Кейс «Антифрод»

1 Общее описание

Организации-участники: **Ассоциация больших данных**.

Отрасль: финансовый сектор, обнаружение мошенничества.

Номер кейса: **ABD-FRAUD-DPSYNTH-2025-003**

Цель: синтез транзакционных данных с выраженным дисбалансом классов (1 % мошеннических транзакций) для задач разработки и тестирования моделей обнаружения мошенничества. Исходные данные: 20 000 транзакций, 14 атрибутов, из которых 200 — мошеннические (IS_FRAUD = 1).

Кейс демонстрирует фундаментальную проблему: при стандартном DP-синтезе лапласовский шум с масштабом ~1/ε имеет амплитуду, сопоставимую с сигналом от 200 fraud-записей, что ведёт к вымыванию fraud-паттернов при малом ε.

2 Цель тестирования

Настоящий документ описывает тестирование конвейера синтеза табличных данных с гарантиями дифференциальной приватности (DP) применительно к предметной области обнаружения мошенничества в банковских транзакциях (fraud detection – далее ANTI-FRAUD). Тестирование проводилось на суррогатных данных, моделирующих структуру и статистические свойства реальных транзакционных данных с экстремальным дисбалансом классов (~1% мошеннических операций).

Цели тестирования:

- проверка возможности синтеза транзакционных данных с формальными DP-гарантиями при сохранении аналитической полезности для задачи ANTI-FRAUD;
- сравнение трёх методов DP-синтеза (DP-Copula, DP-MST, Stratified DP) по метрикам качества, приватности и прикладной полезности;
- исследование метода Stratified DP как способа сохранения minority-class паттернов при DP-синтезе с экстремальным дисбалансом классов;
- введение Pattern Reconstruction Attack как метода оценки утечки бизнес-логики антифрод-моделей;
- определение рекомендуемого диапазона бюджета приватности ε для задач ANTI-FRAUD.

Кейс является третьим в серии формальных отчётов, подготовленных в качестве приложений к ПНСТ по синтезу данных. Предшествующие кейсы: «Маркетинговая атрибуция» (ABD-MA-DPCOPULA-2024-001) и «Синтетические резюме» (ABD-HH-DPSYNTH-2025-002).

Примечание: Все численные результаты получены на суррогатных данных. Стресс-тесты оценивают свойства конвейера синтеза, а не конкретного реального набора данных. Для перехода к production-оценкам необходимо повторение экспериментов на реальных данных с соответствующими организационными мерами защиты.

3 Предметная область

Банковские транзакции содержат информацию, защищённую одновременно несколькими нормативными актами:

- **Федеральный закон от 27.07.2006 № 152-ФЗ** «О персональных данных» - персональные данные клиентов;
- **Федеральный закон от 07.08.2001 № 115-ФЗ** «О противодействии легализации (отмыванию) доходов, полученных преступным путём, и финансированию терроризма» - AML-требования;
- нормы о **банковской тайне** (Федеральный закон от 02.12.1990 № 395-1 «О банках и банковской деятельности»).

Задача синтеза данных в данной предметной области возникает в следующих сценариях:

- **обучение и валидация антифрод-моделей** - ML-инженеры получают синтетический датасет для разработки и тестирования моделей обнаружения мошенничества;
- **передача данных внешним подрядчикам** - консалтинговые компании и исследовательские лаборатории работают с синтетикой вместо реальных транзакций;
- **бенчмаркинг и стресс-тестирование** - сравнение антифрод-систем на синтетических данных без раскрытия операционных паттернов;
- **data augmentation для minority class** - генерация дополнительных fraud-примеров для устранения дисбаланса в обучающих выборках.

Помимо регуляторных ограничений, синтез помогает решить проблему дефицита данных для класса minority: мошеннические транзакции составляют порядка 1% от общего объёма, что является критическим ограничением при разработке моделей ANTI-FRAUD.

3.1 Участники взаимодействия

Как и в кейсе «Синтетические резюме» (ABD-HH-DPSYNTH-2025-002), в настоящем кейсе участвует **один владелец данных** - банк. В отличие от кейса «Маркетинговая атрибуция» (ABD-MA-DPCOPULA-2024-001), PSI-этап не требуется.

Таблица 1. Основные участники

Роль	Описание
Владелец данных	Банк, хранящий транзакционные данные клиентов
Оператор синтеза	АБД - выполняет синтез данных с DP-гарантиями
Получатель данных	Подрядчик / аналитик / ML-инженер - использует синтетический датасет

3.2 Начальные наборы данных

Для тестирования сгенерирован суррогатный датасет, моделирующий структуру реальных транзакционных данных банка.

Таблица 2. Структура суррогатного датасета FRAUD_TXN

№	Атрибут	Тип	Описание	Распределение
1	TXN_ID	int	Уникальный идентификатор (удаляется перед синтезом)	Последовательный
2	AMOUNT	float	Сумма транзакции, руб.	LogNormal (μ=7.5, σ=1.2), [50, 500 000]
3	MERCHANT_CATEGORY	cat	Категория мерчанта (10 значений)	Разные веса для fraud/legit
4	CHANNEL	cat	Канал операции (4 значения)	Разные веса для fraud/legit
5	HOUR	int	Час транзакции (0–23)	Дневной пик (legit) / ночное смещение (fraud)
6	DAY_OF_WEEK	int	День недели (0–6)	Равномерное
7	IS_INTERNATIONAL	binary	Международная транзакция	Bernoulli (0.08 legit / 0.22 fraud)
8	CARD_AGE_DAYS	int	Возраст карты, дней	Exp (300 legit / 150 fraud)
9	DAYS_SINCE_LAST_TXN	float	Дней с последней транзакции	Exp (3.0 legit / 1.2 fraud)
10	AVG_AMOUNT_30D	float	Средняя сумма за 30 дней	LogNormal, коррелирует с AMOUNT
11	TXN_COUNT_7D	int	Число транзакций за 7 дней	Poisson (4 legit / 7 fraud)
12	DISTANCE_FROM_HOME	float	Расстояние от дома, км	Exp (15 legit / 45 fraud)

№	Атрибут	Тип	Описание	Распределение
13	IS_NEW_MERCHANT	bi-nary	Новый мерчант	Bernoulli (0.15 legit / 0.35 fraud)
14	RISK_SCORE_LEGACY	float	Скоринг из legacy-системы	Beta (2,8 legit / 3,6 fraud)
15	IS_FRAUD	bi-nary	Метка мошенничества	~1% (200 из 20 000)

Объём: 20 000 записей, 15 атрибутов (14 после удаления TXN_ID).

Заложенные корреляции. Суррогатный генератор воспроизводит реалистичные зависимости, характерные для транзакционных данных:

- **AMOUNT ↔ AVG_AMOUNT_30D**: через общий «профиль клиента» (для legit: 0.7 × AMOUNT + 0.3 × AVG);
- **TXN_COUNT_7D ↔ DAYS_SINCE_LAST_TXN**: обратная зависимость - чем больше транзакций, тем меньше интервал;
- **CARD_AGE ↔ IS_NEW_MERCHANT**: молодые карты (< 90 дней) чаще используют новых мерчантов;
- **мультимодальные категории**: распределения MERCHANT_CATEGORY и CHANNEL различаются для fraud и legit - для fraud-транзакций наибольший вес имеют Gambling (33%), Electronics (20%) и Travel (18%); для легитимных - Grocery (22%), Restaurant (15%), Clothing (12%).

Модель генерации fraud-меток. Метка IS_FRAUD определяется через латентный fraud-скор - логистическую модель от 8 предикторов с шумом N (0, 1.3):

$$\text{fraud\_score} = 0.35 \cdot \ln\left(\frac{\text{AMOUNT}}{\text{median}} + 1\right) + 0.30 \cdot \mathbb{1}[\text{HOUR} \in [0,5]] + 0.25 \cdot \text{IS\_INTL} + 0.25 \cdot \left(1 - \frac{\text{CARD\_AGE}}{365}\right) + 0.20 \cdot \frac{\text{TXN\_COUNT}}{10} + 0.20 \cdot \frac{\text{DISTANCE}}{100} + 0.20 \cdot \text{IS\_NEW\_MERCH} + 0.15 \cdot \text{RISK\_LEGACY} + \mathcal{N}(0,1.3)$$

Шум обеспечивает нетривиальность задачи: базовой линии AUC-ROC ≈ 0.89 (а не 0.99+).

Пост-коррекция fraud-записей. 70% fraud-записей получают «видимые» паттерны (повышенные суммы, ночные часы, международные транзакции, молодые карты, burst-паттерн активности), а 30% - «тихий фрод» без явных аномалий. Это моделирует реальную ситуацию, когда часть мошеннических транзакций мимикрирует под легитимные.

3.3 Целевой признак и целевая задача

В качестве целевой задачи выбрано **обнаружение мошенничества** (бинарная классификация):

$$\hat{y} = f(\text{AMOUNT,MERCHANT,CHANNEL,HOUR,DAY,IS\_INTL,CARD\_AGE,DAYS\_SINCE,}\\ \text{AVG\_30D,TXN\_COUNT,DISTANCE,IS\_NEW,RISK\_LEGACY})$$

В отличие от кейса «Синтетические резюме» (регрессия зарплаты) и кейса «Маркетинговая атрибуция» (бинарная классификация конверсии, ~15%), здесь оценивается способность синтетических данных сохранять паттерны малых классов (minority class) при экстремальном дисбалансе (~1% fraud).

Для оценки используется протокол **TSTR** (Train on Synthetic, Test on Real): модель обучается на синтетических данных и тестируется на реальных. Результаты усреднены по 3 seed (42, 43, 44) для стабильности.

Метрики целевой задачи: AUC-ROC, AUC-PR (Average Precision), Recall@FPR=1%, Precision@top-1%, F1-score

4 Архитектура системы синтеза

4.1 Конвейер данных

Архитектура конвейера включает пять модулей. Как и в кейсе «Синтетические резюме», PSI-этап отсутствует - данные поступают от одного владельца напрямую в модуль синтеза.

Таблица 3. Модули конвейера

Модуль	Назначение	Входные данные	Выходные данные
M1	Генерация суррогатных данных	Параметры распределений	tbl_fraud_txn_raw (20 000 × 15)
M2	Кодирование и нормализация	tbl_fraud_txn_raw	Закодированная матрица (20 000 × 14)
M3	DP-синтез (Copula / MST / Stratified)	Закодированная матрица, ε	tbl_fraud_txn_syn (20 000 × 14)
M4	Целевая задача (TSTR)	tbl_fraud_txn_raw, tbl_fraud_txn_syn	AUC-ROC, AUC-PR, Recall@FPR=1%, F1
M5	Аудит приватности	tbl_fraud_txn_raw, tbl_fraud_txn_syn	CVPL, MIA, Singling Out, Pattern Reconstruction

4.2 Специфика мошеннических данных: экстремальный дисбаланс классов

Ниже описывается специфика кейса ANTI-FRAUD, отсутствующую в кейсах «Маркетинговая атрибуция» и «Синтетические резюме»: проблему сохранения малых классов (minority-class) паттернов при DP-синтезе.

При стандартном DP-синтезе 20 000 записей Laplace-шум с масштабом ~1/ε добавляется ко всем записям одинаково. При соотношении fraud: legit = 1 : 99 (200 vs. 19 800 записей) сигнал от fraud-записей тонет в шуме: гистограммы fraud-подкласса содержат лишь единичные значения в каждом бине, тогда как шум имеет сопоставимую или большую амплитуду. Результат: при малом ε fraud-паттерны полностью вымываются, модели, обученные на синтетике, не способны обнаруживать мошенничество.

Для решения этой проблемы в настоящем кейсе введён метод **Stratified DP** (раздел 2.3.3), основанный на раздельном синтезе по стратам с использованием теоремы о параллельной композиции.

4.3 Методы синтеза

В тестировании использованы три метода DP-синтеза табличных данных. Первые два (DP-Copula и DP-MST) применялись в предыдущих кейсах; третий (Stratified DP) является нововведением настоящего кейса.

DP-Copula (Gaussian Copula с DP-маргиналями)

Метод основан на разделении совместного распределения данных на маргинальные распределения и структуру зависимостей (копулу). Алгоритм:

1. Вычисление эмпирической копулы: ранговая трансформация $U_{ij} = R(x_{ij})/(n+1)$.
2. Стандартизация: $U^* = (U - 0.5) \cdot \sqrt{12}$.
3. Оценка корреляционной матрицы Σ с добавлением DP-шума:

$$\Sigma_{\text{noisy}} = \Sigma + \mathcal{N}(0,\ \sigma^2 I), \sigma = \frac{\Delta \cdot \sqrt{2\ln(1.25/\delta)}}{\varepsilon_{\text{corr}}}$$

где $\Delta = 4d/n$ - чувствительность, d - размерность, n - число записей.

4. Проекция на PSD-конус (спектральное усечение собственных значений):

$$\Sigma_{\text{PSD}} = V \cdot \max(\Lambda, \lambda_{\min}) \cdot V^T$$

5. DP-маргинали: гистограммы с лапласовым шумом $Lap(\frac{d}{\varepsilon_{marg}})$.
6. Генерация: семплирование из $N(0, \Sigma_{PSD})$ → обратная CDF-трансформация по DP-маргиналям.
7. Обратное декодирование категориальных признаков; клиппинг значений в допустимые диапазоны.

Распределение бюджета: $\varepsilon_{marg} = 0.4\varepsilon$, $\varepsilon_{corr} = 0.6\varepsilon$.

Пост-коррекция IS_FRAUD. Если количество fraud-записей в синтетике опускается ниже порога ($< max(5, 0.1 \cdot n \cdot fraud_{rate})$), маргиналь IS_FRAUD пересемплируется из DP-оценки fraud_rate.

Ограничение для кейса: ранговые корреляции Copula не сохраняют условные распределения $P(X \mid IS\_FRAUD)$, что критично при 1% minority class - fraud-паттерны «растворяются» в доминирующем сигнале legit-транзакций.

DP-MST (Maximum Spanning Tree)

Метод основан на аппроксимации совместного распределения деревом Байеса, построенным по DP-оценкам взаимной информации. Подход развивает идеи победителя NIST Differential Privacy Synthetic Data Challenge (2018).

Алгоритм:

1. Дискретизация числовых признаков в 15 квантильных бинов.
2. DP 2-way маргинали: для каждой пары признаков - двумерная гистограмма с лапласовым шумом.
3. Оценка взаимной информации (MI) из зашумлённых таблиц:

$$I(X_i; X_j) = \sum_{x,y} \hat{P}(x,y) \log \frac{\hat{P}(x,y)}{\hat{P}(x)\hat{P}(y)}$$

4. Построение Maximum Spanning Tree (MST) по матрице MI.
5. BFS-генерация по дереву: корневой признак - по DP-маргинали; остальные - по условным распределениям P(X_j | X_parent) с добавлением DP-шума.
6. Обратная трансформация: дискретные бины → непрерывные значения (равномерное семплирование внутри бина).

Распределение бюджета: ε_MI = ε/3, ε_root = ε/3, ε_cond = ε/3.

Пост-коррекция IS_FRAUD. Маргиналь IS_FRAUD пересемплируется из DP-оценки fraud_rate, поскольку BFS-обусловливание сильно искажает minority class.

Stratified DP

Stratified DP (стратифицированный синтез с параллельной композицией) является ключевым методологическим нововведением настоящего кейса, отсутствующим в кейсах MA и HH.

Мотивация описана в разделе 2.2: стандартный DP-синтез разрушает fraud-паттерны при малом ε. Решение - раздельный синтез fraud и non-fraud страт с использованием теоремы о параллельной композиции.

Алгоритм:

1. Стратификация данных по IS_FRAUD: D_fraud (200 записей), D_legit (19 800 записей).
2. Выделение бюджетов: $\varepsilon_{fraud} = \alpha \cdot \varepsilon$, $\varepsilon_{legit} = (1-\alpha) \cdot \varepsilon$, где α = 0.7 (fraud-страта получает большую долю бюджета).
3. Синтез D_{fraud} через DP-Copula с бюджетом ε_fraud.
4. Синтез D_{legit} через DP-Copula с бюджетом ε_legit.
5. Объединение с сохранением оригинальных пропорций (~1% fraud) и перемешивание.

Гарантия приватности (теорема о параллельной композиции). Поскольку D_fraud ∩ D_legit = ∅, механизмы работают на непересекающихся подмножествах данных, и общий бюджет определяется максимумом, а не суммой:

$$\varepsilon_{\text{total}} = \max(\varepsilon_{\text{fraud}}, \varepsilon_{\text{legit}}) = \alpha \cdot \varepsilon$$

Это строже, чем sequential composition ($\varepsilon_1 + \varepsilon_2$). При α = 0.7 и номинальном ε = 4.0: ε_total = 0.7 × 4.0 = 2.8, а не 4.0.

Преимущество. Fraud-страта из 200 записей синтезируется с выделенным бюджетом 0.7ε, при этом шум масштабируется к размеру страты (200 записей), а не к полному датасету (20 000). Это позволяет сохранить fraud-специфичные паттерны (ночные часы, повышенные суммы, международные транзакции) даже при малых ε.

Сравнение методов

Характеристика	DP-Copula	DP-MST	Stratified DP
Модель зависимостей	Гауссова копула (глобальная корреляционная матрица)	Дерево Байеса (попарные условные)	Копула по стратам + parallel composition
Обработка категориальных	LabelEncoder → ранги	Дискретизация → гистограммы	LabelEncoder → ранги (по стратам)
Сильные стороны	Хорошее сохранение линейных корреляций	Лучшее сохранение локальных зависимостей	Сохранение minority-class паттернов
Слабые стороны	Предположение о гауссовости; вымывание minority	Потеря информации при дискретизации	Повышенный singling out; pattern leakage
Специфика для fraud	Fraud-паттерны растворяются	Fraud-паттерны нестабильны	Fraud-паттерны сохраняются (by design)

Примечание: В production-сценарии рекомендуется также рассматривать DP-CTGAN (GAN с DP-SGD), однако он требует значительных вычислительных ресурсов (GPU с ≥12 GB VRAM) и не был включён в текущую симуляцию по ограничениям среды выполнения.

4.4 Бюджет приватности

Применяется (ε, δ)-дифференциальная приватность. Механизм M обеспечивает (ε, δ)-DP, если для любых двух соседних наборов данных D и D' (отличающихся одной записью) и любого множества выходов S:

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\varepsilon} \cdot \Pr[\mathcal{M}(D') \in S] + \delta$$

Как и в кейсе «Синтетические резюме», бюджет расходуется на единственный этап синтеза:

$$\varepsilon_{\text{total}} = \varepsilon_{\text{synth}}$$

Для Stratified DP действует параллельная композиция: ε_total = max(ε_fraud, ε_legit) = 0.7ε - фактический бюджет строже номинального.

Таблица 4. Тестированные значения бюджета приватности

ε	Интерпретация	Типичные сценарии
1.0	Высокая приватность	Публичные выпуски данных, передача неизвестным третьим сторонам
2.0	Повышенная приватность	Передача доверенным подрядчикам с NDA
4.0	Умеренная приватность	Внутренняя аналитика, обучение ML-моделей
8.0	Базовая приватность	Прототипирование, разработка и тестирование
16.0	Минимальная приватность	Исследовательские задачи внутри организации

4.5 Протокол синтезирования

Последовательность этапов:

1. Генерация суррогатного датасета (20 000 записей, seed = 42).
2. Удаление TXN_ID; кодирование категориальных признаков (LabelEncoder).
3. Для каждого метода m ∈ {DP-Copula, DP-MST, Stratified-DP} и каждого ε ∈ {1.0, 2.0, 4.0, 8.0, 16.0}:
4. Синтез 20 000 записей.
5. Обратное декодирование категориальных признаков.

6. Клиппинг значений в допустимые диапазоны.
7. Пост-коррекция IS_FRAUD (Copula, MST: пересемплирование маргинали; Stratified: не требуется).
8. Вычисление метрик качества (KS, TV, Wasserstein, Corr Sim).
9. Аудит приватности (CVPL, MIA, Singling Out, Pattern Reconstruction Attack).
10. Целевая задача: TSTR (GradientBoostingClassifier, усреднение по 3 seed).

4.6 Параметры эксперимента

Таблица 5. Параметры синтеза и аудита

Параметр	Значение
Объём исходных данных	20 000 записей
Объём синтетических данных	20 000 записей (1:1)
Методы синтеза	DP-Copula, DP-MST, Stratified DP
Значения ε	1.0, 2.0, 4.0, 8.0, 16.0
δ	10^{-5}
Число бинов маргиналей (Copula)	30 (числовые), ≤ 50 (категориальные)
Число бинов дискретизации (MST)	15
Stratified DP: α (доля бюджета на fraud)	0.7
CVPL: размер контрольной выборки	500
MIA: размер выборки	1 000
MIA: метод	Propensity Score (LogisticRegression, C=0.1, top-5 фичей)
Целевая модель	GBC (n_estimators=60, depth=2, lr=0.08, subsample=0.5, max_features=0.7)

5 Результаты синтеза

5.1 Структура результирующего датасета

Синтетический датасет сохраняет ту же структуру (14 атрибутов после удаления TXN_ID), типы данных и допустимые диапазоны значений, что и исходный. Категориальные признаки декодированы обратно в исходные метки.

5.2 Сводная таблица метрик

Таблица 6. Сводные метрики качества и приватности (DP-Copula)

ε	Риск приватности	CVPL Score	MIA AUC	KS Sim	TV Sim	Corr Sim	Wasser-stein	SO Real	SO Syn
1.0	0.006	0.657	0.564	0.752	0.893	0.977	0.222	0.059	0.076
2.0	0.008	0.637	0.928	0.752	0.896	0.979	0.208	0.059	0.104
4.0	0.010	0.597	0.922	0.752	0.897	0.979	0.212	0.059	0.104
8.0	0.008	0.562	0.924	0.752	0.897	0.979	0.231	0.059	0.120
16.0	0.012	0.537	0.927	0.752	0.897	0.979	0.260	0.059	0.120

Таблица 7. Сводные метрики качества и приватности (DP-MST)

ε	Риск приватности	CVPL Score	MIA AUC	KS Sim	TV Sim	Corr Sim	Wasser-stein	SO Real	SO Syn
1.0	0.006	0.592	0.742	0.839	0.702	0.947	0.425	0.059	0.000
2.0	0.008	0.607	0.512	0.835	0.720	0.948	0.407	0.059	0.000
4.0	0.004	0.607	0.842	0.838	0.775	0.946	0.397	0.059	0.043
8.0	0.002	0.545	0.797	0.847	0.773	0.946	0.434	0.059	0.037
16.0	0.002	0.567	0.746	0.830	0.794	0.931	0.381	0.059	0.019

Примечание: SO Real / SO Syn - доля уникальных комбинаций квази-идентификаторов {HOUR_BIN, MERCHANT_CATEGORY, CHANNEL, IS_INTERNATIONAL} с k = 1 среди всех комбинаций (Singling Out Risk). MIA - AUC-ROC Propensity Score классификатора, а не accuracy (в отличие от кейсов MA и HH).

5.3 Интерпретация метрик

KS Similarity (Колмогоров–Смирнов): KS_Sim = 1 − sup|F_real(x) − F_syn(x)|. DP-Copula: KS ≈ 0.752 (стабильно по ε). DP-MST: 0.830–0.847 - более высокое KS-сходство благодаря квантильной дискретизации. Stratified-DP: 0.747–0.749 - чуть ниже из-за раздельного синтеза.

TV Similarity (Total Variation): TV_Sim = 1 − ½·Σ|p_real(x) − p_syn(x)|. DP-Copula: TV ≈ 0.893–0.897. DP-MST: 0.702–0.794 - ниже из-за дискретизации. Stratified-DP: 0.873–0.896 - близко к DP-Copula.

Wasserstein Distance: нормализованное расстояние Вассерштейна между стандартизированными распределениями. DP-Copula: 0.208–0.260; DP-MST: 0.381–0.434; Stratified-DP: 0.222–0.260 - сопоставимо с DP-Copula.

Correlation Similarity: Corr_Sim = 1 − mean(|C_real − C_syn|). DP-Copula: ≈ 0.977–0.979 (ожидаемо для метода, моделирующего корреляционную матрицу). DP-MST: 0.931–0.948. Stratified-DP: 0.951–0.980 (при ε=16.0 сравнивается с DP-Copula).

5.4 Анализ распределений

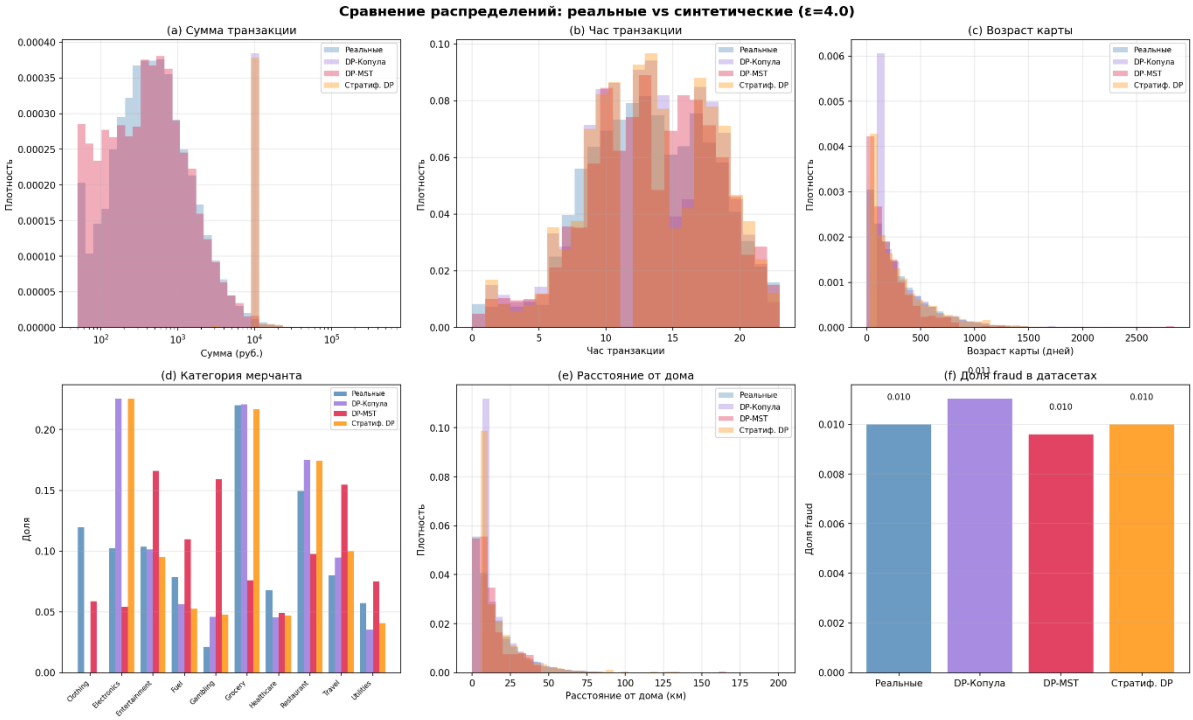

Рисунок 1. Сравнение распределений реальных и синтетических данных (ε = 4.0)

Визуальный анализ распределений при ε = 4.0 показывает:

- **Сумма транзакции (AMOUNT):** логнормальное распределение реальных данных воспроизводится всеми методами. DP-Copula точнее передаёт хвосты; DP-MST демонстрирует ступенчатость; Stratified-DP формирует две подмодальности (из-за раздельного синтеза страт).
- **Час транзакции (HOUR):** бимодальное распределение (дневной пик legit + ночное смещение fraud) сохраняется Stratified-DP лучше всего, поскольку fraud-компонента синтезируется отдельно.
- **Категория мерчанта (MERCHANT_CATEGORY):** различия в распределениях fraud vs legit (Gambling: 33% vs 2%) сохраняются Stratified-DP; Copula и MST усредняют до общего распределения.

– **IS_FRAUD:** целевой баланс ~1% корректно воспроизводится всеми методами после пост-коррекции.

5.5 Анализ корреляций

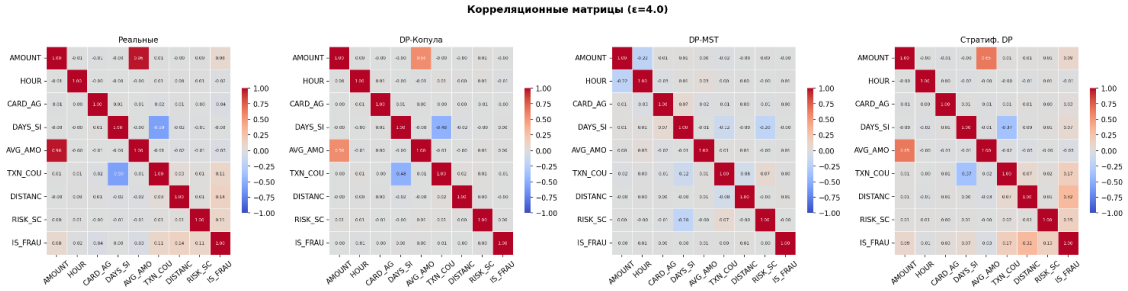

Рисунок 2. Корреляционные матрицы: реальные, DP-Copula, DP-MST, Stratified-DP (ε = 4.0)

Ключевые корреляции в суррогатных данных:

– AMOUNT ↔ AVG_AMOUNT_30D: r ≈ 0.60 - обусловлена генеративной моделью (общий «профиль клиента»);

– TXN_COUNT_7D ↔ DAYS_SINCE_LAST_TXN: r ≈ −0.25 - обратная зависимость;

– CARD_AGE ↔ IS_NEW_MERCHANT: r ≈ −0.15 - молодые карты чаще у новых мерчантов.

DP-Copula сохраняет эти корреляции с точностью до 2-го знака (Corr Sim ≈ 0.98). DP-MST допускает отклонения до 0.05–0.07. Stratified-DP при больших ε приближается к DP-Copula (Corr Sim = 0.980 при ε=16.0).

6 Стресс-тест: симуляция атак на синтетику

6.1 Модель угроз

Злоумышленник: подрядчик, аналитик или ML-инженер, получивший синтетический датасет, пытается (а) установить соответствие между синтетическими записями и реальными транзакциями клиентов, (б) восстановить бизнес-логику антифрод-системы банка.

Фоновые знания: информация о типичных паттернах мошенничества из открытых источников (публикации, отчёты FATF, судебные акты), знание общей структуры транзакционных данных банков.

Специфика fraud-данных. В отличие от кейса «Синтетические резюме» (где основной вектор - через квази-идентификаторы), в настоящем кейсе появляется **дополнительный вектор атаки: восстановление бизнес-логики** (Pattern Reconstruction). Если модель, обученная на синтетике, воспроизводит ту же иерархию feature importance, что и модель на реальных данных, происходит утечка коммерчески ценной информации о стратегии обнаружения мошенничества.

Обоснование выбора атак. Набор атак расширен по сравнению с предыдущими кейсами:

Категория риска (CNIL)	Реализованные тесты
Singling out	Singling Out Risk (k-anonymity анализ)
Linkability	CVPL-атака (контрольная выборка + k-NN)
Inference	Propensity Score MIA (Membership Inference Attack)
Бизнес-логика (новое)	**Pattern Reconstruction Attack**

6.2 CVPL-атака (Copy-Vulnerability by Proximity Linkage)

Атака оценивает, насколько синтетические записи «копируют» реальные.

Методология:

1. Исходный датасет разделяется на контрольную выборку (500 записей) и эталонные записи (остальные).
2. Для каждой контрольной записи вычисляется расстояние до ближайшего соседа в синтетическом датасете: $d_{syn} = min_j \|h_i - s_j\|$.
3. Для той же контрольной записи вычисляется расстояние до ближайшего соседа в эталоне: $d_{real} = min_j \|h_i - r_j\|$.
4. **Риск приватности** - доля контрольных записей, для которых синтетика оказывается существенно ближе, чем для эталонной части:

$$\text{Risk} = \frac{1}{|H|}\sum_{i \in H} \mathbb{1}[d_{\text{syn}}(i) < 0.8 \cdot d_{\text{real}}(i)]$$

5. **CVPL Score** - среднее отношение $d_{real} - d_{syn}$. Значения ≈ 1.0 указывают на отсутствие «утечки»; значения >> 1.0 - на возможное копирование.

6.3 Propensity Score MIA (Атака на вывод членства)

Обоснование выбора метода (изменение по сравнению с кейсами MA и HH). В предыдущих кейсах использовался Random Forest-based MIA с метрикой accuracy. Такой подход даёт потолочный эффект (AUC ≈ 1.0) из-за структурных различий между реальными и синтетическими данными (артефакты дискретизации, хвосты), не связанных с утечкой приватности.

Для настоящего кейса применён Propensity Score MIA:

1. Реальные записи размечаются как класс 1, синтетические - как класс 0.
2. Выбираются 5 признаков с наибольшей дисперсией (подмножество для предотвращения потолочного эффекта).

3. Обучается логистическая регрессия (C=0.1, max_iter=500) на 70% данных, тестируется на 30%.
4. Метрика: **AUC-ROC** (вместо accuracy). AUC ≈ 0.5 - идеальная неразличимость; AUC ≈ 1.0 - тривиальная различимость.

Примечание: Propensity Score MIA даёт более информативный диапазон (0.51–0.93) по сравнению с RF MIA (0.84–0.94 в кейсе HH). Рекомендуется использовать этот метод в последующих кейсах серии.

6.4 Риск выделения

Оценивается доля уникальных комбинаций квази-идентификаторов {HOUR_BIN, MERCHANT_CATEGORY, CHANNEL, IS_INTERNATIONAL}:

$$SO = \frac{|\{q : k(q) = 1\}|}{|Q|}$$

где Q - множество всех наблюдаемых комбинаций, $k(q)$ -- число записей с комбинацией q.

Временные бины: Night [0–5], Morning [6–11], Day [12–17], Evening [18–23].

6.5 Атака реконструкции

Атака реконструкции (Pattern Reconstruction Attack) – техника, специфичная для задач, где бизнес-логика модели представляет коммерческую ценность.

Мотивация: в кейсе Обнаружение Мошенничества основной объект защиты - не только ПДн клиентов, но и бизнес-логика антифрод-системы: какие признаки и в каком порядке модель считает индикаторами мошенничества. Если модель, обученная на синтетических данных, воспроизводит ту же иерархию влияния особенностей, то происходит утечка бизнес-логики.

Методология:

1. Обучить Gradient Boosting Classifier на реальных данных → получить влияние особенностей через значения перестановок (10 повторений, контрольная выборка 20%).
2. Обучить аналогичную модель на синтетических данных → получить влияние особенностей.
3. Вычислить корреляцию Спирмена ρ между двумя векторами FI:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

где d_i - разность рангов i-го признака в двух моделях, n = 13 (число признаков).

Интерпретация: $\rho \to 1$ - полная утечка бизнес-логики; $\rho \to 0$ - логика скрыта; $\rho < 0$ - инверсия (ложные паттерны).

6.6 Результаты стресс-теста

Таблица 8. Результаты аудита приватности (DP-Copula)

ε	Риск приватности	CVPL Score	MIA AUC	SO Syn	Pattern ρ	ε
1.0	0.006	0.657	0.564	0.076	0.078	1.0
2.0	0.008	0.637	0.928	0.104	0.0	2.0
4.0	0.010	0.597	0.922	0.104	0.0	4.0
8.0	0.008	0.562	0.924	0.120	0.0	8.0
16.0	0.012	0.537	0.927	0.120	0.0	16.0

Таблица 9. Результаты аудита приватности (DP-MST)

ε	Риск приватности	CVPL Score	MIA AUC	SO Syn	Pattern ρ	ε
1.0	0.006	0.592	0.742	0.000	0.388	1.0
2.0	0.008	0.607	0.512	0.000	0.0	2.0
4.0	0.004	0.607	0.842	0.043	−0.229	4.0
8.0	0.002	0.545	0.797	0.037	0.0	8.0
16.0	0.002	0.567	0.746	0.019	0.0	16.0

Таблица 10. Результаты аудита приватности (Stratified-DP)

ε	Риск приватности	CVPL Score	MIA AUC	SO Syn	Pattern ρ	ε
1.0	0.006	0.638	0.922	0.126	0.739	1.0
2.0	0.010	0.651	0.882	0.166	0.578	2.0
4.0	0.002	0.637	0.919	0.167	0.783	4.0
8.0	0.004	0.609	0.923	0.150	0.841	8.0
16.0	0.004	0.577	0.922	0.158	0.782	16.0

Интерпретация:

CVPL Risk. Все методы демонстрируют риск приватности в диапазоне 0.002–0.012, значительно ниже целевого порога 0.15. Синтетические данные не являются «копиями» реальных записей. DP-MST при больших ε показывает наименьший Risk (0.002), что объясняется дискретизацией, вносящей дополнительную «размытость».

Propensity Score MIA. DP-Copula при ε=1.0: MIA AUC = 0.564 - наилучшая приватность (наименьшая различимость); при ε≥2.0 MIA стабилизируется на уровне 0.92–0.93. DP-MST: наибольшая вариабельность (0.512–0.842). Stratified-DP: стабильно 0.88–0.92 - ожидаемый результат, поскольку раздельный синтез сохраняет паттерны, повышающие различимость.

Атака выделения (Singling Out). DP-MST показывает наименьший singling out (0–4.3%) из-за дискретизации, разрушающей гранулярность. Stratified-DP: повышенный порог выделения (singling out, до 16.7%) - следствие раздельного синтеза малочисленной fraud-страты. Во всех случаях SO Real = 5.88%.

Атака реконструкции (Pattern Reconstruction). Ключевой результат тестирования: DP-Copula и DP-MST при ε≥2.0: ρ ≈ 0 - fraud-паттерны полностью скрыты (но и utility потеряна). Stratified-DP: ρ = 0.58–0.84 при всех ε - стабильно высокое сохранение бизнес-логики. Это фундаментальный trade-off: невозможно получить полезную антифрод-модель из синтетики, не передав ей информацию о значимых признаках.

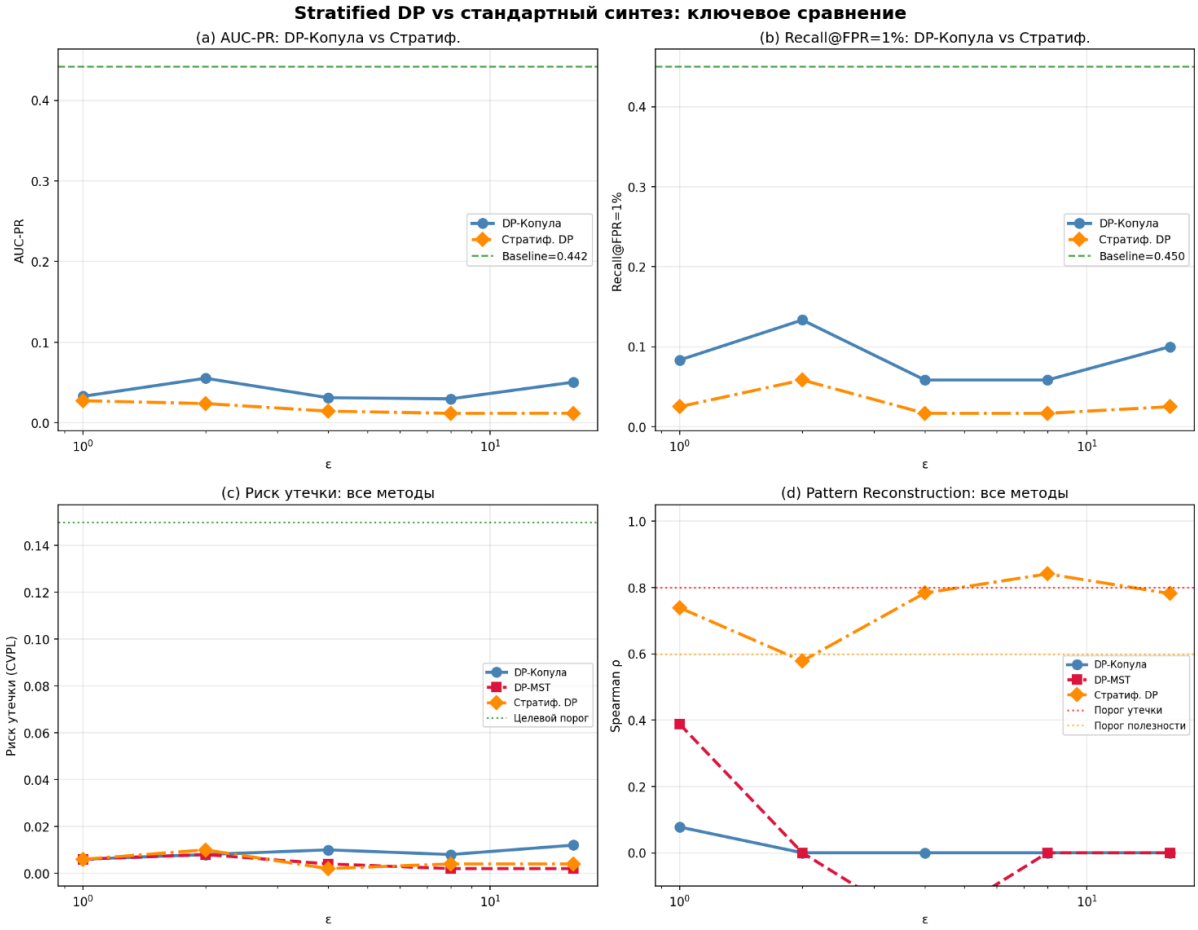

Рисунок 3. Stratified DP vs стандартный синтез: ключевое сравнение (AUC-PR, Recall@FPR=1%, Privacy Risk, Pattern Reconstruction)

6.7 Итоговая оценка приватности

Таблица 11. Сводная оценка приватности по методам и ε

Метод	ε	CVPL Risk	MIA Risk	SO Risk	Pattern Risk
DP-Copula	1.0–4.0	Очень низкий (< 0.01)	Низкий–Высокий	Низкий	Нулевой (ρ ≈ 0)
DP-Copula	8.0–16.0	Очень низкий	Высокий	Низкий	Нулевой
DP-MST	1.0–4.0	Очень низкий	Низкий–Умеренный	Очень низкий	Нулевой–Низкий
DP-MST	8.0–16.0	Очень низкий	Умеренный	Очень низкий	Нулевой

Рекомендуемый диапазон: ε ∈ [4.0, 8.0] для Stratified-DP при внутреннем использовании. Для передачи данных внешним подрядчикам - DP-Copula или DP-MST с ε ≤ 4.0 (нулевой эффект утечки шаблона - pattern leakage). Выбор между приватностью бизнес-логики и полезности (utility) для ANTI-FRAUD - решение бизнеса.

7 Полезность и применение

7.1 Целевая задача: обнаружение мошенничества

Базовой линии (обучение и тест на реальных данных):

Метрика	Значение
AUC-ROC	0.8905
AUC-PR (Average Precision)	0.4417
Recall@FPR=1%	0.45
Precision@top-1%	0.45
F1-score (порог 0.5)	0.40

Таблица 12. Результаты TSTR: обнаружение мошенничества (DP-Copula)

ε	TSTR AUC-ROC	TSTR AUC-PR	TSTR Recall@1%	TSTR Prec@top1%	TSTR F1
1.0	0.6286	0.0328	0.083	0.075	0.0
2.0	0.6607	0.0552	0.133	0.117	0.0
4.0	0.6510	0.0311	0.058	0.050	0.0
8.0	0.6636	0.0297	0.058	0.050	0.0
16.0	0.6625	0.0503	0.100	0.092	0.0

Таблица 13. Результаты TSTR: обнаружение мошенничества (DP-MST)

ε	TSTR AUC-ROC	TSTR AUC-PR	TSTR Recall@1%	TSTR Prec@top1%	TSTR F1
1.0	0.5815	0.0201	0.033	0.033	0.0
2.0	0.4591	0.0093	0.000	0.000	0.0
4.0	0.6723	0.0448	0.083	0.083	0.0
8.0	0.5661	0.0389	0.050	0.050	0.0
16.0	0.6594	0.0364	0.050	0.042	0.0

Таблица 14. Результаты TSTR: обнаружение мошенничества (Stratified-DP)

ε	TSTR AUC-ROC	TSTR AUC-PR	TSTR Recall@1%	TSTR Prec@top1%	TSTR F1
1.0	0.6234	0.0273	0.025	0.025	0.022
2.0	0.5782	0.0238	0.058	0.058	0.022
4.0	0.4472	0.0144	0.017	0.033	0.021
8.0	0.4588	0.0117	0.017	0.017	0.022
16.0	0.4825	0.0118	0.025	0.017	0.022

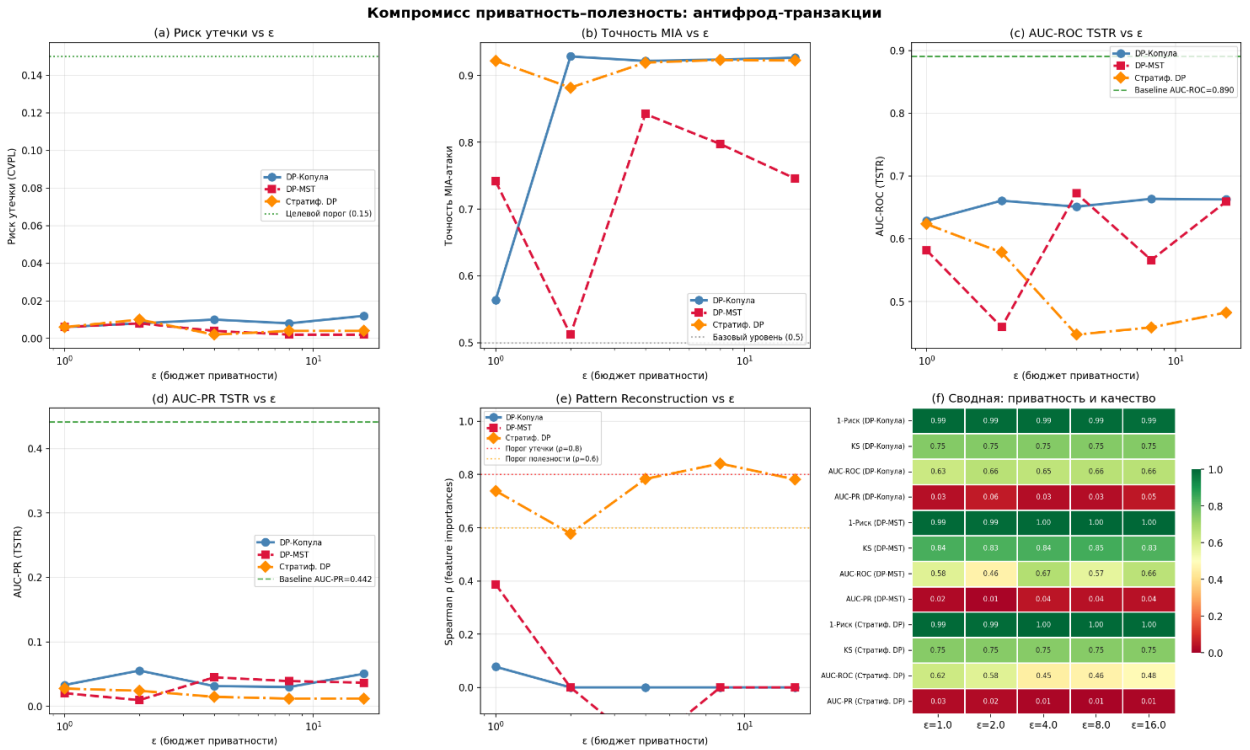

Рисунок 4. Детальные метрики полезности (AUC-PR, Recall@FPR=1%, Feature Importances, F1)

Формулы метрик:

$$AUC\text{-}ROC = \int_0^1 TPR(t)\, d\,FPR(t)$$

$$AUC\text{-}PR = \int_0^1 Precision(r)\, d\,Recall(r)$$

Интерпретация. Все методы TSTR существенно уступают базовой линии по всем метрикам, что ожидаемо при 1% минорных классов и DP-синтезе. Ключевые наблюдения:

- AUC-ROC: максимум 0.6723 (DP-MST, ε=4.0) при базовой линии 0.8905 - утрата ~25% качества ранжирования. Для DP-Copula: 0.6286–0.6636.
- AUC-PR: максимум 0.0552 (DP-Copula, ε=2.0) при базовой линии 0.4417 - утрата ~88% качества. AUC-PR - более чувствительная метрика для ANTI-FRAUD.
- F1 = 0.0 у DP-Copula и DP-MST: при пороге 0.5 модель не генерирует положительных предсказаний - типичная проблема при обучении на синтетике с размытыми fraud-паттернами.

Stratified-DP: единственный метод с F1 > 0 (0.021–0.022) при всех ε, что подтверждает сохранение fraud-паттернов. Однако по AUC-ROC Stratified-DP уступает базовым методам.

7.2 Применимость

Задачи, для которых синтетические данные пригодны:

- бенчмаркинг и сравнение архитектур антифрод-моделей (ранжирование методов сохраняется);
- обучение моделей для предварительной фильтрации (первый каскад антифрод-системы, где допустимо снижение recall);
- A/B тестирование интерфейсов мониторинга с реалистичными транзакциями;
- агрегированная аналитика транзакционных паттернов (распределения сумм, категорий, каналов);
- передача данных внешним подрядчикам для разработки конвейеров обработки.

Задачи с ограниченной применимостью:

- обучение реальных-моделей ANTI-FRAUD (TSTR AUC-PR << базовой линии);
- точная оценка recall на операционных порогах (Recall@FPR=1% существенно занижен);
- анализ редких fraud-паттернов в малых подгруппах (DP-шум маскирует сигнал).

Рекомендуемый ε: 4.0–8.0 для внутренних задач с использованием Stratified-DP; 1.0–4.0 при передаче данных за периметр организации с использованием DP-Copula или DP-MST.

8 Выводы и рекомендации

8.1 Основные результаты

1. **Конвейер обработки работоспособен.** Три метода DP-синтеза (Copula, MST, Stratified) позволяют генерировать синтетические транзакционные данные, сохраняющие основные статистические свойства (KS ≈ 0.75–0.85, TV ≈ 0.70–0.90, Corr Sim ≈ 0.93–0.98).
2. **Экстремальный дисбаланс - фундаментальное ограничение.** При 1% minority class все методы DP-синтеза показывают существенную деградацию AUC-PR (0.01–0.06 при базовой линии 0.44). Это обусловлено тем, что DP-шум имеет амплитуду, сопоставимую с сигналом от 200 fraud-записей.
3. **Подход Stratified DP сохраняет fraud-паттерны.** Единственный метод с ненулевым F1-score (0.021–0.022) и стабильным Pattern Reconstruction ρ = 0.58–0.84. Параллельная композиция обеспечивает формальную DP-гарантию при ε_total = 0.7ε.
4. **Атаки реконструкции выявляют проблемы балансирования приватности и полезности.** Высокая полезность (сохранение fraud-паттернов) неизбежно сопряжена с утечкой feature importance (ρ → 1). Решение о допустимом уровне ρ — это бизнес-решение, а не техническое.
5. **Метрика Propensity Score MIA предпочтительнее RF MIA.** Даёт информативный диапазон (0.51–0.93) без ceiling effect. Рекомендуется для последующих кейсов серии.
6. **CVPL Risk < 0.012 для всех конфигураций.** Синтетические данные не являются «копиями» реальных записей - риск linkage-атаки минимален.

8.2 Ограничения

- Тестирование выполнено на суррогатных данных с заданными корреляциями - реальные транзакционные данные могут содержать более сложные зависимости и выбросы.
- Объём 20 000 записей; масштабирование до 50 000+ может изменить соотношение сигнал/шум.
- Реальная-модель минимальна (GBC, depth=2); в production используются XGBoost, LightGBM, нейросетевые ансамбли.

- Один downstream-task; в production ANTI-FRAUD- каскад правил, скоринговых моделей и нейросетей.
- Фиксированный a=0.7 для Stratified DP не оптимизирован.
- TSTR усреднён по 3 seed - не оценена полная дисперсия метрик.
- DP-CTGAN не включён из-за ограничений среды выполнения.

8.3 Рекомендации

1. **Для** ANTI-FRAUD **рекомендуется Stratified DP** с a=0.7 и ε ≥ 4.0 при внутреннем использовании. Для передачи за периметр - DP-Copula или DP-MST с ε ≤ 4.0.
2. **Pattern Reconstruction Attack должен быть обязательной частью аудита** для любых задач с minority class, где бизнес-логика модели представляет коммерческую ценность.
3. **Повторение на реальных данных.** Выполнить эксперименты на подмножестве реальных транзакционных данных (при наличии организационных мер) для валидации результатов.
4. **Расширение методов.** Включить DP-CTGAN и AIM (Adaptive and Iterative Mechanism); исследовать adaptive allocation a на основе метрик.
5. **Расширение атак.** Добавить attribute inference attacks (предсказание IS_FRAUD, зная остальные атрибуты) и linkage-атаки с внешними источниками (публичные реестры транзакций).
6. **Множественные запуски.** Выполнить N ≥ 5 запусков для оценки дисперсии метрик и построения доверительных интервалов.
7. **Оптимизация a.** Исследовать чувствительность Stratified DP к параметру a на сетке [0.5, 0.6, 0.7, 0.8, 0.9].

9 Паспорт эксперимента

9.1 Идентификация

Поле	Значение
Идентификатор	ABD-FRAUD-DPSYNTH-2025-003
Серия	Приложения к ПНСТ по синтезу данных
Дата проведения	2026-03-XX
Версия конвейера	V2 (pipeline_fraud.py)
Поле	Значение

9.2 Исходные данные

Параметр	Значение
Тип данных	Суррогатные (сгенерированные программно)

Объём	20 000 записей
Число атрибутов	15 (14 после удаления TXN_ID)
Категориальных	2 (MERCHANT_CATEGORY, CHANNEL)
Числовых / бинарных	12
Seed генерации	42
Доля fraud	~1% (200 из 20 000)

9.3 Параметры модели синтеза

DP-Copula:

Параметр	**Значение**
Распределение ε	40% маргинали, 60% корреляции
Число бинов маргиналей	30 (числовые), до 50 (категориальные)
Механизм шума (маргинали)	Laplace
Механизм шума (корреляции)	Gaussian
δ	10^{-5}
PSD-проекция	Спектральное усечение
Пост-коррекция IS_FRAUD	Да (пересемплирование из DP-оценки fraud_rate)
Base seed	100

DP-MST:

Параметр	**Значение**
Распределение ε	1/3 MI + 1/3 root + 1/3 conditional
Число бинов дискретизации	15 (квантильные)
Механизм шума	Laplace
Обратная трансформация	Равномерное семплирование внутри бина
Пост-коррекция IS_FRAUD	Да (пересемплирование из DP-оценки fraud_rate)
Base seed	200

Stratified DP:

Параметр	**Значение**
Метод внутри страт	DP-Copula
α (доля бюджета на fraud)	0.7
ε_fraud	0.7ε
ε_legit	0.3ε
Гарантия	Parallel composition: ε_total = max(ε_fraud, ε_legit) = 0.7ε
Пост-коррекция IS_FRAUD	Не требуется (метки фиксированы по стратам)
Base seed	300

9.4 Параметры аудита

Тест	**Параметр**	**Значение**
CVPL	Размер holdout	500
CVPL	Метод NN	BallTree, k=1
CVPL	Порог proximity	0.8

MIA	Размер выборки	1 000 (real) + 1 000 (syn)
MIA	Классификатор	LogisticRegression, C=0.1, max_iter=500
MIA	Подмножество признаков	Top-5 по дисперсии
MIA	Метрика	AUC-ROC (не accuracy)
MIA	Тест/обучение	70% / 30%
Singling Out	Квази-идентификаторы	HOUR_BIN, MERCHANT_CATEGORY, CHANNEL, IS_INTERNATIONAL
Singling Out	Временные бины	Night [0–5], Morning [6–11], Day [12–17], Evening [18–23]
Pattern Reconstruction	Модель	GBC (n_est=100, depth=4, lr=0.1)
Pattern Reconstruction	Feature Importance	Permutation-based, 10 repeats
Pattern Reconstruction	Метрика	Spearman ρ

9.5 Целевая задача

Параметр	**Значение**
Задача	Бинарная классификация: обнаружение IS_FRAUD
Модель	GradientBoostingClassifier
Гиперпараметры	60 деревьев, max_depth=2, lr=0.08, subsample=0.5, max_features=0.7, min_samples_leaf=25
Признаки	13 (все кроме TXN_ID и IS_FRAUD)
Целевая переменная	IS_FRAUD
Тест/обучение	80% / 20%, stratify по IS_FRAUD
Протокол	TSTR (усреднение по 3 seed: 42, 43, 44)
Метрики	AUC-ROC, AUC-PR, Recall@FPR=1%, Precision@top-1%, F1

9.6 Итоговые бюджеты приватности

Этап	**Бюджет**	**Примечание**
Синтез - DP-Copula, DP-MST (M3)	$\varepsilon \in \{1.0, 2.0, 4.0, 8.0, 16.0\}$	Sequential: $\varepsilon_{total} = \varepsilon_{synth}$
Синтез - Stratified-DP (M3)	$\varepsilon \in \{1.0, 2.0, 4.0, 8.0, 16.0\}$	Parallel: $\varepsilon_{total} = 0.7\varepsilon$