**Документация для решения задачи: "Выявление аномалий в финансовых транзакциях"**

**Описание задачи**

Цель — автоматическое выявление аномалий (нетипичных операций) в транзакционных данных с использованием методов машинного обучения (ML). Анализ включает исследовательский этап (EDA), построение модели KNN, а также ансамблевого классификатора (VotingClassifier) для детектирования аномалий.

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

1. **Загрузка и обработка данных**
   * Данные загружаются из CSV-файла и обрабатываются для подготовки к обучению.
2. **Категориальные данные**
   * Преобразование категориальных признаков (device\_type, tran\_code, card\_type, oper\_type, card\_status) в числовые с помощью LabelEncoder.
3. **Создание временных признаков**
   * Из временной метки (datetime) извлекаются признаки:
     + hour — час операции,
     + day\_of\_week — день недели.
4. **Нормализация числовых данных**
   * Признаки sum, balance, pin\_inc\_count масштабируются в диапазон [0, 1] с использованием MinMaxScaler.
5. **Формирование выборок**
   * Используемые признаки: sum, balance, pin\_inc\_count, device\_type, hour, day\_of\_week, tran\_code, oper\_type, card\_status.
   * Данные разделяются на обучающую (X\_train) и тестовую (X\_test) выборки в пропорции 80/20.

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

**1. Модель KNN (ближайшие соседи)**

* **Описание**: Используется для нахождения "ближайших соседей" и определения аномальных транзакций, основываясь на расстоянии до соседей.
* **Оптимизация параметров**:
  + Параллельно обучаются модели с различными значениями параметров:
    - n\_neighbors (от 5 до 50),
    - Метрики расстояний (euclidean, manhattan, chebyshev, cosine).
  + Используется библиотека joblib для ускорения расчётов.
* **Порог для аномалий**:
  + Устанавливается на уровне 80% от максимального расстояния (threshold).
* **Выбор лучшей модели**:
  + Модель с наибольшей F1-метрикой выбирается как оптимальная.

**2. Ансамблевый классификатор (VotingClassifier)**

* **Состав**:
  + **RandomForestClassifier** — обеспечивает устойчивость к шуму.
  + **LogisticRegression** — добавляет интерпретируемость.
* **Механизм**:
  + Используется мягкое голосование (voting='soft'), объединяющее предсказания обеих моделей.

**Часть 3: Результаты и оценка модели**

**Оценка качества**

1. **F1-Score**:
   * Тренировочная выборка: 0.99 (высокая точность детектирования аномалий).
   * Тестовая выборка: Результаты аналогичны, что свидетельствует о хорошем обобщении модели.
2. **Классификационный отчёт**:
   * Отчёт включает оценку точности, полноты и F1-метрики для классов (аномалии/нормальные транзакции).

**Идентификация аномалий**

* **Аномальные транзакции**:
  + Обнаружены транзакции с высокой суммой (sum) и низким балансом (balance).
  + Чаще всего аномалии происходят в определённые часы и дни недели.

**Часть 4: Визуализация**

1. **Scatterplot: Сумма и баланс**
   * Красные точки (аномалии) выделяются на фоне нормальных транзакций (синие точки).
2. **Корреляционная матрица**
   * Демонстрирует взаимосвязь между признаками и флагом аномалий (anomaly\_flag).

**Часть 5: Оценка решения**

**Качество модели**

* Оптимальная модель KNN + VotingClassifier демонстрирует высокую F1-метрику, подходящую для задачи аномалий.
* Параллельная оптимизация гиперпараметров позволила ускорить процесс обучения.

**Заключение**

Методология эффективно выявляет аномальные транзакции с использованием KNN и VotingClassifier. Результаты визуализации подтверждают корректность модели и предлагают возможные направления для дальнейшего улучшения.