**Документация для решения задачи: "Выявление нетипичных операций по транзакционной активности"**

**Обзор задачи**

Задача — разработать инструмент для выявления аномальных паттернов в финансовом поведении клиентов, используя модели машинного обучения. Мы анализируем транзакционные данные, проводим исследовательский анализ (EDA) и строим модель для автоматической детекции аномалий.

**Часть 1: Исследовательский анализ данных (EDA)**

1. **Загрузка и очистка данных**
   * Пропуски были удалены из данных с помощью data.dropna().
2. **Категориальные данные**
   * Колонки device\_type, tran\_code, card\_type, oper\_type, и card\_status закодированы в числовой формат с помощью LabelEncoder.
3. **Обработка временных признаков**
   * Из временного признака datetime извлечены час (hour) и день недели (day\_of\_week), что позволяет учитывать сезонные паттерны в транзакционной активности.
4. **Нормализация данных**
   * Числовые признаки sum, balance и pin\_inc\_count приведены к единому масштабу с помощью MinMaxScaler.
5. **Вывод**
   * Признаки для обучения модели: sum, balance, pin\_inc\_count, device\_type, hour, day\_of\_week, tran\_code, oper\_type, card\_status.

**Часть 2: Построение моделей и методы обучения**

**1. Модель ближайших соседей (KNN)**

* **Назначение**: Используется для поиска ближайших точек (транзакций) и определения аномалий.
* **Гиперпараметры**: Подбираются с помощью GridSearchCV, включая:
  + Количество соседей (n\_neighbors),
  + Метрики расстояния (metric),
  + Алгоритмы поиска (algorithm),
  + Размеры листьев для деревьев поиска (leaf\_size).
* **Аномалии**: Транзакции, расстояние которых до ближайших соседей превышает оптимальный порог, классифицируются как аномалии.

**2. Ансамблевые методы**

* **Voting Classifier**:
  + Включает модели RandomForestClassifier и LogisticRegression.
  + Использует мягкое голосование для предсказания меток.

**Часть 3: Оценка и интерпретация модели**

1. **Порог для аномалий**
   * Оптимальный порог: 80% от максимального расстояния до ближайших соседей.
2. **Метрики качества**
   * С помощью программы, которая была нам предоставлена, мы достигли результата метрики в 0.9909500556194498
3. **Вывод аномалий**
   * **Обнаруженные паттерны**:
     + Аномалии чаще связаны с высокой суммой транзакции (sum) и низким балансом (balance).
     + Большое число аномальных операций происходит в специфические часы суток.

**Часть 4: Визуализация результатов**

1. **Scatterplot: Сумма и баланс**
   * Красные точки (аномалии) четко выделяются среди нормальных транзакций.
2. **Корреляционная матрица**
   * Визуализированы взаимосвязи между признаками и метками аномалий.

**Вывод**

Разработанная модель KNN в сочетании с ансамблевыми методами эффективно выявляет аномальные операции. Дальнейшая оптимизация порога и включение новых источников данных могут еще больше повысить эффективность модели.