данни за България	Тези данни са статични и отразяват броя на децата под 18 години към момента на преброяването през 2021 г.	Не е приложимо	Не е приложимо

Име на функция	Описание	Метод на изчисление	Метод на изчисление. Източници на данни	Мерна единица	Уместност	Грешка при извадката	Грешка, свързана с липса на семплиране	Географско	Временна дезагрегация	Сравнимост	Регионално
Националност	Показва законната или самоопределената национална принадлежност на дадено лице.	Синтезирана променлива	Разпределението е извлечено от преброяването на населението в България през 2021 г., като данните са събрани от въпросници за домакинствата.	**Номинална**** От преброяването през 2021 г.	Ключова демографска характеристика за анализ на разнообразието на населението, културния състав и интегритета на общността.	Официални данни – Преброяване 2021 г.	Потенциалните грешки включват нежелание за показване или грешки във въвеждането на данни.	Данни за България	Тези данни са статични и отразяват самоопределената националност към момента на преброяването през 2021 г.	Не е приложимо	Не е приложимо
Религия	Показва религиозната принадлежност, системата от вярвания на дадено лице или липса на такава.	Синтезирана променлива	Разпределението е получено от преброяването на населението в България през 2021 г.	**Номинална**** От преброяването през 2021 г.	Демографски фактор за разбиране на културното многообразие, социалната динамика и нейното влияние върху традициите и ангажираността на общността.	Официални данни – Преброяване 2021 г.	Потенциалните грешки включват нежелание за разкриване или грешки във въвеждането на данни.	Данни за България	Тези данни са статични и отразяват религиозната принадлежност по време на преброяването през 2021 г.	Не е приложимо	Не е приложимо

Име на функция	Описание	Метод на изчисление	Метод на изчисление. Източници на данни	Мерна единица	Уместност	Грешка при извадката	Грешка, свързана с липса на семплиране	Географско	Временна дезагрегация	Сравнимост	Регионално
Професия/Индустрия	Разглежда професията или индустрията, в която е заето дадено лице.	Синтезирана променлива	Разпределението е базирано на информация от Националния статистически институт.	**Номинална**:** От НСИ	Професията/ индустрията - важен демографски фактор за разбиране на тенденциите в заетостта, икономическата структура и разпределението на труда в различните сектори.	Официални данни – НСИ.	Грешки могат да възникнат, ако хората предоставят неточни отговори.	Данни за България	Тези данни са статични и отразяват статуса на професията/индустрията през референтния период на преброяването от 2021 г.	Не е приложимо	Не е приложимо
Професионален статус	Показва вида заетост на дадено лице, категоризирана въз основа на неговата роля на пазара на труда.	Синтезирана променлива	Разпределението е базирано на информация от Националния статистически институт.	**Номинална**:** От Инфостат [17]	Професионалният статус предоставя информация за социално-икономическото положение на индивида, ролята му на пазара на труда, тенденциите в заетостта и икономическите неравенства.	Официални данни – НСИ	Грешки могат да възникнат поради неправилно попълване на анкетни карти, грешки при въвеждането или класифицирането на данни.	Данни за България	Тези данни са статични и отразяват социално-икономическия статус през референтния период на преброяването от 2021 г.	Не е приложимо	Не е приложимо

Име на функция	Описание	Метод на изчисление	Метод на изчисление. Източници на данни	Мерна единица	Уместност	Грешка при извадката	Грешка, свързана с липса на семплиране	Географско	Времева дезагрегация	Сравнимост	Регионално
Брой апартаменти/къщи	Разглежда броя на жилищните единици (апартаменти или къщи), притежавани от домакинство.	Синтезирана променлива	Разпределението е базирано на преброяването на населението на България от 2021 г., административни и статистически отчети.	**Интервална:** 0; 1; 2+	Предоставя информация относно достъпа до жилища и условията. Помага за анализа на разпределението на жилищните ресурси и жизнения стандарт в различните социални групи.	Официални данни – Преброяване 2021 г.	Грешки могат да възникнат, ако хората предоставят неточни отговори.	Данни за България	Тези данни са статични и отразяват броя на апартаментите/ къщите през референтния период на преброяването от 2021 г.	Не е приложимо	Не е приложимо
Парцели земя	Измерител за собствеността на земеделска земя, от страна на физическо лице, като се изследва процентът на собственост.	Синтезирана променлива	Разпределението е извлечено от поземления регистър, данни от Министерството на земеделието и допускания.	**Двоична:** Да/Не	Предоставя информация относно достъпа до земя, собствеността и разпределението на селскостопанските ресурси в регионите и социалните групи.	Възможни са разминавания в предположенията	Грешки могат да възникнат поради пропуски при регистрация, неточности при въвеждане на данни или липсваща информация.	Данни за България	Тези данни са статични и отразяват броя на земеделските парцели през референтния период на преброяването през 2021 г.	Не е приложимо	Не е приложимо

Име на функция	Описание	Метод на изчисление	Метод на изчисление. Източници на данни	Мерна единица	Уместност	Грешка при извадката	Грешка, свързана с липса на семплиране	Географско	Времева дезагрегация	Сравнимост	Регионално
Брой автомобил	Разглежда броя на автомобилите, притежавани от едно домакинство.	Синтезирана променлива	Разпределението е извлечено от преброяването на населението в България през 2021 г.	**Интервална:** 0; 1; 2; 3+	Броят на автомобилите помага за анализа на мобилността и условията на живот. Също така разкрива социално-икономическите различия между домакинствата.	Официални данни – Преброяване 2021 г.	Грешката включва потенциални неточности в самоотчитането, неразбиране на въпроси или липсващи данни.	Данни за България	Тези данни са статични и отразяват броя на автомобилите в домакинствата през референтния период на преброяването от 2021 г.	Не е приложимо	Не е приложимо
Образование	Образованието показва разпределението на индивидите в различните области на обучение.	Синтезирана променлива	Разпределението е базирано на информация от Националния статистически институт.	**Номинална**:** От Инфостат	Разкрива тенденции в образователната структура на населението, като откроява разликите в достъпа до образователни ресурси и възможностите за професионално развитие.	Официални данни – НСИ	Неправилно отчитане на образователни нива, пристрастия при липса на отговор, грешки при обработката на данни и неточности при класифицирането на образователните нива	Данни за България	Тези данни са статични и отразяват образованието през референтния период на преброяването - 2021 г.	Не е приложимо	Не е приложимо

Име на функция	Описание	Метод на изчисление	Метод на изчисление. Източници на данни	Мерна единица	Уместност	Грешка при извадката	Грешка, свързана с липса на семплиране	Географско	Временна дезагрегация	Сравнимост	Регионално
Темперамент	Темпераментът отразява разпределението на индивидите по различни личностни черти.	Синтезирана променлива	Разпределението е получено от международно проучване с повече от 15 000 респонденти.	**Номинална:** - Холерик - Флегматичен - Сангвиничен - Меланхоличен	Предоставя информация за личностни черти и поведенчески модели, като подчертава тяхното влияние върху социалните взаимодействия.	Минимална вероятност, в данните от проучването	Грешките в темперамента могат да включват пристрастия в самооценката или субективност.	Данни за България	Необходими са периодични актуализации на данните през относително дълъг период (>5 години)	Не е приложимо	Не е приложимо
Индивидуална склонност към риск	Отразява разпределението на индивидите въз основа на тяхната готовност да поемат инвестиционни рискове.	Синтезирана променлива	Разпределението е получено от вътрешно проучване с повече от 900 респонденти.	**Непрекъсната:** 0 - 1	Индивидуалните предпочитания за риск отразяват изработването на решения в условия на несигурност, подчертавайки поведението на индивидите на финансовите пазари.	Безопасна среда. Трудности при оценката на поведението в реална ситуация.	Може да включва неточно самоотчитане. Безопасна среда. Трудности при оценката на поведението в реална ситуация.	Данни за България	Текущи изследвания. Стабилни резултати. Не се очакват резки колебания.	Не е приложимо	Не е приложимо

Име на функция	Описание	Метод на изчисление	Метод на изчисление. Източници на данни	Мерна единица	Уместност	Грешка при извадката	Грешка, свързана с липса на семплиране	Географско	Времева дезагрегация	Сравним ост	Региони
Инвестиционен опит	Предишният инвестиционен опит, измерен в години, отразява историята на индивида в инвестирането и изработването на финансови решения.	Синтезирана променлива	Правдоподобно предположение и априорна симулация	**Ординална*:** 0; 1-5; 6-10; 11-15; 16-25 *Интервална променлива, преобразувана в ординална	Отразява вземането на решения в областта на инвестициите, като подчертава финансовото поведение и неговото въздействие върху икономическите избори.	Минимални в данните от проучването	Грешките могат да включват неправилно тълкуване на финансови термини, субективно отчитане или неточна самооценка.	Данни за България	Необходими са периодични актуализации на данните за относително дълъг период (>5 години)	Не е приложимо	Не е приложимо
Акции	Показва разпределението на лицата, които инвестират в акции.	Синтезирана променлива	Разпределението е извлечено от публично достъпни данни от БНБ.	**Двоична:** Да/Не	Сегментът на акциите отразява наличието или отсъствието на инвестиции в акции, като подчертава финансовото поведение на инвеститорите.	Официални данни от БНБ	Грешки за сегмента на акциите могат да възникнат поради неточно самоотчитане.	Данни за България	Необходими са периодични актуализации на данните. Не по-често от веднъж годишно.	Не е приложимо	Не е приложимо

Име на функция	Описание	Метод на изчисление	Метод на изчисление. Източници на данни	Мерна единица	Уместност	Грешка при извадката	Грешка, свързана с липса на семплиране	Географско	Временна дезагрегация	Сравнимост	Регионално
Облигации	Показва разпределението на лицата, които инвестират в облигации.	Синтезирана променлива	Разпределението е извлечено от публично достъпни данни от БНБ.	**Двоична:** Да/Не	Отразява инвестициите в дългови инструменти, като подчертава финансовото поведение на инвеститорите	Официални данни от БНБ	Потенциалните неточности могат да произтичат от неправилно отчитане или непълно представяне.	Данни за България	Необходими са периодични актуализации на данните. Не по-често от веднъж годишно.	Не е приложимо	Не е приложимо
Други	Показва разпределението на лицата, които инвестират в други инвестиционни инструменти	Синтезирана променлива	Разпределението е извлечено от публично достъпни данни от БНБ.	**Двоична:** Да/Не	Другите инвестиции отразяват наличието или отсъствието на инвестиции в други инвестиционни инструменти.	Официални данни от БНБ	Потенциалните неточности могат да произтичат от неправилно отчитане или непълно представяне.	Данни за България	Необходими са периодични актуализации на данните. Не по-често от веднъж годишно.	Не е приложимо	Не е приложимо
Инвестиционни фондове	Показва разпределението на инвеститорите, които имат инвестиционен опит с валути	Синтезирана променлива	Разпределението е извлечено от публично достъпни данни от БНБ.	**Двоична:** Да/Не	Осигурява данни, необходими за оценка на персоналните инвестиционни подходи, управлението на капитал и анализ на финансовия рисков профил.	Официални данни от БНБ	Неточности могат да възникнат поради неправилна класификация.	Данни за България	Необходими са периодични актуализации на данните. Не по-често от веднъж годишно.	Не е приложимо	Не е приложимо

Име на функция	Описание	Метод на изчисление	Метод на изчисление. Източници на данни	Мерна единица	Уместност	Грешка при извадката	Грешка, свързана с липса на семплиране	Географско	Временна дезагрегация	Сравнимост	Регионално
Пари в брой	Това показва разпределението на инвеститорите, които имат инвестиционен опит с валути.	Синтезирана променлива	Разпределението е извлечено от публично достъпни данни от БНБ.	**Двоична:** Да/Не	Инвестирането във валута предоставя информация за диверсификация на портфолиото и познания за валутните пазари.	Официални данни от БНБ	Неточности могат да възникнат поради неправилна класификация на валутните инвеститори	Данни за България	Необходими са периодични актуализации на данните. Не по-често от веднъж годишно.	Не е приложимо	Не е приложимо
Криптовалута	Предоставя информация за инвеститори, които имат опит с криптовалути.	Синтезирана променлива	Разпределението е извлечено от публично достъпни данни от БНБ.	**Двоична:** Да/Не	Индикатор за участието на дадено лице на пазара на цифрови активи. Той отразява и по-широките тенденции в приемането на децентрализирани финанси.	Официални данни от БНБ	Неточности могат да възникнат от неправилно отчитане, липса на видимост в частните портфейли с криптовалута.	Данни за България	Необходими са периодични актуализации на данните. Не по-често от веднъж годишно.	Не е приложимо	Не е приложимо

Име на функция	Описание	Метод на изчислен ие	Метод на изчисление. Източници на данни	Мерна единица	Уместност	Грешка при извадката	Грешка, свързана с липса на семплиране	Географс ко	Временна дезагрегация	Сравним ост	Регионн а
Държавни облигации	Предоставя информация дали разглежданите инвеститори имат опит с инвестиции в държавни облигации.	Синтезирана променлива	Разпределението е извлечено от публично достъпни данни от БНБ.	**Двоична:** Да/Не	Индикатор за индивидуално инвестиционно поведение в нискорискови, стабилни финансови инструменти. Те предоставят информация за финансовите стратегии и доверието в държавните облигации.	Официални данни от БНБ	Възможно е да възникнат неточности поради неправилно отчитане.	Данни за България	Необходими са периодични актуализации на данните. Не по-често от веднъж годишно.	Не е приложимо	Не е приложимо
Депозити	Предоставя информация дали разглежданите инвеститори имат опит с инвестиции в банкови депозити.	Синтезирана променлива	Разпределението е извлечено от публично достъпни данни от БНБ.	**Двоична:** Да/Не	Предоставя информация за финансовите навици на хората, проникването на банкови продукти и доверието на потребителите в банковата система.	Официални данни от БНБ	Неточности могат да бъдат резултат от неправилно отчитане.	Данни за България	Необходими са периодични актуализации на данните. Не по-често от веднъж годишно.	Не е приложимо	Не е приложимо

Име на функция	Описание	Метод на изчисление	Метод на изчисление. Източници на данни	Мерна единица	Уместност	Грешка при извадката	Грешка, свързана с липса на семплиране	Географско	Времева дезагрегация	Сравнимост	Регионално
Доход	Показва разпределението на доходите между различните доходни групи в рамките на дадена популация. Данните предоставят информация за съответния процент на лицата, попадащи във всяка от тях.	Синтезирана променлива	Разпределението на личните доходи се основава на данни, събрани чрез анкета сред 491 души.	**Ординална*:** до 19 200 19 201 - 27 600 27 601 - 54 000 54 001 - 82 800 от 82 801 *Интервална променлива, преобразувана в ординална	Разпределението на доходите е от решаващо значение за разбирането на икономическото неравенство, социалната стратификация и концентрацията на богатство в рамките на населението.	Възможни са несъответствия в данните, ако извадката от проучването не съответства напълно на популацията.	Може да възникне изкривяване, ако данните са попълнени неправилно	Данни за България	Динамична променлива. Необходими са актуализации на данните веднъж годишно.	Не е приложимо	Не е приложимо

Име на функция	Описание	Метод на изчисление	Метод на изчисление. Източници на данни	Мерна единица	Уместност	Грешка при извадката	Грешка, свързана с липса на семплиране	Географско	Времева дезагрегация	Сравнимост	Регионално
Лични разходи	Представлява общите лични разходи на дадено лице. Тези данни помагат да се разбере поведението на харчене на хора в различните доходни групи.	Синтезиран променлива	Разпределението на разходите се извлича от априорни знания и предположения за видовете домакински разходи.	**Ординална*:** Група 1 Група 2 Група 3 Група 4 *Интервална променлива, преобразувана в ординална	Индикатор за оценка на финансовото здраве, икономическото поведение и моделите на потребление в различни демографски групи. Той помага да се открият потенциални области за подобрение на поведенческите модели при спестяване.	Възможни несъответствия в допусканията, ако средата се промени. Пример – инфлация.	Потенциалните проблеми включват непоследователни класификации на разходите или пристрастия при категоризирането.	Данни за България	Динамична променлива. Необходими са актуализации на данните веднъж годишно.	Не е приложимо	Не е приложимо

Име на функция	Описание	Метод на изчисление	Метод на изчисление. Източници на данни	Мерна единица	Уместност	Грешка при извадката	Грешка, свързана с липса на семплиране	Географско	Времева дезагрегация	Сравнимост	Регионално
Жилищни разходи	Отнася се до общите разходи, направени от дадено лице за разходи, свързани с жилище, включително наем, ипотечни плащания, комунални услуги и поддръжка. Тези разходи са от основно значение за разбирането на финансовата стабилност и поведенческите модели.	Синтезирана променлива	Разпределението на разходите се извлича от априорни знания и предположения за видовете домакински разходи.	**Ординална*:** Група 1 Група 2 Група 3 Група 4 *Интервална променлива, преобразувана в ординална	Мярка за финансово благополучие, оценяваща достъпността до жилище в различни икономически контексти. Тя помага да се установи каква част от дохода на домакинството е предназначена за жилище.	Възможни несъответствия в допусканията, ако средата се промени. Пример - инфлация.	Възможно е да възникнат несъответствия в определението или категоризацията.	Данни за България	Динамична променлива. Необходими са актуализации на данните веднъж годишно.	Не е приложимо	Не е приложимо

Име на функция	Описание	Метод на изчисление	Метод на изчисление. Източници на данни	Мерна единица	Уместност	Грешка при извадката	Грешка, свързана с липса на семплиране	Географско	Времева дезагрегация	Сравнимост	Регионално
Данъци	Отнася се до общата сума на данъците и осигурителните вноски, платени от дадено лице, включително данъци върху доходите, социално осигуряване, пенсионни вноски, здравно осигуряване и други задължителни плащания.	Синтезирана променлива	Разпределението на разходите се извлича от априорни знания и предположения за видовете домакински разходи.	**Ординална*:** Група 1 Група 2 Група 3 Група 4 *Интервална променлива, преобразувана в ординална	Данъците и осигурителните вноски са важни за оценката на индивидуалните финансови задължения и разбирането на въздействието на данъчното облагане върху разполагаемия доход.	Възможни несъответствия в допусканията, ако средата се промени. Пример - инфлация.	Възможно е да възникнат несъответствия в определението или категоризацията.	Данни за България	Динамична променлива. Необходими са актуализации на данните веднъж годишно.	Не е приложимо	Не е приложимо

Име на функция	Описание	Метод на изчисление	Метод на изчисление. Източници на данни	Мерна единица	Уместност	Грешка при извадката	Грешка, свързана с липса на семплиране	Географско	Временна дезагрегация	Сравнимост	Регионално
Разходи за транспорт и телекомуникации	Представя общите разходи, които дадено лице прави за транспорт (напр. обществен транспорт, разходи за автомобил, таксита) и комуникационни услуги (напр. сметки за мобилен телефон, интернет, пощенски услуги).	Синтезирана променлива	Разпределението на разходите се извлича от априорни знания и предположения за видовете домакински разходи.	**Ординална*:** Група 1 Група 2 Група 3 Група 4 *Интервална променлива, преобразувана в ординална	От съществено значение е по отношение на разбирането на моделите на мобилност на индивидите и техния достъп до комуникация. Тези разходи подчертават разликите във финансовите възможности между групите.	Възможни несъответствия в допусканията, ако средата се промени. Пример - инфлация.	Възможно е да възникнат несъответствия в определението или категоризацията.	Данни за България	Динамична променлива. Необходими са актуализации на данните веднъж годишно.	Не е приложимо	Не е приложимо

Име на функция	Описание	Метод на изчисление	Метод на изчисление. Източници на данни	Мерна единица	Уместност	Грешка при извадката	Грешка, свързана с липса на семплиране	Географско	Времева дезагрегация	Сравнимост	Регионално
Разходи за хоби и лично време	Представя общите разходи, които дадено лице отделя за развлекателни дейности, хобита и забавления. Тази функция предоставя информация за начина на живот и приоритетите на дадено лице.	Синтезирана променлива	Разпределението на разходите се извлича от априорни знания и предположения за видовете домакински разходи.	**Ординална*:** Група 1 Група 2 Група 3 Група 4 *Интервална променлива, преобразувана в ординална	Разходите за хобита са важни за разбирането на предпочитанията за начин на живот на даден човек. Те са показател за икономическо благополучие.	Възможни несъответствия в допусканията, ако средата се промени. Пример - инфлация.	Възможно е да възникнат несъответствия в определението или категоризацията.	Данни за България	Динамична променлива. Необходими са актуализации на данните веднъж годишно.	Не е приложимо	Не е приложимо

Име на функция	Описание	Метод на изчислен ие	Метод на изчисление. Източници на данни	Мерна единица	Уместност	Грешка при извадката	Грешка, свързана с липса на семплиране	Географс ко	Времена дезагрегация	Сравним ост	Регионн а
Предпочитан метод за банкиране	Показва предпочитания начин на банкиране за физическите лица, независимо дали предпочитат онлайн банкиране или банкиране в банката. Тази функция предоставя информация за тенденциите в дигиталното внедряване и регионалните или демографските различия в банковото поведение.	Синтезирана променлива	Разпределението е извлечено от данни на НСИ.	**Номинална:** Онлайн/Офлайн	Тази характеристика е от решаващо значение за разбирането на потребителското поведение и насочването на решенията относно разпределението на ресурсите, както и за описание на нивото на доверие в цифровите платежни системи.	Официални данни от НСИ	Възможно е да възникнат несъответствия, ако има лица с редовно банково обслужване както онлайн, така и офлайн.	Данни за България	Актуализациите на данните са необходими веднъж годишно.	Не е приложимо	Не е приложимо

Име на функция	Описание	Метод на изчисление	Метод на изчисление. Източници на данни	Мерна единица	Уместност	Грешка при извадката	Грешка, свързана с липса на семплиране	Географско	Временна дезагрегация	Сравнимост	Регионално
Среден брой банкови операции	Предоставя информация за активния потребител на банкови услуги, като анализира средния брой банкови транзакции, извършени от дадено лице за един месец. Това включва депозити, тегления, преводи, плащания на сметки и др.	Синтезирана променлива	Априори симулирано разпределение	**Ординална:** до 10 11 - 14 15 - 20 21 - 26 от 27	Честотата на банковите транзакции е от решаващо значение за финансовите институции, за да оценят активността на клиентите. Тези данни могат да помогнат и при определянето на най-често използваните банкови услуги, както и в процеса на сегментиране на клиентите.	Възможни несъответствия в допусканията. Възможност за пристрастност в извадката (фокусиране само върху определена група клиенти).	Потенциалните грешки включват неправилно категоризиране на трансакции или пропуснати трансакции, които са се случили на платформи извън регистрираните системи на банката.	Данни за България	Актуализация на данните е необходима веднъж годишно.	Не е приложимо	Не е приложимо

Име на функция	Описание	Метод на изчисление	Метод на изчисление. Източници на данни	Мерна единица	Уместност	Грешка при извадката	Грешка, свързана с липса на семплиране	Географско	Временна дезагрегация	Сравнимост	Регионална
Дебитна карта	Представя процентът на хората, които притежават една или повече дебитни карти. Функцията предоставя информация за разпространението на банкови услуги сред населението.	Синтезирана променлива	Априори симулирано разпределение	**Интервална:** 0; 1; 2; 3	Броят на притежаваните дебитни карти е важен за разбиране на поведението на клиентите. Потребителите, които притежават няколко дебитни карти, може да имат различни банкови нужди, като например отделни карти за лична и бизнес употреба или за различни категории разходи.	Възможни несъответствия в предположенията.	Потенциалните грешки включват неправилно отчитане или липса на яснота.	Данни за България	Актуализациите на данните са необходими на всеки три/пет години.	Не е приложимо	Не е приложимо

Име на функция	Описание	Метод на изчислен ие	Метод на изчисление. Източници на данни	Мерна единица	Уместност	Грешка при извадката	Грешка, свързана с липса на семплиране	Географс ко	Временна дезагрегация	Сравним ост	Регионн а
Банкова сметка	Показва процентът на лицата, които притежават банкова сметка в евро. Тази функция помага да се разбере проникването на основните банкови услуги сред различните клиентски сегменти, особено по отношение на наличието на сметка.	Синтезирана променлива	Разпределението е извлечено от публично достъпна информация за банковата система.	**Двоична:** Да/Не	Притежаването на банкова сметка е най-важният показател за икономически активния клиент. Банковата сметка е основен инструмент за управление на финансите и взаимодействие с по-широката икономика, което прави тази характеристика критична за разбирането на финансовите навици.	Възможни са несъответствия в данните, ако извадката от проучването не съответства напълно на популацията.	Неразбиране на видовете акаунти или неточно отчитане на собствеността.	Данни за България	Актуализациите на данните са необходими на всеки три/пет години.	Не е приложимо	Не е приложимо

* Синтезирано чрез симулационен метод, базиран на разпределения и бизнес логика

Приложение 3 - Номинални променливи

Име на функцията	Тип	Източник на данни	Възможни стойности
Пол	Номинална	От преброяването през 2021 г.	- М - Ф
Трудов статус	Номинална	От преброяването през 2021 г. & проучване на работната сила	-Работодатели -Самостоятелно наети -Служители в частни предприятия -Служители в държавни предприятия -Неплатени семейни работници -Безработен
Семейно положение	Номинална	От преброяването през 2021 г.	- Неженен/а - Женен/а - Разведен/а - Вдовец/Вдовица
Националност	Номинална	От преброяването през 2021 г.	- България - Европейски съюз - Други
Религия	Номинална	От преброяването през 2021 г.	- Православен - Протестант - Католик - Мюсюлманин - Друг - Атеист/Без религия - Не искам да отговоря

Име на функцията	Тип	Източник на данни	Възможни стойности
Професия/Индустрия	Номинална	От НСИ	- Селско, горско и рибно стопанство - Добивна промишленост и преработваща промишленост - Снабдяване с електроенергия, газ, пара и климатизация. Водоснабдяване, канализация, управление на отпадъци и дейности по възстановяване - Строителство - Търговия на едро и дребно; ремонт на автомобили и мотоциклети - Транспорт и складови дейности - Хотелиерство и ресторантьорство - Информация и комуникации - Финансови и застрахователни дейности. Операции с недвижими имоти - Образователни, професионални, научни и технически дейности. - Административни и спомагателни дейности. Държавна администрация и отбрана; задължително социално осигуряване - Дейности в областта на хуманното здравеопазване и социалната работа - Изкуства, развлечения и отдих. Други обслужващи дейности
Професионален статус	Номинална	От НСИ	- Договор за управление - Трудов договор - Граждански договор - Самостоятелно нает - Безработен - Пенсионер
Собственик на апартамент/къща	Номинална	От преброяването на населението през 2021 г. и НСИ	- Да - Не

Име на функцията	Тип	Източник на данни	Възможни стойности
Образование	Номинална	От НСИ	- Педагогически науки - Хуманитарни науки - Социални, икономически и правни науки - Природни науки, математика и информатика - Технически науки - Селскостопански науки и ветеринарна медицина - Здраве и спорт - Изкуства - Сигурност и отбрана
Темперамент	Номинална	Международно проучване [4]	- Холерик - Флегматик - Сангвиник - Меланхолик
Предпочитан метод за банкиране	Номинална	От НСИ	- Онлайн - Офлайн

** Номинални променливи с подробни възможни стойност