

Применение ML для обнаружения мошеннических транзакций

Команда



Ботялина Дарья

Работа с аномалиями и генерация признаков



Стариков Егор

Интерпретация и диагностика моделей



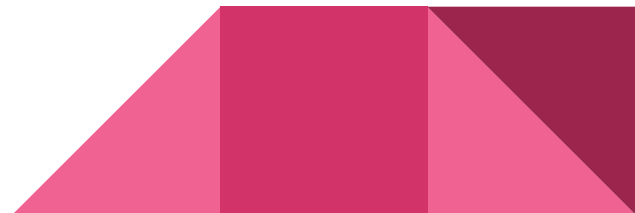
Ткаченко Александр

Baseline и презентация



Полковникова Мария

EDA и саммари



Датасет


1. E-Commerce Fraud Detection (Synthetic Dataset):

- a. Размер: 299,695 строк
- b. Признаков: 17 колонок
- c. Пропусков нет
- d. Доля fraud: ~2.206%


1. Группы признаков

- a. User/profile: account_age_days, total_transactions_user, avg_amount_user
- b. Transaction: amount, shipping_distance_km, promo_used, merchant_category, channel
- c. Security: avs_match, cvv_result, three_ds_flag
- d. Geo: country, bin_country
- e. Time: transaction_time

1. Цель проекта: предсказать is_fraud для транзакции, используя:

- a. транзакционные признаки
 - b. поведенческие признаки (много операций на пользователя)
 - c. аномалии и контекст
- 

EDA: ключевые открытия

- Вероятность мошенничества не сильно зависит от страны - только в 2 странах меньше 2% фрода (NL, DE)
 - Вероятность мошенничества при переводе за границу целых 11%, хотя при внутреннем переводе всего 1.5%
 - Матрица корреляций показала самую сильную зависимость между фродом и количеством денег(0.27), а также дистанцией перевода(0.2) и возрастом аккаунта (-0.12)
 - Количество мошеннических операций сильно зависит от времени суток, максимальное в 23, 1, 2, минимальное в 11, 15, 22
 - Большинство мошеннических операций совершается в четверг (2.3% от общего кол-ва операций), минимальное в среду и вторник (2.16%)
- 

BaseLine (CatBoost)

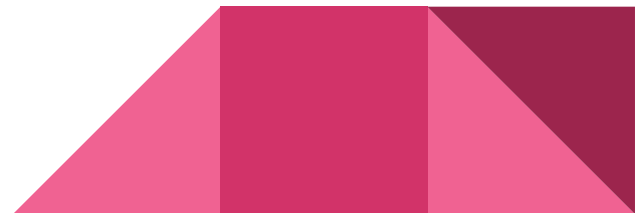
Без сверхглубокого анализа и брейнсторма решение через **CatBoost + stratified split 80/20 + class balancing** дало следующий результат:

- **ROC-AUC: 0.97784**
- **PR-AUC (Average Precision): 0.85269**

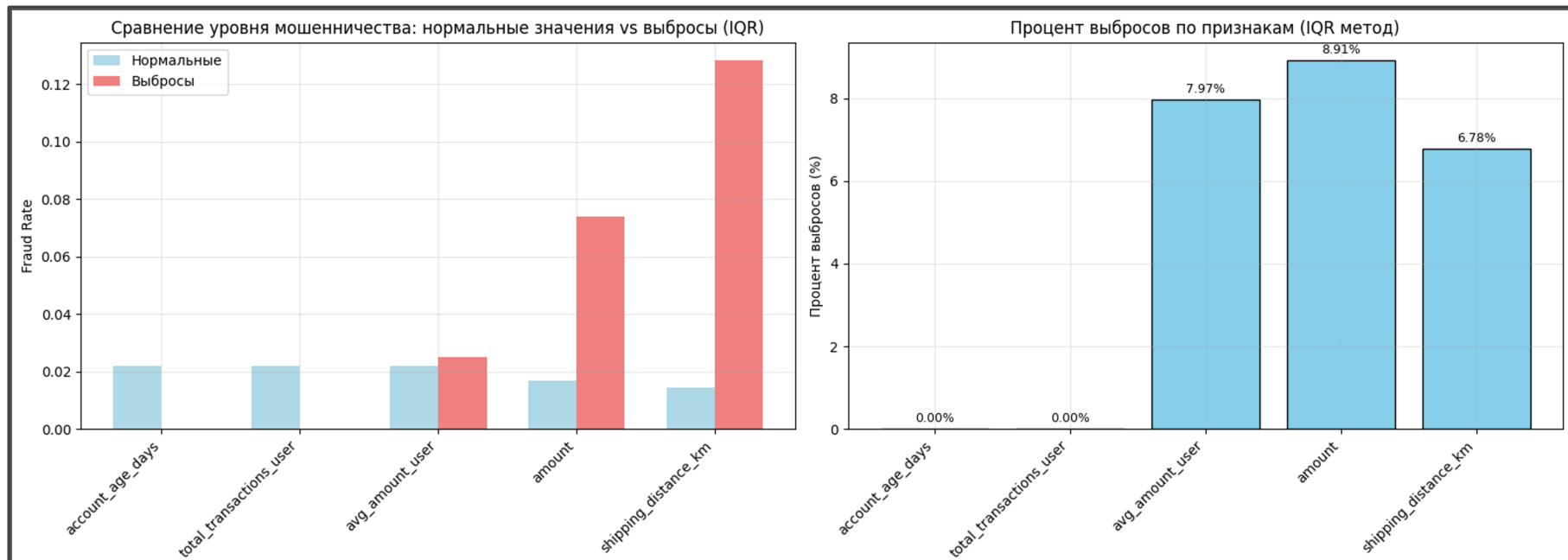


Статистические методы

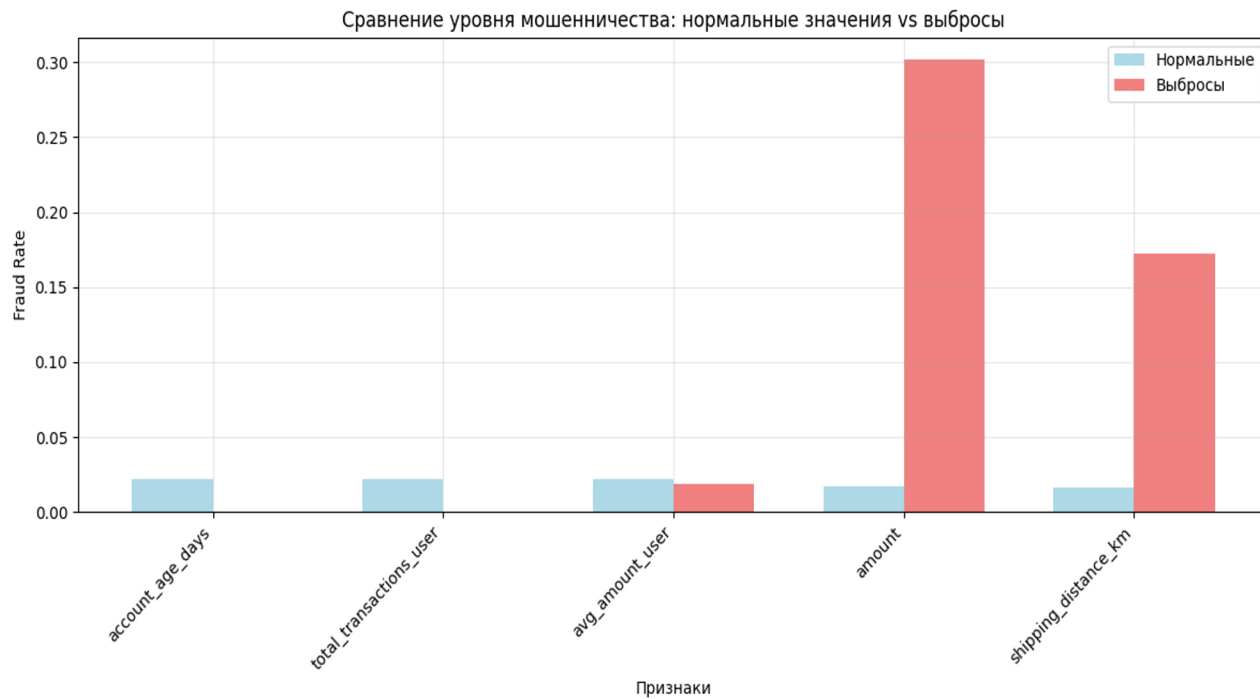
- Z-score:
 - Для amount доля z-выбросов 1.79%, но среди них доля мошенничества 30.16%, тогда как среди нормальных значений 1.70%.
 - Для shipping_distance_km доля z-выбросов 3.61%, fraud rate среди них 17.27%, среди нормальных 1.64%.
- IQR:
 - Для shipping_distance_km доля выбросов 6.78%, fraud rate среди них 12.81%, среди нормальных 1.44%
 - Для amount outliers доля IQR-выбросов 8.91%, fraud rate среди них 7.40%, среди нормальных значений 1.70%
- тест Габбаса для вспомогательной проверки крайних точек, так как у нас не нормальное распределение:
 - amount: Найдено выбросов по тесту Граббса: 100
 - shipping_distance_km: Найдено выбросов по тесту Граббса: 100



Визуализация (IQR)



Визуализация (Z-score)



Признаки на основе аномалий: статистические выбросы (Z-score, IQR)

- Для ключевых числовых признаков (amount, shipping_distance_km) выделили аномалии двумя устойчивыми критериями:
 - Z-score ($|z| > 3$) - фиксирует самые экстремальные значения
 - IQR ($1.5 \cdot IQR$) - выделяет нетипичные значения при асимметричных распределениях
- Добавили в датасет бинарные флаги выбросов вида “есть выброс по признаку X” (для суммы и дистанции, отдельно для Z-score и IQR)
- Построили счётчики аномалий (сколько признаков у транзакции попало в зону выбросов), чтобы учитывать “накопление подозрительности”

Смысл признаков: переводят редкость значения в явный риск-сигнал, который модель использует проще и устойчивее, чем сырые хвосты распределений



Признаки на основе аномалий: сложные выбросы и локальная плотность окружения

- Для поиска мультифакторных аномалий (редких комбинаций признаков) применили ML-детекторы: IsolationForest, LOF, One-Class SVM, Elliptic Envelope.
- Для каждого метода добавили флаг аномалии, затем собрали интегральные признаки: `anomaly_count` (сколько методов согласилось) и `consensus/strong_anomaly` (аномалия подтверждена ≥ 2 методами).
- LOF дополнительно отражает идею “плотности окружения”: транзакция считается аномальной, если она находится в области низкой локальной плотности относительно ближайших соседей.

Практический эффект: признаки аномалий повышают способность модели выделять `fraud` в редком классе, а также улучшают интерпретируемость (“аномальная сумма/доставка”, “аномалия подтверждена несколькими методами”).



Итоги

- Isolation Forest показывает лучший баланс: находит 45.4% мошеннических операций среди аномалий с точностью 20%. One-Class SVM дает самую высокую точность (22.5%), но находит только 8.5% мошенничества.
- Главный инсайт: сильные аномалии, обнаруженные ≥ 2 методами (11,808 транзакций), имеют fraud rate 23.76% - почти каждая 4-я транзакция в этой группе является мошеннической. Это позволяет фокусировать ручную проверку на наиболее рискованных операциях.

Method	Precision	Recall	F1
IsolationForest	0.200	0.454	0.278
LOF	0.106	0.241	0.148
One-Class SVM	0.225	0.085	0.123
EllipticEnvelope	0.184	0.416	0.255

Обработка категориальных переменных

- Для признаков с большим числом категорий применен Target Encoding:
merchant_category → merchant_category_te, country → country_te, bin_country → bin_country_te
- Для признака channel выполнен One-Hot Encoding

Промежуточный контроль: метрики фиксировались на baseline и после добавления новых признаков для корректного сравнения.



Признаки на основе похожих наблюдений (аналог ближайших соседей, но без координат)

- Построены поведенческие признаки на уровне пользователя (агрегации по `user_id`): статистики по сумме и дистанции, а также характеристики активности.
- Сформированы признаки нетипичности транзакции относительно профиля пользователя: `amount_zscore_user` и `dist_zscore_user`.
- Добавлен локальный контекст по категории мерчанта (агрегации по `merchant_category`): `amount_vs_merchant` как мера типичности суммы для категории.

Смысл: вместо kNN-расстояний используются устойчивые прокси «близости» - сравнение с исторической нормой пользователя и типичностью по категории.



Временные и контекстные признаки

- Временные признаки: извлечены компоненты времени и выполнено sin/cos-кодирование циклов hour_sin/cos, dow_sin/cos, month_sin/cos; добавлены режимные индикаторы is_night, is_business_hours, is_evening, is_weekend.
- Контекстные признаки: is_cross_border (country \neq bin_country); security_score = AVS + CVV + 3DS и флаги all_security_passed/no_security; отношения к среднему пользователя amount_to_avg_ratio и amount_diff_from_avg; логистика is_long_distance (по 0.9-квантили) и взаимодействия факторов.

Сформирован интерпретируемый risk_score как агрегирование ключевых факторов риска и их комбинаций. Мы объединили гео-контекст, безопасность, логистику и время, задав повышенные веса наиболее сильным факторам: $\text{risk_score} = 3 \cdot \text{is_cross_border} + 2 \cdot \text{no_security} + (1 - \text{three_ds_flag}) + \text{is_long_distance} + \text{is_night}$.



Отбор признаков и итоговый набор

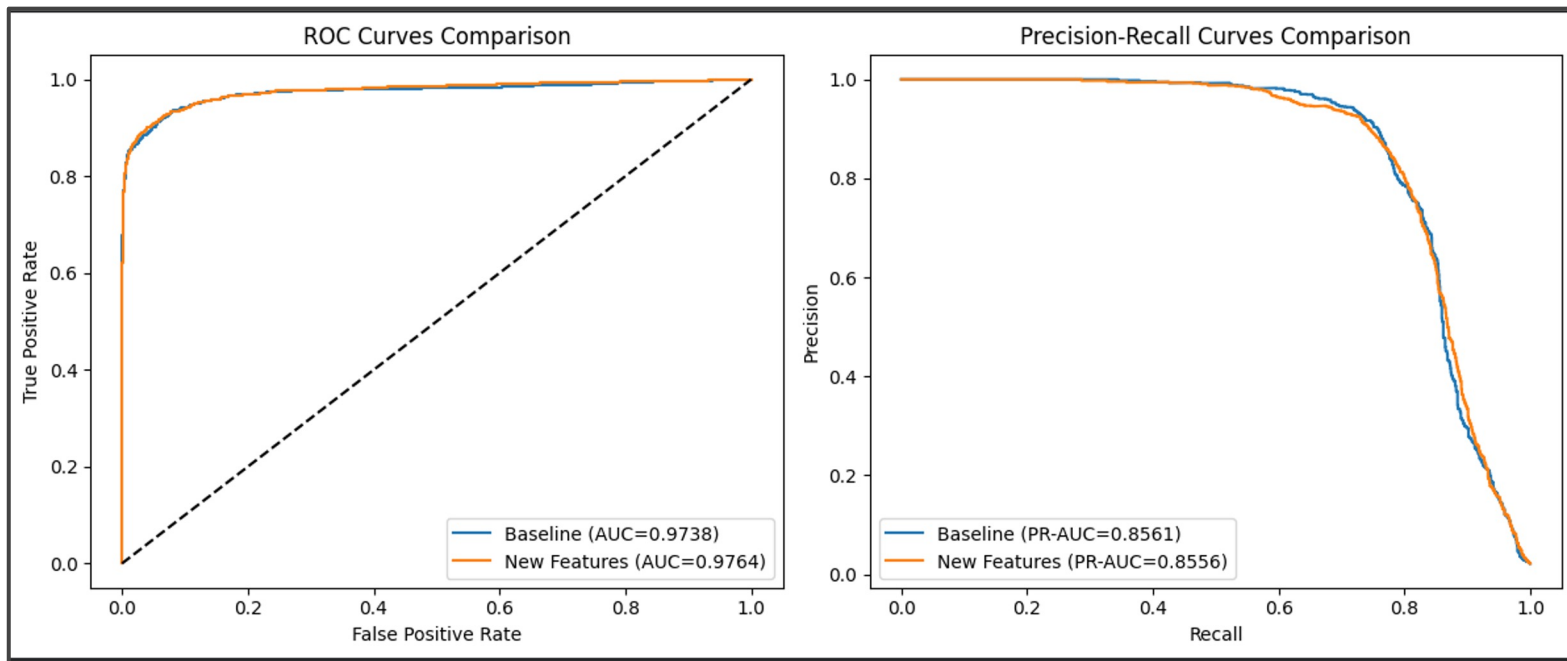
- Выполнена фильтрация по статистической связности: рассчитана абсолютная корреляция признаков с целевой переменной и сформирован ранжированный список (sanity-check).
- Выполнен встроенный отбор через модель: обучен CatBoost и извлечены feature importances; по важности выбран компактный набор признаков.

Итог: сформирован набор из 25 признаков (top-25 по CatBoost importance), модель переобучена на этом наборе, метрики зафиксированы и сопоставлены с baseline.



Результаты модели с новыми признаками

Модель улучшилась относительно бейзлайна



Локальные объяснения: сравнение SHAP и LIME на одной транзакции

- Выбран конкретный пример fraud-наблюдения из тестовой выборки ($y_{\text{test}} = 1$) и построены локальные объяснения двумя методами.
- SHAP показал вклад каждого признака в итоговую вероятность fraud для выбранной транзакции (какие признаки подтолкнули решение вверх или вниз).
- LIME аппроксимировал поведение модели в окрестности точки и выдал список интерпретируемых условий (диапазоны значений признаков), которые повышают/понижают вероятность fraud.

Сравнение результатов: методы согласуются по ключевым признакам риска, однако LIME даёт более похожее на правило объяснение, а SHAP - более стабильную и теоретически обоснованную декомпозицию вкладов.




Интерпретация моделей: глобальные объяснения (SHAP + LIME)

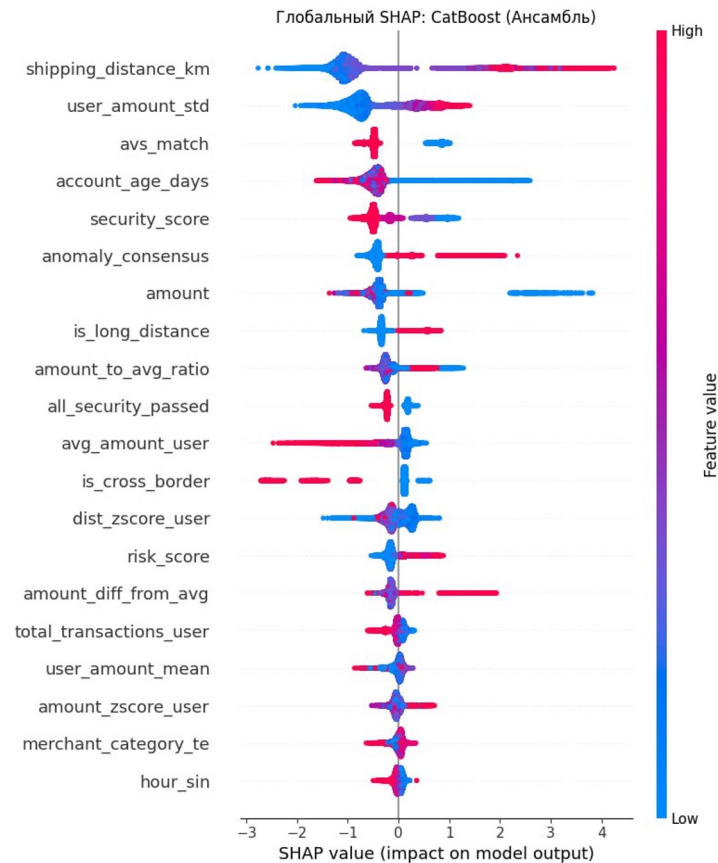
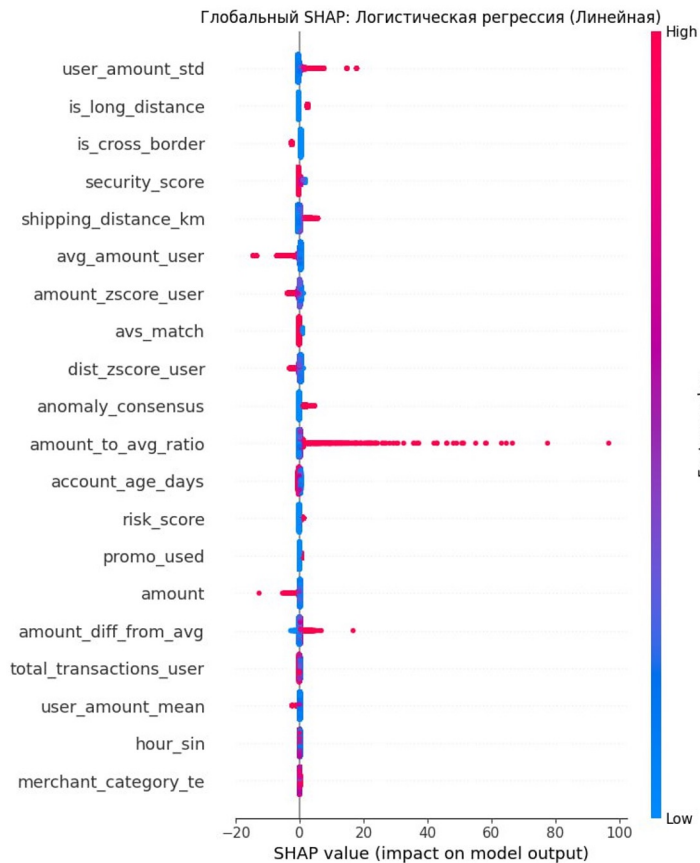
Цель этапа: проверить корректность поведения моделей и понять, какие признаки вносят основной вклад в предсказание мошенничества.

- Рассмотрены две модели разных классов: линейная (Logistic Regression) и ансамблевая (CatBoost). Для линейной модели выполнена стандартизация признаков (StandardScaler).
- Построены глобальные интерпретации: SHAP summary plot для CatBoost и для логистической регрессии, чтобы ранжировать факторы по среднему влиянию на вероятность fraud.
- Дополнительно применён LIME (Tabular) для получения интерпретируемых локальных правил и последующей агрегации выводов по нескольким наблюдениям (SP-LIME) как “портретов” типовых сценариев.

Вывод глобального анализа: ключевые факторы риска согласуются с логикой задачи (безопасность платежа, cross-border, нетипичность поведения, аномальные суммы/доставка); CatBoost лучше отражает взаимодействия признаков, чем линейная модель.



Интерпретация моделей - глобальный SHAP



SHAP-эмбединги: построение и поиск аномалий/сдвигов

- Получены SHAP-эмбединги для модели CatBoost в виде отдельной функции `get_shap_embeddings(model, X_data)`, возвращающей матрицу вкладов `shap_<feature>` для каждой транзакции (train и test).
- SHAP-эмбединги использованы как альтернативное пространство представления транзакций, где аномалии интерпретируются как нетипичные профили влияния признаков
- Для выявления аномалий в SHAP-пространстве применён IsolationForest (`contamination=0.01`).
- Выполнена очистка обучающей выборки: удалено 2398 наблюдений; модель переобучена на очищенных данных.

Сравнение качества: ROC-AUC после очистки = 0.97340, что немного ниже исходного уровня, получается, что при агрессивной очистке теряется часть информативных fraud-паттернов



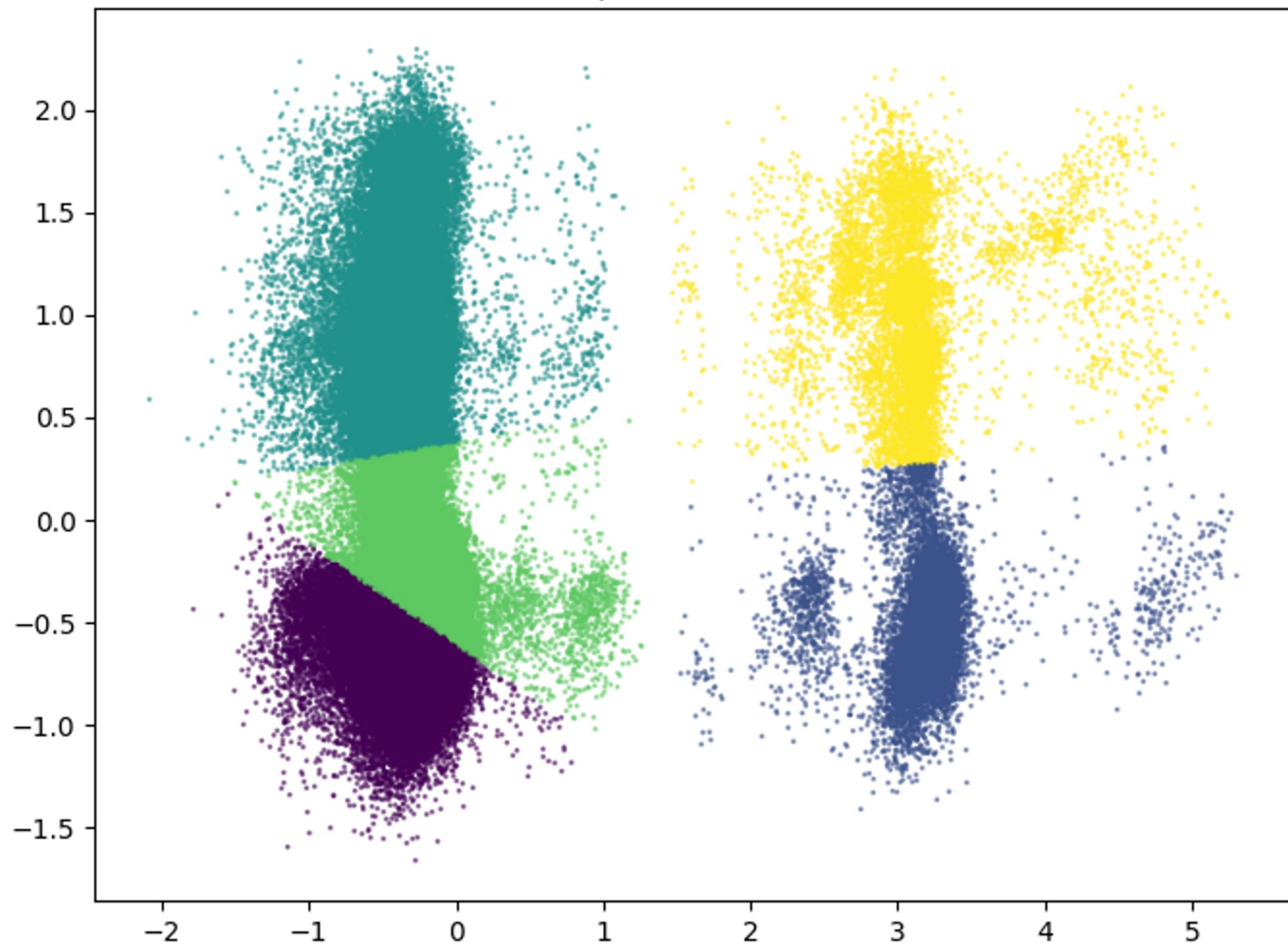
Кластеризация SHAP-эмбеддингов и признак «кластер»

- Для кластеризации SHAP-эмбеддингов выполнено снижение размерности PCA до 2 компонент и применён k-Means ($k=5$); получена визуализация кластеров в проекции PCA.
- Дополнительно протестирован DBSCAN на стандартизированных SHAP-эмбеддингах; при выбранных параметрах DBSCAN не выделил устойчивых кластеров (0 кластеров).
- Номер кластера добавлен в датасет как категориальный признак cluster; CatBoost переобучен с учётом `cat_features=['cluster']`.

Результат: ROC-AUC с кластерами = 0.97566 - прирост относительно исходной модели минимальный; основной эффект кластеризации - интерпретируемая сегментация транзакций по типовым SHAP-профилям, а не существенное улучшение качества.



Кластеры SHAP (K-Means)



Валидация с SHAP-эмбедингами: сравнение представлений признаков

- В качестве валидации использовано сравнение на отложенной тестовой выборке (hold-out).
- Построены SHAP-эмбединги как матрица вкладов `shap_<feature>` для каждой транзакции (`df_shap_train` / `df_shap_test`) на основе CatBoost (TreeExplainer).
- Обучен CatBoost на исходных признаках и отдельно CatBoost только на SHAP-эмбедингах; результаты на test: ROC-AUC (исходные признаки) 0.97640, ROC-AUC (только SHAP-эмбединги) 0.96381.

Вывод: SHAP-эмбединги сохраняют существенную информацию о транзакции, но в текущей реализации уступают исходному пространству; их целесообразно использовать как инструмент диагностики/анализа, а не как полную замену признаков.



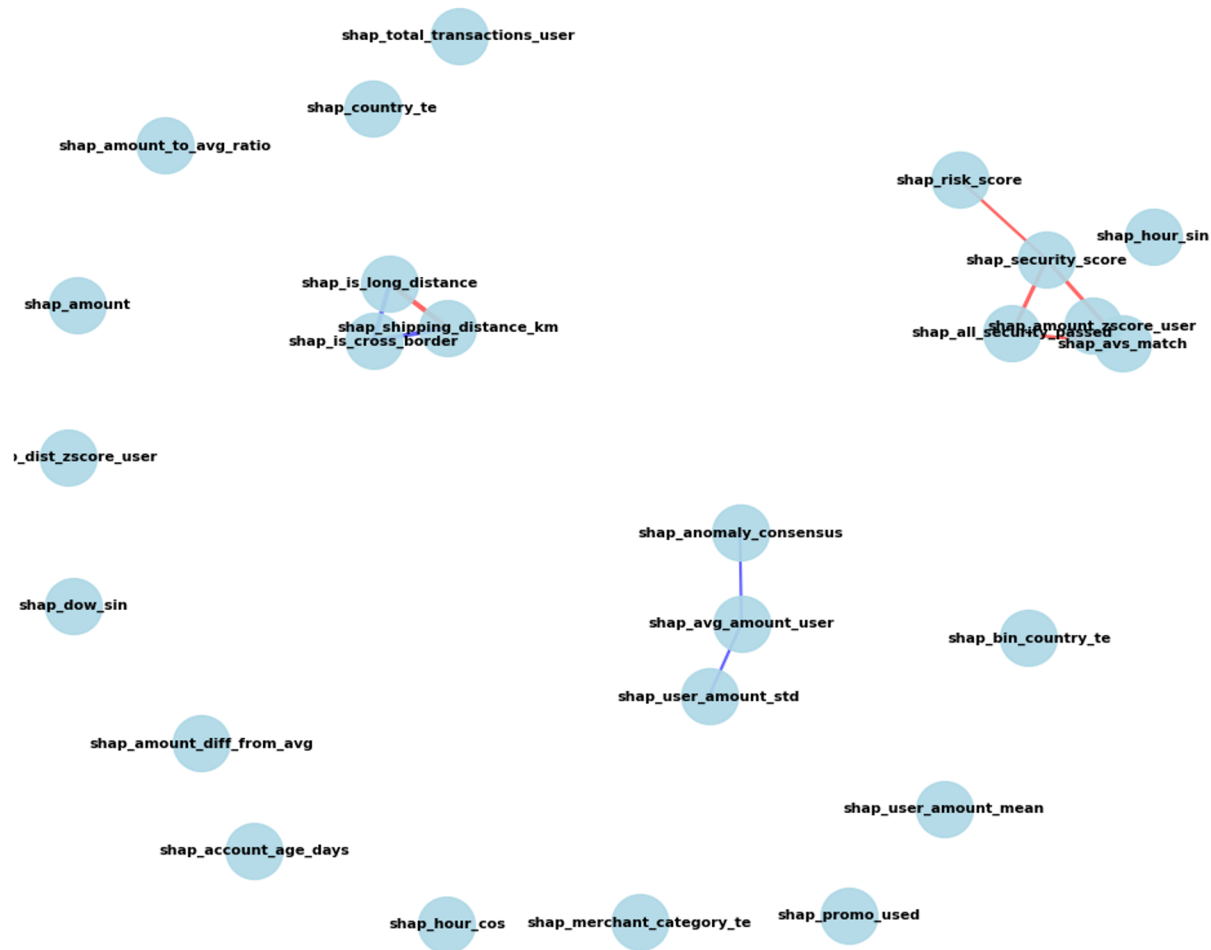
Shapley Flow: граф взаимосвязей признаков по SHAP-корреляциям

- На неочищенных данных построена корреляционная матрица SHAP-признаков (train) и сформирован граф NetworkX: узлы - shap-признаки, ребра добавляются при $|\text{corr}| > 0.5$; знак связи кодируется цветом, сила - толщиной ребра.
- Рассчитана центральность по степени: наиболее центральный признак shap_security_score (связан с 3 другими), что подчёркивает роль “безопасности” как узлового фактора в объяснениях модели.
- Выполнено разбиение на сообщества (greedy modularity): найдено 18 групп, примеры интерпретируемых сообществ:
 - безопасность: shap_security_score, shap_avs_match, shap_risk_score, shap_all_security_passed
 - гео/логистика: shap_shipping_distance_km, shap_is_cross_border, shap_is_long_distance
 - поведение/аномалии: shap_user_amount_std, shap_avg_amount_user, shap_anomaly_consensus

Вывод: Shapley Flow выявляет сценарии совместного влияния факторов риска, а не только важность отдельных признаков.



Shapley Flow: Граф взаимосвязей признаков (по SHAP-корреляциям)



Сдвиги и устойчивость структуры (train vs test) и итог по качеству

- Построен аналогичный граф связей на test и выполнено сравнение множеств ребер: стабильных связей 5, исчезло 4, появилось новых 4; пример новой связи: (shap_is_cross_border, shap_shipping_distance_km)
- Интерпретация: часть структуры объяснений сохраняется, однако присутствуют изменения, что указывает на чувствительность корреляционной структуры SHAP-признаков к выборке и необходимость аккуратной трактовки сдвигов

По качеству улучшения не получено: очистка в SHAP-пространстве (IsolationForest) снизила ROC-AUC до 0.97340, добавление кластеров SHAP дало 0.97566; обучение только на SHAP-эмбедингах хуже, чем на исходных признаках (0.96381 vs 0.97640)



Итоги

- Проведён полный цикл решения задачи fraud detection: EDA → feature engineering → обучение модели → интерпретация (SHAP/LIME) → эксперименты в SHAP-пространстве
- Выявлены ключевые факторы риска: слабая безопасность (AVS/CVV/3DS), cross-border, аномальные сумма/доставка, поведенческие отклонения пользователя
- Добавлены интерпретируемые признаки (risk score, security score, поведенческие и аномальные фичи) и выполнен отбор признаков

Финальная модель (CatBoost):

- ROC-AUC 0.97640
- PR-AUC 0.85556


Fraud:

- Precision 0.34
- Recall 0.90
- F1 0.49

Улучшение относительно baseline:

- ROC-AUC: -0.00144
- PR-AUC: +0.00287

Fraud:

- Precision: +0.04
 - Recall: -0.02
 - F1: +0.04
- 



Спасибо за внимание!