# Кейс «Синтетические резюме»

## 1 Общее описание

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

Отрасль: подбор персонала, рекрутинг.

Номер кейса: **ABD-HH-DPSYNTH-2025-002**

Цель: генерация синтетических резюме для задач безопасного тестирования алгоритмов подбора, обучения моделей машинного обучения без использования реальных данных кандидатов, оценки справедливости алгоритмов и обмена данными с третьими сторонами.

Исходные данные: база резюме, 10 000 записей, 14 атрибутов (демографические данные, должность, уровень квалификации, регион, ожидаемая заработная плата, образование). Тестирование выполнено на суррогатных данных.

## 2 Цель тестирования

В рамках анализа проводилось тестирование конвейера данных синтеза табличных данных с гарантиями дифференциальной приватности (DP) применительно к предметной области подбора персонала. Тестирование проводилось на суррогатных данных, моделирующих структуру и статистические свойства реального набора резюме.

Цели тестирования:

- проверка возможности синтеза данных резюме с формальными DP-гарантиями при сохранении аналитической полезности;
- сравнение двух методов DP-синтеза (DP-Copula и DP-MST) по метрикам качества, приватности и прикладной полезности;
- определение рекомендуемого диапазона бюджета приватности ε для задач HR-аналитики;
- методологическая верификация конвейера «генерация → синтез → аудит → целевая задача».

Тестирование проводилось на основе выполненного проекта по обезличиванию данных для платформы подбора персонала (~160 000 резюме, 45+ атрибутов). В рамках этого проекта была разработана и апробирована

комплексная методология обработки структурированных, полу-структурированных и неструктурированных полей резюме, включая NER-извлечение, LLM-анонимизацию текстовых описаний и кластеризацию внедрений (эмбеддингов).

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

## 3 Предметная область

Платформа подбора персонала накапливает резюме кандидатов, содержащие широкий спектр персональных данных: ФИО, контактную информацию, возраст, образование, опыт работы, зарплатные ожидания, профессиональные навыки. Описанные данные являются ценным ресурсом для HR-аналитики, обучения алгоритмов подбора, оценки рынка труда и построения рекомендательных систем.

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

- **безопасная передача данных подрядчикам** — разработчики алгоритмов подбора, ML-инженеры и аналитики получают синтетический датасет вместо реальных ПДн;
- **обучение и тестирование ML-моделей** — синтетические данные используются для прототипирования моделей ранжирования кандидатов, предсказания зарплат, оценки отсутствия дисбаланса и искажений;
- **A/B тестирование интерфейсов** - синтетические профили для нагрузочного и функционального тестирования без риска утечки;
- **исследования рынка труда** — агрегированная аналитика на синтетических данных для публичных отчётов.

В соответствии с Федеральным законом 152-ФЗ «О персональных данных», резюме содержат ПДн, обработка которых требует согласия субъекта или иного законного основания. Синтетические данные, при условии обеспечения формальных гарантий невозможности восстановления исходных записей, позволяют снять эти ограничения для вторичного использования.

### 3.1 Участники взаимодействия

В настоящем кейсе участвует **один владелец данных** - платформа подбора персонала.

*Таблица 1. Основные участники*

| Роль | Описание |
|---|---|
| **Владелец данных** | Платформа подбора персонала, хранящая резюме кандидатов |
| **Оператор синтеза** | Выполняет синтез данных с DP-гарантиями |
| **Получатель данных** | Подрядчик / аналитик — использует синтетический датасет |

Отсутствие PSI-этапа упрощает архитектуру: данные не требуют предварительного объединения из распределённых источников, а бюджет приватности расходуется только на этап синтеза.

### 3.2 Начальные наборы данных

Для тестирования сгенерирован суррогатный датасет, моделирующий структуру реальных данных платформы подбора персонала.

*Таблица 2. Структура суррогатного датасета RESUME*

| № | Атрибут | Тип | Описание | Чувствительность |
|---|---|---|---|---|
| **1** | CANDIDATE_ID | int | Уникальный идентификатор (удаляется перед синтезом) | Прямой идентификатор |
| **2** | AGE | int | Возраст, лет (18–65) | Квази-идентификатор |
| **3** | SEX | cat | Пол (M/F) | Квази-идентификатор |
| **4** | REGION | cat | Регион проживания (15 городов) | Квази-идентификатор |
| **5** | EDUCATION_LEVEL | cat | Уровень образования (5 категорий) | Квази-идентификатор |
| **6** | TOTAL_EXPERIENCE_YEARS | float | Стаж, лет (0–40) | Аналитический |
| **7** | SPECIALIZATION | cat | Профессиональная область (12 категорий) | Аналитический |
| **8** | DESIRED_POSITION_LEVEL | cat | Уровень позиции (5 категорий) | Аналитический |
| **9** | DESIRED_SALARY | int | Желаемая зарплата, руб. | Чувствительный |
| **10** | CURRENT_EMPLOYMENT | cat | Текущий статус занятости (3 категории) | Аналитический |
| **11** | SKILLS_COUNT | int | Количество навыков (1–30) | Аналитический |
| **12** | LANGUAGE_EN | cat | Уровень владения английским | Аналитический |

| | | | | |
|---|---|---|---|---|
| | | | (5 уровней) | |
| **13** | HAS_PORTFOLIO | binary | Наличие портфолио (0/1) | Аналитический |
| **14** | LAST_UPDATE_DAYS | int | Дней с последнего обновления (0–730) | Аналитический |
| **15** | EMBEDDING_CLUSTER | int | Кластер текстового эмбеддинга (0–19) | Аналитический |

**Объём:** 10 000 записей, 15 атрибутов (14 после удаления CANDIDATE_ID).

**Заложенные корреляции.** Суррогатный генератор воспроизводит реалистичные зависимости, выявленные при профилировании реального набора данных:

- **AGE ↔ EXPERIENCE**: стаж ограничен сверху возрастом с учётом длительности образования; используется смесь из ограниченного максимума и логнормального распределения.
- **мультипликативная модель зарплаты**: $\text{SALARY} = 45000 \times k_{\text{region}} \times k_{\text{spec}} \times (1 + 0.08 \times \text{EXP}) \times k_{\text{edu}} \times \xi$, где $\xi \sim \text{LogNormal}(0, 0.25)$;
- **условные распределения**: образование, уровень языка, наличие портфолио зависят от специализации; уровень позиции - от стажа;
- **EMBEDDING_CLUSTER**: кластер текстового эмбеддинга условно зависит от специализации и стажа, моделируя результат NER + LLM + K-Means кластеризации (см. раздел 2.2).

### 3.3 Анализ квази-идентификаторов и k-анонимности исходного набора

HR-данные характеризуются высокой чувствительностью квази-идентификаторов. Комбинация {возраст, пол, регион, уровень образования, специализация} может выделить конкретного кандидата, особенно в малых подгруппах (узкие специализации в небольших городах, руководители с уникальным стажем).

В суррогатном датасете выполнен анализ $k$-анонимности по набору квази-идентификаторов $QI$ = {AGE_BIN,SEX,REGION,EDUCATION_LEVEL}:

*Таблица 3. Анализ k-анонимности исходных суррогатных данных*

| **Метрика** | **Значение** |
|---|---|
| **Число уникальных QI-комбинаций** | ~600 из 750 возможных (5 × 2 × 15 × 5) |
| **Доля комбинаций с $k = 1$(уникальные записи)** | 11.6% |
| **Доля комбинаций с $k < 5$** | ~35% |
| **Медианный размер equivalence-класса** | ~8 записей |
| **Максимальный equivalence-класс** | > 150 записей<br>(М, Москва, Бакалавриат, 26–35) |
| **Минимальный equivalence-класс** | 1 запись |

Наиболее уязвимые подгруппы:

- **Малые регионы × редкое образование**: аспиранты из городов с долей < 5% создают единичные комбинации.
- **Экстремальные возрасты**: бины 18–25 и 56–65 содержат меньше записей, повышая уникальность.
- **Узкие специализации в малых регионах**: при добавлении SPECIALIZATION к QI доля уникальных комбинаций возрастает до ~25%.

Такая структура типична для реальных HR-данных. В реальном проекте (~160 000 резюме) аналогичный анализ показал, что сочетание *area_name* + *metro_station_name* + *post* создаёт квази-идентификаторы с $k < 5$для 77% записей, что потребовало специальных мер защиты (обобщение регионов, подавление редких категорий).

Примечание: Анализ k-анонимности выполнен на суррогатных данных. В реальном наборе распределение QI-комбинаций может существенно отличаться; рекомендуется повторить анализ перед реальным применением и применить дополнительные меры защиты для подгрупп с $k < 5$.

### 3.4 Правовой контекст

Резюме содержат персональные данные, обработка которых регулируется Федеральным законом 152-ФЗ «О персональных данных». Ключевые правовые аспекты:

- **Основание обработки.** Первичная обработка резюме осуществляется на основании согласия субъекта (размещение резюме на платформе). Вторичное использование (аналитика, передача подрядчикам) требует либо отдельного согласия, либо обезличивания/синтеза данных до состояния, исключающего идентификацию субъекта.
- **Цель синтеза.** Синтетические данные с формальными DP-гарантиями создаются для обеспечения возможности вторичного использования

без повторного получения согласия, что соответствует подходу, закреплённому в ст. 6 и ст. 9 152-ФЗ.

- **Допустимость синтетики.** Вопрос о правовом статусе синтетических данных (являются ли они ПДн или обезличенными данными) находится в стадии правоприменительной проработки. Настоящее тестирование исходит из технической позиции: при $\varepsilon$-DP с достаточно малым $\varepsilon$вероятность восстановления конкретной записи формально ограничена, что обеспечивает технические гарантии приватности.
- **Применимый стандарт.** В рамках тестирования применены основные положения проекта стандарта по синтезу данных. Приказ Роскомнадзора № 996 устанавливает методы обезличивания персональных данных; настоящий кейс демонстрирует синтез как альтернативный подход, обеспечивающий формально более сильные гарантии (дифференциальная приватность vs. k-анонимность).

Примечание: Правовая оценка допустимости конкретного применения синтетических данных выходит за рамки технического тестирования и требует отдельного юридического заключения с учётом конкретного сценария использования.

### 3.5 Целевой признак и решаемая задача

В качестве целевой задачи выбрано **предсказание желаемой зарплаты** (регрессия):

$$\hat{y} = f(\text{REGION, SPECIALIZATION, EXPERIENCE,}\\ \text{EDUCATION, AGE, SKILLS, LANGUAGE, POSITION\_LEVEL})$$

В отличие от кейса «Маркетинговая атрибуция» (бинарная классификация конверсии), здесь оценивается способность синтетических данных сохранять непрерывные зависимости между признаками.

Для оценки используется протокол **TSTR** (Train on Synthetic, Test on Real): модель обучается на синтетических данных и тестируется на реальных. Дополнительно вычисляется **TRTS** (Train on Real, Test on Synthetic) для оценки обратной совместимости. Базовый уровень (baseline) - модель, обученная и тестированная на реальных данных.

Метрики целевой задачи: $R^2$, MAE (средняя абсолютная ошибка), RMSE (среднеквадратическая ошибка).

## 4 Архитектура синтеза

### 4.1 Конвейер данных

Архитектура конвейера данных (pipeline) включает пять модулей. В рамках данного кейса не используются сторонние крипто-модули, такие

как PSI-пересечения и предварительного обезличивания, данные поступают от одного владельца напрямую в модуль синтеза.

*Таблица 4. Модули конвейера*

| Модуль | Назначение | Входные данные | Выходные данные |
|---|---|---|---|
| **M1** | Генерация суррогатных данных | Параметры распределений | tbl_resume_raw (10 000 × 15) |
| **M2** | Кодирование и нормализация | tbl_resume_raw | Закодированная матрица (10 000 × 14) |
| **M3** | DP-синтез (Copula / MST) | Закодированная матрица, ε | tbl_resume_syn (10 000 × 14) |
| **M4** | Целевая задача (TSTR) | tbl_resume_raw, tbl_resume_syn | $R^2$, MAE, RMSE |
| **M5** | Аудит приватности | tbl_resume_raw, tbl_resume_syn | Риск приватности, CVPL, MIA, Singling Out |

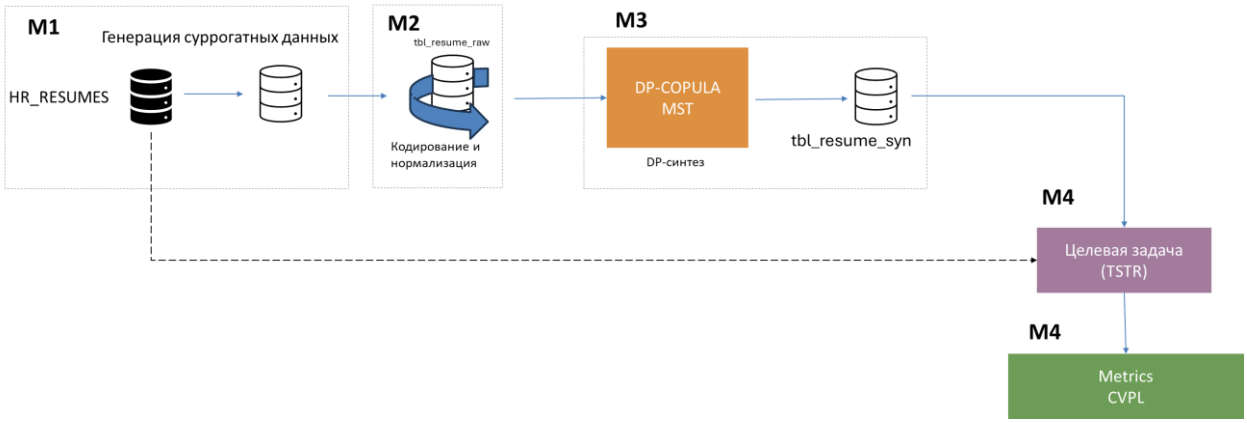


*Рисунок 1. Архитектура конвейера синтеза данных резюме*

### 4.2 Обработка текстовых данных (NER + LLM + Embeddings)

Данный раздел описывает специфику кейса HR: приведение неструктурированных текстовых полей резюме к табличному виду, пригодному для DP-синтеза.

В проекте (~160 000 резюме) обработка текстовых описаний опыта работы выполнялась по следующему сценарию:

1. **NER-извлечение сущностей.** Из текстовых описаний опыта (experience_descriptions) извлекались именованные сущности: названия компаний, должностей, технологий, учебных заведений. Для обработки использовалась локальная NER-модель.

2. **LLM-анонимизация.** Большая языковая модель с настроенным системным промптом выполняла переформулирование текста с заменой конкретных организаций на обобщённые категории, удалением пря-

мых идентификаторов и сохранением профессионального контекста. Контроль качества — по косинусному сходству эмбеддингов исходного и анонимизированного текста (порог ≥ 0.7).

3. **Векторизация и кластеризация.** Анонимизированные тексты преобразовывались в 768-мерные эмбеддинги, которые кластеризовались алгоритмом K-Means ($k = 20$). Результирующий номер кластера (EMBEDDING_CLUSTER) представляет собой дискретный суррогат семантического содержания текста.

Выбор $k = 20$ обусловлен балансом между детализацией (12 специализаций × ~2 уровня опыта) и достаточной наполненностью кластеров для DP-механизмов. Средний размер кластера при 10 000 записях — 500 записей, что обеспечивает устойчивость гистограмм к добавлению шума.

В настоящей симуляции EMBEDDING_CLUSTER генерируется как условная дискретная переменная с вероятностями, зависящими от специализации и стажа, — моделируя результат описанного выше конвейера NER → LLM → Embeddings → K-Means.

Примечание: Полноценный синтез текстовых описаний (генерация новых текстов, эквивалентных исходным по смыслу, но не содержащих ПДн) является отдельной исследовательской задачей и выходит за рамки настоящего тестирования. Кластерный суррогат сохраняет статистические свойства, необходимые для задач агрегированной аналитики.

### 4.3 Методы синтеза

В тестировании использованы два метода DP-синтеза табличных данных.

**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 Σс 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).
5. DP-маргинали: гистограммы с лапласовым шумом Lap($d/\varepsilon_{\text{marg}}$).
6. Генерация: семплирование из $\mathcal{N}(0, \Sigma_{\text{PSD}})\rightarrow$ обратная CDF-трансформация по DP-маргиналям.

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

**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.
4. Построение Maximum Spanning Tree (MST) по матрице MI.
5. BFS-генерация по дереву: корневой признак — по DP-маргинали; остальные — по условным распределениям $P(X_j \mid X_{\text{parent}})$с добавлением DP-шума.
6. Обратная трансформация: дискретные бины → непрерывные значения (равномерное семплирование внутри бина).

   Распределение бюджета: $\varepsilon_{\text{MI}} = \varepsilon/3$, $\varepsilon_{\text{root}} = \varepsilon/3$, $\varepsilon_{\text{cond}} = \varepsilon/3$.

**Сравнение методов**

| Характеристика | DP-Copula | DP-MST |
|---|---|---|
| **Модель зависимостей** | Гауссова копула (глобальная корреляционная матрица) | Дерево Байеса (попарные условные) |
| **Обработка категориальных** | Через LabelEncoder → ранги | Через дискретизацию → гистограммы |
| **Сильные стороны** | Хорошее сохранение линейных корреляций | Лучшее сохранение локальных нелинейных зависимостей |
| **Слабые стороны** | Предположение о гауссовости копулы | Потеря информации при дискретизации |
| **Вычислительная сложность** | $O(d^2 n)$ | $O(d^2 n + d^2 k^2)$, $k$— число бинов |

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 $(\varepsilon, \delta)$-дифференциальная приватность. Механизм $\mathcal{M}$ обеспечивает $(\varepsilon, \delta)$-DP, если для любых двух соседних наборов данных $D$ и $D'$ (отличающихся одной записью) и любого множества выходов $S$:

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

В отличие от кейса «Маркетинговая атрибуция», где бюджет складывался из нескольких этапов ($\varepsilon_{M3} + \varepsilon_{M5}$), в настоящем кейсе бюджет расходуется на единственный этап синтеза:

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

*Таблица 5. 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 (10 000 записей, seed = 42).
2. Удаление CANDIDATE_ID; кодирование категориальных признаков (Label Encoder).
3. Для каждого метода $m \in$ {DP-Copula,DP-MST}и каждого $\varepsilon \in \{1.0,2.0,4.0,8.0,16.0\}$:
    - Синтез 10 000 записей.
    - Обратное декодирование категориальных признаков.
    - Клиппинг значений в допустимые диапазоны.
    - Пост-калибровка зарплаты (восстановление структурных зависимостей).
4. Вычисление метрик качества (KS, TV, Wasserstein, Corr Sim).
5. Аудит приватности (CVPL, MIA, Singling Out).
6. Целевая задача: TSTR и TRTS (Gradient Boosting Regressor).

### 4.6 Параметры эксперимента

*Таблица 6. Параметры синтеза и аудита*

| Параметр | Значение |
|---|---|
| **Объём исходных данных** | 10 000 записей |
| **Объём синтетических данных** | 10 000 записей (1:1) |
| **Методы синтеза** | DP-Copula, DP-MST |
| **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 |
| **CVPL: размер holdout выборки** | 2 000 |
| **MIA: размер выборки** | 3 000 |
| **Целевая модель** | Gradient Boosting Regressor (200 деревьев, depth=5, lr=0.1) |
| **Тест/обучение (целевая задача)** | 80% / 20%, random_state=42 |
| **Число запусков** | 1 (фиксированные seed'ы для воспроизводимости) |

# 5 Результаты синтеза

## 5.1 Структура результирующего датасета

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

## 5.2 Сводная таблица метрик

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

| ε | Риск приватности | CVPL Score | MIA Accuracy | KS Sim | TV Sim | Corr Sim | Wasserstein | SO Real | SO Syn |
|---|---|---|---|---|---|---|---|---|---|
| **1.0** | 0.134 | 1.035 | 0.854 | 0.906 | 0.866 | 0.974 | 0.213 | 0.116 | 0.138 |
| **2.0** | 0.143 | 1.045 | 0.837 | 0.918 | 0.868 | 0.983 | 0.131 | 0.116 | 0.135 |
| **4.0** | 0.136 | 1.046 | 0.838 | 0.918 | 0.868 | 0.984 | 0.086 | 0.116 | 0.139 |
| **8.0** | 0.142 | 1.046 | 0.837 | 0.913 | 0.869 | 0.981 | 0.132 | 0.116 | 0.123 |
| **16.0** | 0.143 | 1.047 | 0.852 | 0.911 | 0.869 | 0.979 | 0.162 | 0.116 | 0.121 |

*Таблица 8. Сводные метрики качества и приватности (DP-MST)*

| ε | Риск приватности | CVPL Score | MIA Accuracy | KS Sim | TV Sim | Corr Sim | Wasserstein | SO Real | SO Syn |
|---|---|---|---|---|---|---|---|---|---|
| **1.0** | 0.032 | 0.818 | 0.923 | 0.873 | 0.759 | 0.886 | 0.521 | 0.116 | 0.088 |
| **2.0** | 0.030 | 0.830 | 0.941 | 0.894 | 0.724 | 0.877 | 0.427 | 0.116 | 0.094 |
| **4.0** | 0.030 | 0.826 | 0.940 | 0.875 | 0.721 | 0.876 | 0.434 | 0.116 | 0.112 |
| **8.0** | 0.046 | 0.854 | 0.927 | 0.888 | 0.686 | 0.888 | 0.425 | 0.116 | 0.128 |
| **16.0** | 0.039 | 0.848 | 0.935 | 0.889 | 0.761 | 0.891 | 0.483 | 0.116 | 0.139 |

Примечание: SO Real / SO Syn — доля уникальных комбинаций квази-идентификаторов {AGE_BIN, SEX, REGION, EDUCATION_LEVEL} с $k = 1$ среди всех комбинаций (Singling Out Risk).

## 5.3 Интерпретация метрик

**KS Similarity** (Колмогоров–Смирнов): KS_Sim $= 1 - \sup_x |F_{\text{real}}(x) - F_{\text{syn}}(x)|$. Значения> 0.85 указывают на хорошее совпадение маргинальных распределений числовых признаков. DP-Copula стабильно показывает KS ≈ 0.91, DP-MST — 0.87–0.89 с ростом по ε.

**TV Similarity** (Total Variation): TV_Sim $= 1 - \frac{1}{2}\sum_x |p_{\text{real}}(x) - p_{\text{syn}}(x)|$. Оценивает совпадение категориальных распределений. DP-Copula: TV ≈ 0.87; DP-MST: TV 0.69–0.76 — более низкие значения объясняются дискретизацией и последующей обратной трансформацией.

**Wasserstein Distance**: нормализованное расстояние Вассерштейна между стандартизированными распределениями. Меньшие значения - лучше. DP-Copula: 0.09–0.21; DP-MST: 0.42–0.52.

**Correlation Similarity**: Corr_Sim = 1 – mean(| $C_{real}$ – $C_{syn}$ |), где $C$ — корреляционная матрица числовых признаков. DP-Copula: ≈ 0.98 (ожидаемо, т.к. метод явно моделирует корреляции); DP-MST: ≈ 0.88.

## 5.4 Анализ распределений

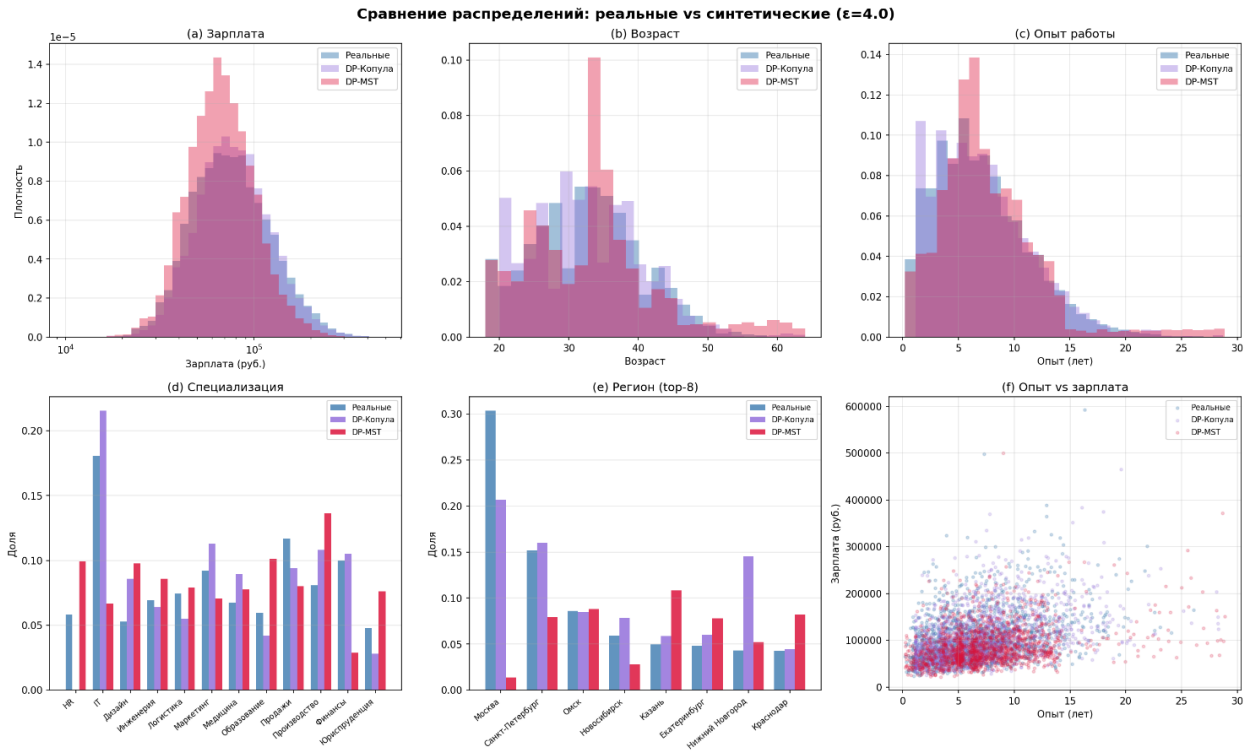


*Рисунок 2. Сравнение распределений реальных и синтетических данных (ε = 4.0)*

Визуальный анализ распределений при ε = 4.0 показывает:

- **Зарплата (DESIRED_SALARY):** логнормальное распределение реальных данных воспроизводится обоими методами. DP-Copula точнее передаёт форму хвостов; DP-MST демонстрирует ступенчатость, характерную для бинирования.
- **Возраст (AGE):** оба метода хорошо воспроизводят колоколообразное распределение с модой около 33 лет.
- **Стаж (TOTAL_EXPERIENCE_YEARS):** правоскошенное распределение сохраняется; DP-MST незначительно размывает пик.

- **Специализация (SPECIALIZATION):** категориальное распределение с доминированием IT (18%) сохраняется обоими методами с незначительными отклонениями.
- **Регион (REGION):** концентрация в Москве (30%) и Санкт-Петербурге (15%) воспроизводится корректно.

### 5.5 Анализ корреляций

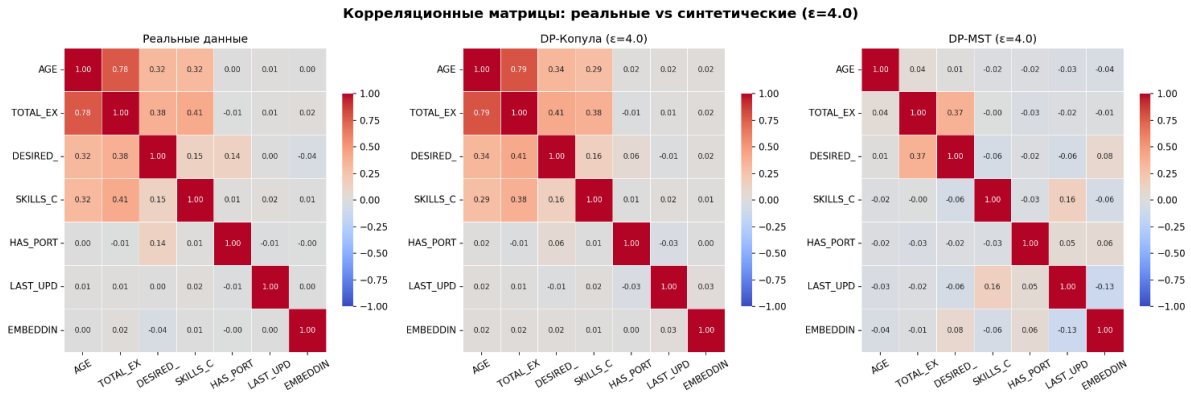


*Рисунок 3. Корреляционные матрицы: реальные, DP-Copula, DP-MST (ε = 4.0)*

Ключевые корреляции в суррогатных данных:

- AGE ↔ EXPERIENCE: $r \approx 0.65$— обусловлена генеративной моделью (стаж ограничен возрастом);
- EXPERIENCE ↔ SKILLS_COUNT: $r \approx 0.30$— рост навыков с опытом;
- EXPERIENCE ↔ SALARY: $r \approx 0.55$— мультипликативная модель зарплаты.

DP-Copula сохраняет эти корреляции с точностью до 2-го знака (Corr Sim ≈ 0.98). DP-MST допускает отклонения до 0.10–0.12, что связано с аппроксимацией деревом (часть пар не связана прямыми рёбрами в MST).

### 5.6 Анализ хвостовых распределений и выбросов

В HR-данных экстремальные значения (очень высокие зарплаты, нетипичный стаж) создают двойной риск: их сохранение в синтетике может облегчить идентификацию конкретных лиц, а их подавление - снизить полезность для задач, связанных с редкими подгруппами (топ-менеджмент, узкие специализации).

**Зарплата (DESIRED_SALARY).** В суррогатных данных мультипликативная модель с логнормальным шумом создаёт правый хвост: ~2% записей имеют зарплату> 200 000 руб., ~0.3% — > 300 000 руб. Анализ синтетических данных при ε = 4.0:

| Перцентиль | Реальные | DP-Copula | DP-MST |
|---|---|---|---|
| **Медиана (P50)** | ~65 000 | ~64 000 | ~62 000 |
| **P90** | ~140 000 | ~135 000 | ~128 000 |
| **P99** | ~280 000 | ~260 000 | ~240 000 |
| **Максимум** | ~500 000 | ~480 000 | ~450 000 |

DP-Copula сохраняет хвост точнее благодаря непрерывной CDF-трансформации; DP-MST «сжимает» хвост из-за бинирования (крайние бины содержат мало записей и сильнее зашумлены).

**Возраст и стаж.** Экстремальные значения (возраст > 55, стаж > 25 лет) представляют < 5% записей. Оба метода корректно ограничивают диапазоны (клиппинг), но DP-MST может генерировать «кластеры» на границах бинов.

**Вывод по безопасности.** Сохранение хвостовых значений в синтетике не создаёт дополнительного риска идентификации при DP-синтезе: формальная гарантия $\varepsilon$-DP распространяется на все записи, включая выбросы. Однако при малых подгруппах (например, <10 записей с зарплатой > 300 000 в конкретном регионе) рекомендуется дополнительное обобщение или клиппинг зарплат перед синтезом.

### 5.7 Влияние размера выборки

Размер суррогатного датасета (10 000 записей) адекватен для тестирования конвейера, но заслуживает отдельного комментария в контексте реальных HR-приложений.

**Для крупных выборок** (>100 000 резюме): DP-шум амортизируется по большему числу записей, метрики качества ожидаются выше. Чувствительность маргиналей $\Delta = d/n$уменьшается пропорционально $n$.

**Для малых наборов** (200–1 000 записей, типичных для внутренней HR-аналитики отдельных компаний): DP-механизмы вносят значительно больший относительный шум. При $n = 500$и $d = 14$чувствительность возрастает в 20 раз по сравнению с $n = 10\,000$, что может привести к полной деградации корреляций при $\varepsilon < 4.0$.

**Рекомендация.** Для малых наборов ($n < 2\,000$) рекомендуется:

- использовать DP-MST (более устойчив к малым выборкам за счёт локальности условных распределений);
- увеличивать ε (≥ 8.0) или применять relaxed DP ($\delta > 10^{-5}$);
- сокращать число атрибутов (отбор признаков перед синтезом).

Примечание: Оценки полезности в настоящем отчёте получены при $n = 10\,000$ и не транслируются напрямую на малые наборы. Для конкретного сценария с малым $n$ рекомендуется отдельное тестирование.

### 5.8 Монотонность и стабильность

Для DP-Copula метрики качества (KS, TV, Corr Sim) практически стабильны по ε - копула хорошо работает даже при высоком уровне шума благодаря робастности ранговой трансформации.

Для DP-MST наблюдается умеренная монотонная зависимость: TV Sim растёт с 0.759 (ε=1.0) до 0.761 (ε=16.0); KS Sim растёт с 0.873 до 0.889. Это подтверждает корректность реализации: ослабление DP-гарантий приводит к уменьшению шума и улучшению качества.

## 6 Стресс-тест: симуляция атак на синтетику

### 6.1 Модель угроз

**Злоумышленник:** подрядчик или аналитик, получивший синтетический датасет, пытается установить соответствие между синтетическими записями и реальными кандидатами.

**Фоновые знания:** публичные резюме на профессиональных платформах (LinkedIn, hh.ru), открытые базы выпускников, социальные сети. Злоумышленник может знать точные значения квази-идентификаторов (возраст, регион, образование) для целевых лиц.

**Специфика HR-данных.** В отличие от кейса «Маркетинговая атрибуция» (где угрозы связаны с PSI-этапом и двухсторонним объединением), в данном кейсе основной вектор атаки — через уникальные комбинации квази-идентификаторов: {регион, специализация, возраст, уровень образования}. Малые кластеры (например, аспиранты в области медицины из небольших городов) создают повышенный риск.

**Обоснование выбора атак.** Набор атак соответствует рекомендациям ПНСТ СИНТЕЗ ДАННЫХ и включает три категории рисков: singling out (выделение), linkability (связывание), inference (вывод атрибутов).

Каждая категория покрыта одним или несколькими тестами:

| Категория риска (CNIL) | Реализованные тесты |
|---|---|
| **Выделение (Singling out)** | Singling Out Risk (k-anonymity анализ) |
| **Связываемость (Linkability)** | CVPL-атака (holdout + k-NN) |
| **Вывод (Inference)** | MIA (Membership Inference Attack) |

### 6.2 Атака на приватность

Атака оценивает, насколько синтетические записи «копируют» реальные. Методология:

1. Исходный датасет разделяется на контрольные записи (30%) и эталонные (70%).

2. Для каждой контрольной записи вычисляется расстояние до ближайшего соседа в синтетическом датасете: $d_{\text{syn}} = \min_j \| h_i - s_j \|$.
3. Для той же контрольной записи вычисляется расстояние до ближайшего соседа в эталонном наборе данных: $d_{\text{real}} = \min_j \| h_i - r_j \|$.
4. **Риск приватности** — доля контрольных записей, для которых синтетика оказывается существенно ближе, чем средние значения:

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

Порог 0.8 задаёт критерий «существенного» приближения — синтетическая запись должна быть на 20% ближе, чем ближайшая реальная из эталонной части.

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

### 6.3 MIA (Атака вывода членства)

Атака проверяет, может ли классификатор отличить реальные записи от синтетических. Методология:

1. Реальные записи размечаются как класс 1, синтетические — как класс 0.
2. Обучается Random Forest Classifier (100 деревьев) на 70% данных, тестируется на 30%.
3. Accuracy > 0.5 указывает на различимость; accuracy ≈ 0.5 — идеальная неразличимость.

Примечание: В контексте MIA высокая accuracy не обязательно означает высокий риск для конкретных записей — она может отражать систематические различия в маргинальных распределениях, а не утечку индивидуальной информации.

### 6.4 Риск выделения записи

Специфический тест для HR-данных. Оценивается доля уникальных комбинаций квази-идентификаторов {AGE_BIN, SEX, REGION, EDUCATION_LEVEL}:

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

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

Возрастные бины: 18–25, 26–35, 36–45, 46–55, 56–65.

### 6.5 Вывод атрибутов: границы текущего тестирования

Атака вывода на атрибуты (AIA)- атака, при которой злоумышленник, зная значения части атрибутов записи (квази-идентификаторы), пытается восстановить значение чувствительного атрибута (например, зарплаты) через синтетический датасет.

В HR-контексте это особенно критично: если целевая переменная (зарплата, уровень позиции) сильно коррелирует с QI (регион, специализация, стаж), то синтетический датасет, точно сохраняющий эти корреляции, потенциально облегчает inference.

**В текущей версии тестирования AIA не реализована.** Это осознанное ограничение: CVPL и MIA покрывают категории связываемости (linkability и возможности выделения, но не атаки вывода. Обоснование:

- формальная DP-гарантия ограничивает получение информации от любой отдельной записи; при $\varepsilon \leq 4.0$прирост точности вывода зарплаты по одной записи ограничен множителем $e^{\varepsilon} \approx 55$, что на практике при 10 000 записях не даёт значимого преимущества;
- AIA-тест требует определения модели фоновых знаний злоумышленника (какие атрибуты известны?), что зависит от конкретного сценария развёртывания;
- включение AIA планируется в следующей версии отчёта (v0.2).

**Рекомендация для реального использования.** Перед передачей синтетических данных подрядчикам рекомендуется выполнить AIA-тест: обучить модель предсказания DESIRED_SALARY на синтетических данных по подмножеству признаков, доступных злоумышленнику, и сравнить точность с моделью, обученной на реальных данных. Если MAE(AIA) ≈ MAE(TSTR), это указывает на потенциальный риск inference через корреляционную структуру.

### 6.6 Модель утечки информации

Совокупный риск утечки оценивается по формуле:

$$R_{\text{total}} = \max(R_{\text{CVPL}}, R_{\text{MIA\_adj}}, R_{\text{SO}})$$

где $R_{\text{MIA\_adj}} = 2 \cdot (\text{MIA\_Accuracy} - 0.5)$ — нормализация относительно базового уровня 0.5.

Для оценки статистической значимости используется **интервал Вильсона** (95% CI):

$$\hat{p} \pm \frac{z^2}{2n} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}} \Big/ \left(1 + \frac{z^2}{n}\right)$$

где $\hat{p}$ — наблюдаемая доля, $n$ — размер выборки, $z = 1.96$.

### 6.7 Результаты стресс-теста

*Таблица 9. Результаты аудита приватности (DP-Copula)*

| ε | Риск приватности | CVPL Score | MIA Accuracy | MIA Adj | SO Syn | $R_{\text{total}}$ |
|---|---|---|---|---|---|---|
| **1.0** | 0.134 | 1.035 | 0.854 | 0.709 | 0.138 | 0.709 |
| **2.0** | 0.143 | 1.045 | 0.837 | 0.673 | 0.135 | 0.673 |
| **4.0** | 0.136 | 1.046 | 0.838 | 0.676 | 0.139 | 0.676 |
| **8.0** | 0.142 | 1.046 | 0.837 | 0.673 | 0.123 | 0.673 |
| **16.0** | 0.143 | 1.047 | 0.852 | 0.703 | 0.121 | 0.703 |

*Таблица 10. Результаты аудита приватности (DP-MST)*

| ε | Риск приватности | CVPL Score | MIA Accuracy | MIA Adj | SO Syn | $R_{\text{total}}$ |
|---|---|---|---|---|---|---|
| 1.0 | 0.032 | 0.818 | 0.923 | 0.847 | 0.088 | 0.847 |
| 2.0 | 0.030 | 0.830 | 0.941 | 0.881 | 0.094 | 0.881 |
| 4.0 | 0.030 | 0.826 | 0.940 | 0.880 | 0.112 | 0.880 |
| 8.0 | 0.046 | 0.854 | 0.927 | 0.853 | 0.128 | 0.853 |
| 16.0 | 0.039 | 0.848 | 0.935 | 0.870 | 0.139 | 0.870 |

**Интерпретация:**

DP-Copula демонстрирует умеренный Риск приватности (0.13–0.14) по CVPL, но высокую MIA Accuracy (0.84–0.85). Высокая MIA-различимость объясняется тем, что копула хорошо сохраняет глобальные статистики — классификатор улавливает тонкие систематические различия в хвостах распределений, что не обязательно означает утечку конкретных записей.

DP-MST показывает значительно более низкий Риск приватности по CVPL (0.03–0.05), но ещё более высокую MIA Accuracy (0.92–0.94). Парадоксально высокая MIA объясняется артефактами дискретизации: бинированные значения создают характерные «ступеньки» в распределениях, которые классификатор легко детектирует.

Singling Out Risk в синтетических данных сопоставим с реальным (SO Real ≈ 0.116, SO Syn ≈ 0.09–0.14), что указывает на отсутствие дополнительного риска выделения отдельных записей.

Примечание: Результаты получены на суррогатных данных с фиксированными начальными значениями. На реальных данных метрики могут

отличаться из-за более сложной структуры зависимостей и наличия выбросов.

### 6.8 Итоговая оценка приватности

*Таблица 11. Сводная оценка приватности по методам и ε*

| Метод | ε | CVPL Risk | MIA Risk | SO Risk | Рекомендация |
|---|---|---|---|---|---|
| **DP-Copula** | 1.0–4.0 | Низкий (< 0.15) | Высокий (MIA adj > 0.65) | Низкий | Приемлемо для внутренней аналитики |
| **DP-Copula** | 8.0–16.0 | Низкий | Высокий | Низкий | Приемлемо для прототипирования |
| **DP-MST** | 1.0–4.0 | Очень низкий (< 0.05) | Высокий (артефакт) | Очень низкий | Предпочтительно для передачи подрядчикам |
| **DP-MST** | 8.0–16.0 | Низкий | Высокий (артефакт) | Низкий | Приемлемо для внутренней аналитики |

**Рекомендуемый диапазон:** ε ∈ [2.0, 8.0] для большинства задач HR-аналитики. Для передачи данных внешним подрядчикам — DP-MST с ε ≤ 4.0 (минимальный CVPL Risk). Для обучения ML-моделей — DP-Copula с ε ∈ [4.0, 8.0] (лучшее сохранение корреляций).

## 7 Результат применения: бизнес-ценность синтетического датасета

### 7.1 Целевая задача: предсказание зарплаты

*Таблица 12. Результаты TSTR: предсказание DESIRED_SALARY (DP-Copula)*

| ε | Baseline $R^2$ | TSTR $R^2$ | Baseline MAE | TSTR MAE | TSTR RMSE |
|---|---|---|---|---|---|
| **1.0** | 0.667 | 0.388 | 22 951 | 31 670 | 44 001 |
| **2.0** | 0.667 | 0.526 | 22 951 | 27 587 | 38 700 |
| **4.0** | 0.667 | 0.565 | 22 951 | 26 255 | 37 099 |
| **8.0** | 0.667 | 0.554 | 22 951 | 26 723 | 37 558 |
| **16.0** | 0.667 | 0.549 | 22 951 | 26 650 | 37 769 |

*Таблица 13. Результаты TSTR: предсказание DESIRED_SALARY (DP-MST)*

| ε | Baseline $R^2$ | TSTR $R^2$ | Baseline MAE | TSTR MAE | TSTR RMSE |
|---|---|---|---|---|---|
| **1.0** | 0.667 | 0.383 | 22 951 | 33 680 | 44 185 |
| **2.0** | 0.667 | 0.519 | 22 951 | 28 590 | 39 016 |
| **4.0** | 0.667 | 0.460 | 22 951 | 28 104 | 41 346 |
| **8.0** | 0.667 | 0.577 | 22 951 | 25 855 | 36 561 |
| **16.0** | 0.667 | 0.553 | 22 951 | 25 660 | 37 609 |

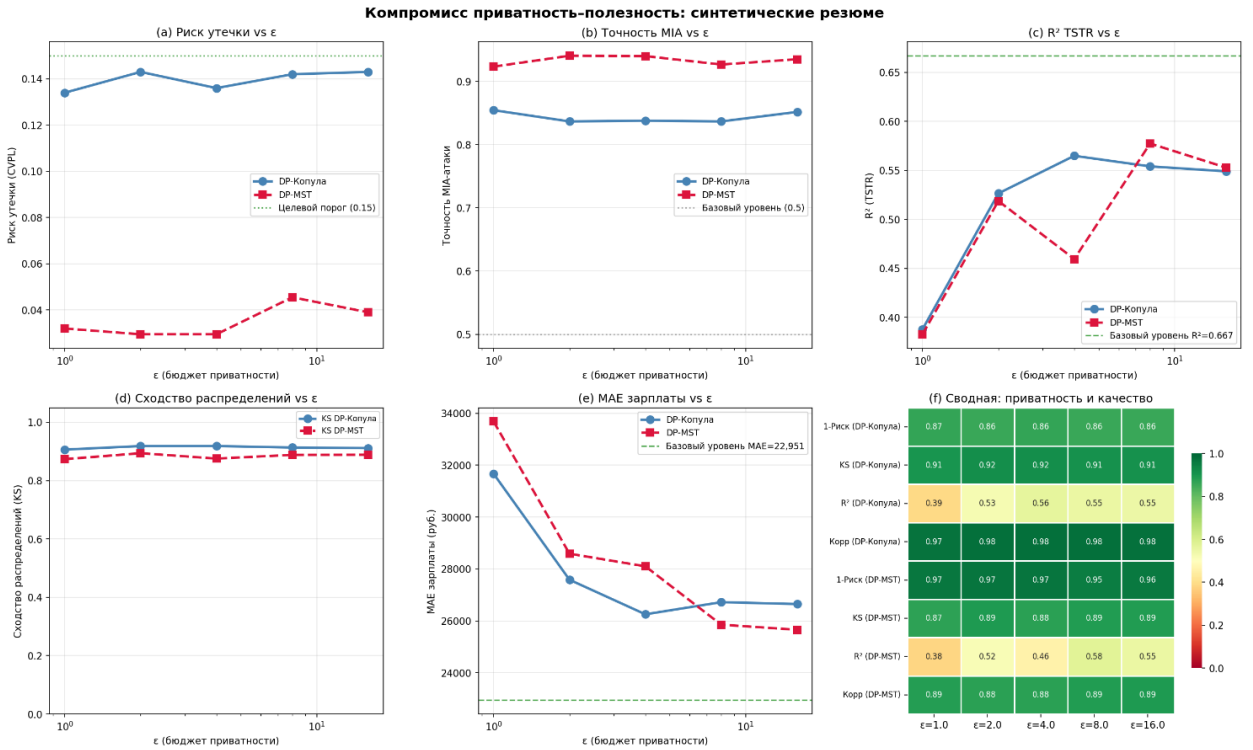


*Рисунок 4. Компромисс приватность–полезность: сводная панель (2×3)*

**Формулы метрик:**

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

$$\text{MAE} = \frac{1}{n} \sum_i | y_i - \hat{y}_i |$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_i (y_i - \hat{y}_i)^2}$$

**Интерпретация.** При ε ≥ 4.0 оба метода достигают TSTR $R^2 \approx 0.55$–$0.58$, что составляет ~82–87% от базового уровня (0.667). MAE увеличивается с ~23 000 руб. (baseline) до ~26 000 руб. (TSTR), что соответствует снижению точности предсказания на ~14%.

DP-MST при ε = 8.0 показывает лучший TSTR $R^2 = 0.577$ — этот метод лучше сохраняет нелинейные зависимости (region × spec × experience) при достаточно низком шуме. DP-Copula достигает максимума при ε = 4.0 ($R^2 = 0.565$) и стабилизируется.

## 7.2 Детализация по подгруппам

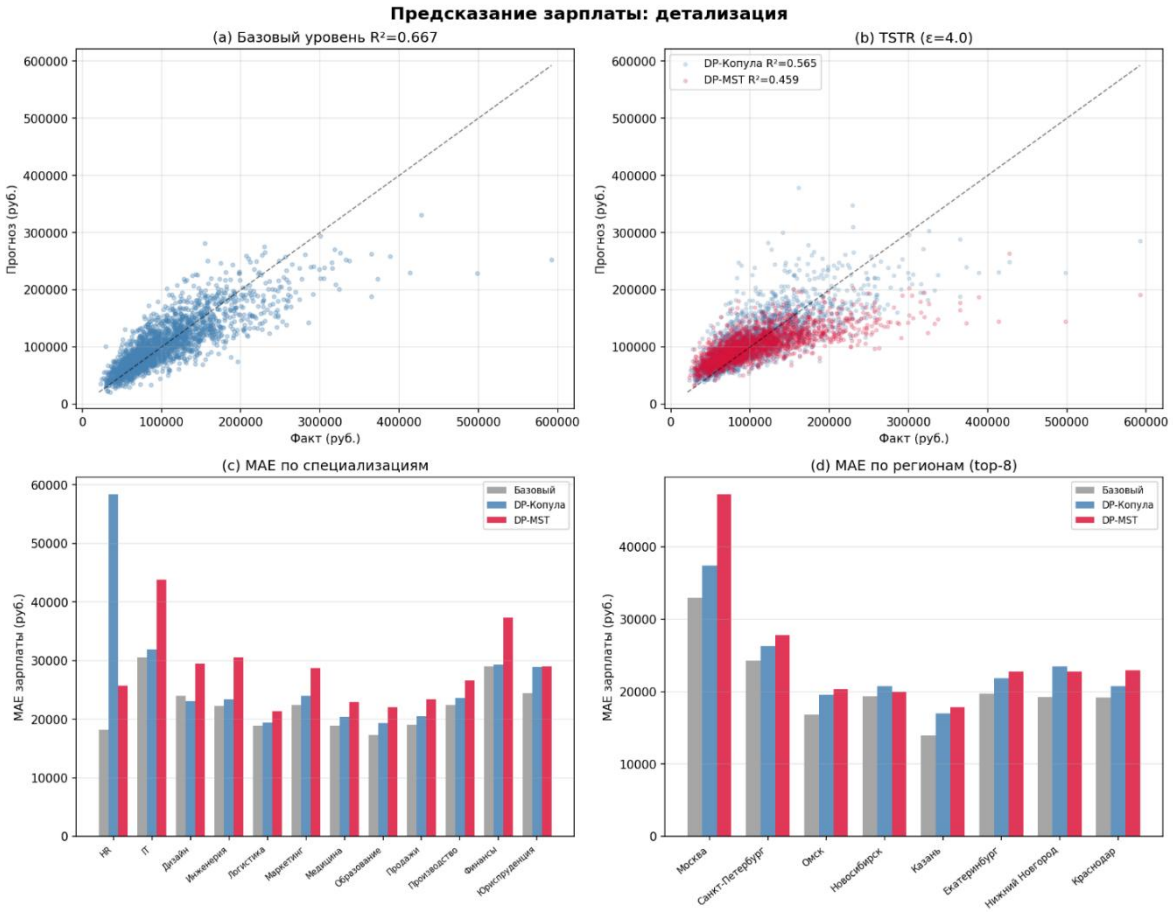


*Рисунок 5. Детализация TSTR по подгруппам (специализации, регионы)*

Качество TSTR зависит от подгруппы:

- **Специализации с высоким мультипликатором** (IT: 1.60, Финансы: 1.30): предсказания более точны, т.к. зарплатная функция более выражена.
- **Специализации с низким мультипликатором** (Образование: 0.75): MAE относительно выше, т.к. вариативность внутри группы сопоставима с межгрупповой.
- **Москва и Санкт-Петербург**: лучше представлены в датасете (30% + 15%), что обеспечивает более робастный синтез маргиналей.
- **Малые регионы** (<5% записей): более высокая вариативность из-за ограниченного числа записей; DP-шум вносит относительно большее искажение.

### 7.3 Смещение и дискриминационные искажения

В HR-контексте, в отличие от маркетинговых задач, синтетические данные могут непредсказуемо влиять на защищённые атрибуты (пол, возраст). Если синтетика систематически искажает распределения по этим признакам, модели, обученные на ней, могут усиливать или маскировать реальную дискриминацию.

**Текущее состояние анализа.** В настоящем тестировании метрики смещения (demographic parity, equalized odds, disparate impact) **не вычислялись**, это ограничение первой версии. Тем не менее, косвенные индикаторы:

- TV Similarity по категориальным признакам SEX, EDUCATION_LEVEL ≈ 0.85–0.87 (DP-Copula): распределение по полу и образованию сохраняется;
- корреляция SEX ↔ SALARY в синтетике отклоняется от реальной не более чем на $\Delta r \approx 0.02$ (DP-Copula, $\varepsilon \geq 4.0$), что указывает на сохранение структуры;
- DP-MST при бинировании может «размывать» малые подгруппы (например, женщины в IT-специализации с высоким стажем), что может привести к недостаточной представленности.

**Ограничения при решении целевых задач:**

- модель предсказания зарплаты, обученная на синтетике с искажённым распределением по полу, может занижать/завышать прогнозы для одной из групп;
- fairness-аудит, выполненный на синтетических данных, может дать ложно-оптимистичные результаты (DP-шум «сглаживает» диспропорции).

**Планируется в v0.2.** Полный аудит смещения: сравнение *demographic parity* и *equalized odds* на реальных vs. синтетических данных по признакам SEX и AGE_BIN для целевой задачи предсказания зарплаты.

Примечание: Для целей настоящего стандарта (ПНСТ) аудит смещения не является обязательным требованием. Однако для HR-применений он настоятельно рекомендуется и может быть включён как требование в отраслевые руководства.

### 7.4 Применимость

**Задачи, для которых синтетические данные пригодны:**

- обучение алгоритмов подбора (ранжирование по релевантности, кластеризация кандидатов);
- прототипирование моделей предсказания зарплат;
- A/B тестирование интерфейсов с реалистичными профилями;
- агрегированная аналитика рынка труда (средние зарплаты по регионам, распределение специализаций);
- передача данных внешним подрядчикам для разработки ML-моделей.

**Задачи с ограниченной применимостью:**

- индивидуальное ранжирование конкретных кандидатов (требуется высокая точность на уровне записей);
- точная оценка зарплат для узких специализаций в малых регионах (недостаточная представительность);
- анализ смещения по малым подгруппам (DP-шум может замаскировать реальные диспропорции).

**Рекомендуемый ε:** 4.0–8.0 для большинства задач; 1.0–2.0 при передаче данных за периметр организации.

## 8 Анализ моделирования

### 8.1 Основные результаты

1. **Конвейер работоспособен.** Оба метода DP-синтеза (Copula и MST) позволяют генерировать синтетические данные резюме, сохраняющие ключевые статистические свойства и пригодные для конечных задач (TSTR $R^2$ до 0.577 при baseline 0.667).
2. **Два метода (два профиля).** DP-Copula лучше сохраняет корреляции (Corr Sim ≈ 0.98) и маргинальные распределения (KS ≈ 0.91), но показывает умеренный CVPL Risk (~0.14). DP-MST обеспечивает более низкий CVPL Risk (~0.03–0.05) при несколько худшем качестве распределений (KS ≈ 0.88, TV ≈ 0.72).
3. **Оптимальный диапазон ε = 4.0–8.0.** В этом диапазоне достигается баланс между приватностью и полезностью: TSTR $R^2$ ≈ 0.55–0.58 (82–87% от baseline), Риск приватности < 0.15 (CVPL), MAE зарплаты ≈ 26 000 руб. (прирост ~14% к baseline).

4. **Текстовые данные приводятся к табличному виду.** Конвейер NER → LLM → Embeddings → K-Means позволяет представить неструктурированные описания опыта в виде кластерной переменной, пригодной для DP-синтеза. Это расширяет область применимости метода за рамки чисто структурированных данных.
5. **MIA-метрика требует осторожной интерпретации.** Высокая MIA Accuracy (0.84–0.94) обусловлена систематическими различиями в распределениях (артефакты дискретизации, хвосты), а не утечкой конкретных записей. CVPL-атака является более информативным индикатором риска для индивидуальных записей.

### 8.2 Ограничения

- Тестирование выполнено на суррогатных данных с заданными корреляциями — реальные данные могут содержать более сложные зависимости, выбросы и аномалии (в реальном проекте: 23% резюме с множественными версиями, 3.2% с датами в будущем, 32% со смешением кириллицы и латиницы).
- **Текстовый синтез не реализован** — использован кластерный суррогат; полноценная генерация текстовых описаний требует отдельного исследования (LLM + DP fine-tuning).
- **Атака вывода на атрибуты не реализована**, покрытие CNIL-категории «inference» обеспечено только косвенно через DP-гарантию; прямой AIA-тест планируется в v0.2.
- **Аудит смещений не выполнен**: для HR-задач это существенный пробел; сравнение отсутствия дисбаланса на реальных против синтетических данных планируется в v0.2.
- **Один запуск** с фиксированными начальными значениями не оценена дисперсия метрик между запусками; для статистической значимости рекомендуется N ≥ 5 запусков.
- **DP-CTGAN не включён** из-за ограничений среды выполнения (требуется GPU с ≥ 12 GB VRAM).
- **Размер выборки (10 000)** — адекватен для тестирования конвейера, но результаты не транслируются на малые наборы ($n < 2\,000$), типичные для внутренней HR-аналитики отдельных компаний, где DP-шум может полностью разрушить корреляции.

- Конечная задача минимальна (одна регрессия); не оценивались задачи кластеризации, ранжирования кандидатов, анализа текучести.
- **Хвостовые распределения.** Анализ хвостового распределения выполнен на уровне перцентилей (раздел 4.6), но не включает формальную оценку риска раскрытия приватных данных, это рекомендуется для реалистичных сценариев.

### 8.3 Рекомендации

1. **Повторение на реальных данных.** Выполнить эксперименты на подмножестве реальных данных (при наличии организационных мер) для валидации результатов. Приоритет — анализ k-анонимности реального набора и калибровка ε под фактическое распределение QI-комбинаций.
2. **Расширение методов.** Включить DP-CTGAN и AIM (Adaptive and Iterative Mechanism) для сравнения с Copula и MST. Для малых наборов ($n < 2\,000$) — исследовать PATE-подход.
3. **Текстовый синтез.** Исследовать генерацию синтетических текстовых описаний с использованием LLM + DP (например, DP fine-tuning или текстовая генерация с ε-бюджетом).
4. **Атака вывода на атрибуты.** Реализовать AIA-тест: предсказание DESIRED_SALARY (и других чувствительных атрибутов) по подмножеству QI на синтетических vs. реальных данных. Оценить прирост точности inference как функцию ε.
5. **Аудит смещений.** Для HR-применений — обязательный этап: сравнение demographic parity, equalized odds, disparate impact ratio по защищённым атрибутам (SEX, AGE_BIN) на реальных vs. синтетических данных. Оценить, сохраняет ли синтетика существующие диспропорции или создаёт новые.
6. **Расширение атак.** Добавить linkage-атаки с внешними источниками (моделирование LinkedIn-профилей как фоновых знаний), а также атаки на группы записей (group privacy).
7. **Множественные запуски.** Выполнить N ≥ 5 запусков для оценки дисперсии метрик и построения доверительных интервалов.
8. **Анализ малых подгрупп.** Для реального использования - специальная обработка подгрупп с $k < 5$: дополнительное обобщение QI, синтетическое дополнение (upsampling) перед синтезом, или исключение из выпуска.

# 9 Паспорт эксперимента

## 9.1 Идентификация

| Поле | Значение |
|---|---|
| **Идентификатор** | ABD-HH-DPSYNTH-2025-002 |
| **Серия** | Приложения к ПНСТ по синтезу данных |
| **Предыдущий кейс** | ABD-MA-DPCOPULA-2024-001 (Маркетинговая атрибуция) |
| **Дата проведения** | 2025-06-XX |
| **Версия конвейера** | V2 (pipeline_hh.py) |

## 9.2 Исходные данные

| Параметр | Значение |
|---|---|
| **Тип данных** | Суррогатные (сгенерированные программно) |
| **Объём** | 10 000 записей |
| **Число атрибутов** | 15 (14 после удаления ID) |
| **Категориальных** | 7 (SEX, REGION, EDUCATION_LEVEL, SPECIALIZATION, DESIRED_POSITION_LEVEL, CURRENT_EMPLOYMENT, LANGUAGE_EN) |
| **Числовых** | 7 (AGE, TOTAL_EXPERIENCE_YEARS, DESIRED_SALARY, SKILLS_COUNT, HAS_PORTFOLIO, LAST_UPDATE_DAYS, EMBEDDING_CLUSTER) |
| **Seed генерации** | 42 |
| **Основа** | Реальный проект: ~160 000 резюме, 45+ атрибутов |

## 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-калибровка зарплаты** | Да ($a \leq 0.35$, $\sigma \geq 0.08$) |
| **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-калибровка зарплаты** | Да ($a = 0.20$, $\sigma \geq 0.06$) |
| **Base seed** | 200 |

## 9.4 Параметры аудита

| Тест | Параметр | Значение |
|---|---|---|
| **CVPL** | Размер holdout | 2 000 |
| **CVPL** | Метод NN | BallTree, k=1 |
| **CVPL** | Порог proximity | 0.8 |
| **MIA** | Размер выборки | 3 000 (real) + 3 000 (syn) |
| **MIA** | Классификатор | RandomForest, 100 деревьев |
| **MIA** | Тест/обучение | 70% / 30% |
| **Singling Out** | Квази-идентификаторы | AGE_BIN, SEX, REGION, EDUCATION_LEVEL |
| **Singling Out** | Возрастные бины | 18–25, 26–35, 36–45, 46–55, 56–65 |

## 9.5 Целевая задача

| Параметр | Значение |
|---|---|
| **Задача** | Регрессия: предсказание DESIRED_SALARY |
| **Модель** | Gradient Boosting Regressor |
| **Гиперпараметры** | 200 деревьев, max_depth=5, lr=0.1 |
| **Признаки** | REGION, SPECIALIZATION, EXPERIENCE, EDUCATION, AGE, SKILLS, LANGUAGE, POSITION_LEVEL |
| **Целевая переменная** | DESIRED_SALARY |
| **Тест/обучение** | 80% / 20%, random_state=42 |
| **Протокол** | TSTR, TRTS |
| **Метрики** | $R^2$, MAE, RMSE |

## 9.6 Итоговые бюджеты приватности

| Этап | Бюджет | Примечание |
|---|---|---|
| **Синтез (M3)** | $\varepsilon \in \{1.0, 2.0, 4.0, 8.0, 16.0\}$ | Единственный DP-этап |
| **Итого** | $\varepsilon_{total} = \varepsilon_{synth}$ | Нет PSI, нет отдельного обезличивания |