Прогнозирование аномалий во временных рядах

Желудев К.И.

Осенний семестр 2025-2026 гг.

Предметные области и данные

1. **Авиадиспетчеризация** (уровень топлива, расход топлива, температура в двигателях)

- Временные ряды датчиков системы впрыска топлива
- Один период нормального и четыре различных периода аномального режимов работы двигателя
- Синтетическая сборка длинных рядов
    a. Клонируем “нормальные” периоды
    b. Произвольно вставляем “аномальные” периоды, предварительно их аугментируя
    c. На обучение идет ряд без аномалий, для теста – с аномалиями

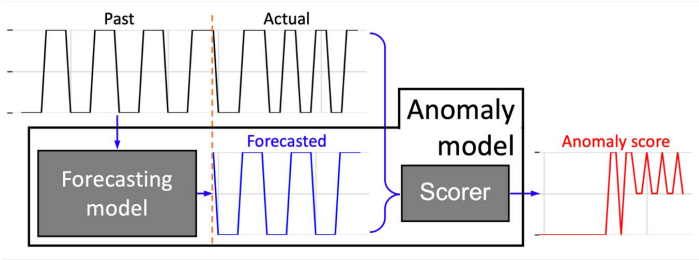
2. **Финансовые транзакции** (fraud/норма)

- Транзакционный датасет Kaggle Credit Card Fraud Detection
- 0.172% всех транзакций – мошеннические
- Данные чувствительные => преобразование PCA до 28 главных компонент
- Кол-во транзакций в единицу времени непостоянно => агрегация транзакций в интервалы (например по 1 мин)
- Аномалией называем наличие fraud в интервале

Общий pipeline

1. Подготовка данных
   a. загрузка исходных данных
   b. формирование временного ряда
   c. разделение на обучающую (без аномалий) и тестовую выборки
2. Моделирование нормального поведения
   a. обучение прогнозной модели (LinearRegression / RandomForest)
   b. построение прогноза временного ряда
3. Анализ отклонений от нормы
   a. вычисление ошибок прогноза (residuals)
   b. вычисление точечной меры аномальности (Norm-score)
4. Контекстный анализ аномалий
   a. обучение модели по Norm-score на нормальных данных (KMeans по окнам)
   b. вычисление контекстного anomaly score
5. Оценка качества

   Расчет метрик ROC-AUC (Какова вероятность, что anomaly score для аномальной точки будет больше anomaly score для нормальной)



Авторегрессия

Рассматривается многомерный временной ряд: $\mathbf{x}_t \in \mathbb{R}^d$

Предполагается, что текущее состояние системы зависит от L предыдущих состояний:

$$\mathbf{x}_t = f(\mathbf{x}_{t-1}, \mathbf{x}_{t-2}, \ldots, \mathbf{x}_{t-L}) + \boldsymbol{\varepsilon}_t$$
$$\mathbf{r}_t = \mathbf{x}_t - \hat{\mathbf{x}}_t$$

Функция f аппроксимируется:

- линейной регрессией
- нелинейной моделью Random Forest

Причем эта функция – результат обучения на нормальных данных.

Аномалии проявляются как изменение структуры ошибки прогноза $\mathbf{r}_t$

Методы обнаружения аномалий

Два оценщика

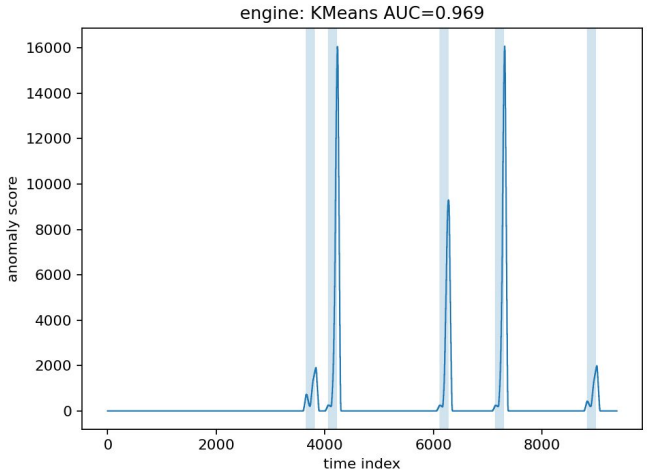
1. **Norm score (точечная мера)** $s_t^{(\text{norm})} = \|\mathbf{r}_t\|_1$

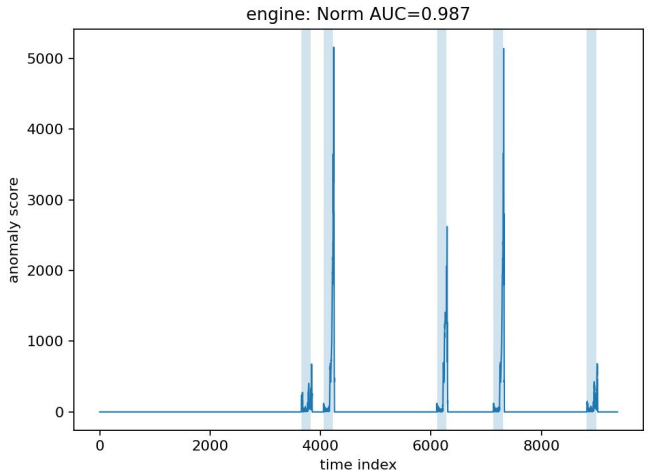
- агрегирует ошибку прогноза по всем признакам
- выявляет резкие амплитудные отклонения
- служит базовой статистикой аномальности

2. **KMeans window score (контекстная мера)**

- применяется к “временным окнам” Norm-score $\mathbf{z}_t = [s_{t-w+1}, \ldots, s_t]$
- k-means обучается на окнах с данными без аномалий
- аномальность определяется как расстояние до ближайшего кластера

Результаты: Engine (forecaster - linear)



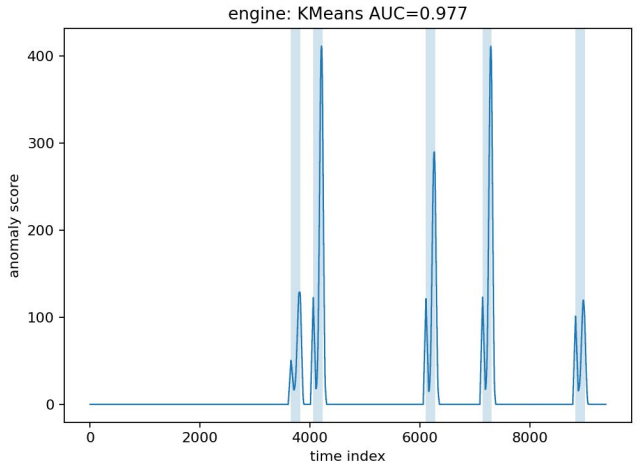


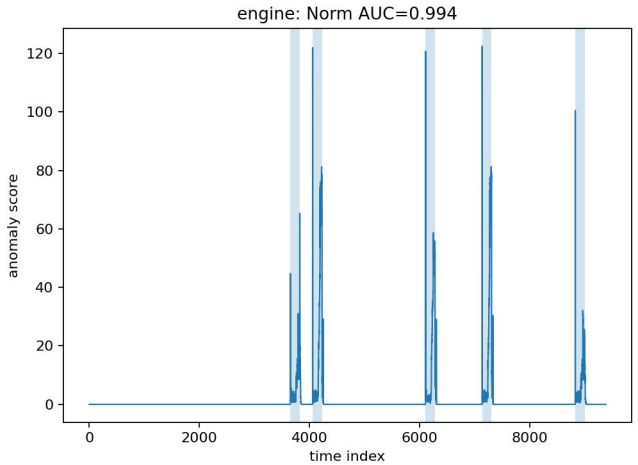
Интерпретация

- Аномальные режимы нарушают динамику сигналов
- Ошибка прогноза резко возрастает
- Метод эффективно выявляет аномалии

python detect.py ^
--domain engine ^
--forecasting linear ^
--lags 25 ^
--normal-periods 200 ^
--anom-inserts 5 ^
--out artifacts/engine_linear

Результаты: Engine (forecaster - random forest)



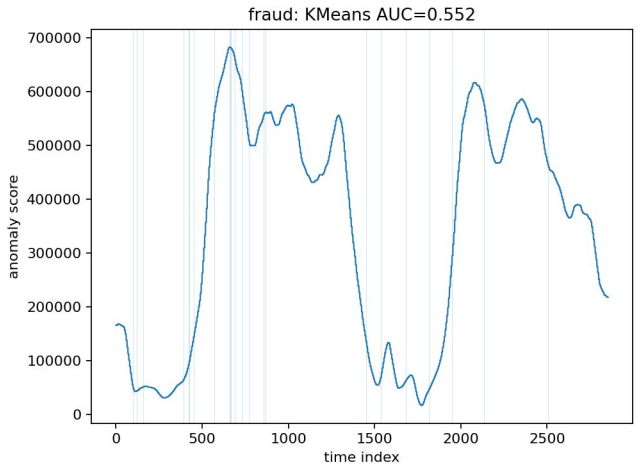


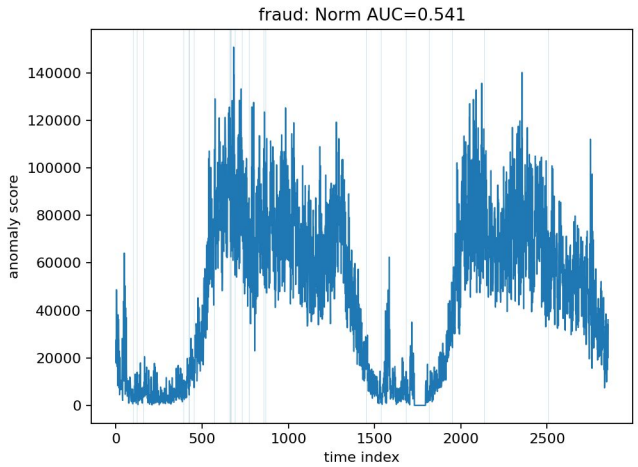
Интерпретация

- Аномальные режимы нарушают динамику сигналов
- Ошибка прогноза резко возрастает
- Метод эффективно выявляет аномалии

python detect.py ^
--domain engine ^
--forecasting rf ^
--lags 25 ^
--rf-n 200 ^
--normal-periods 200 ^
--anom-inserts 5 ^
--out artifacts/engine_rf

Результаты: Fraud (forecaster - linear)



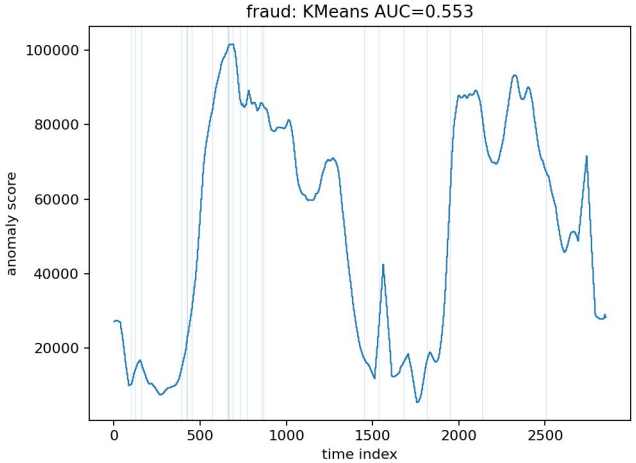


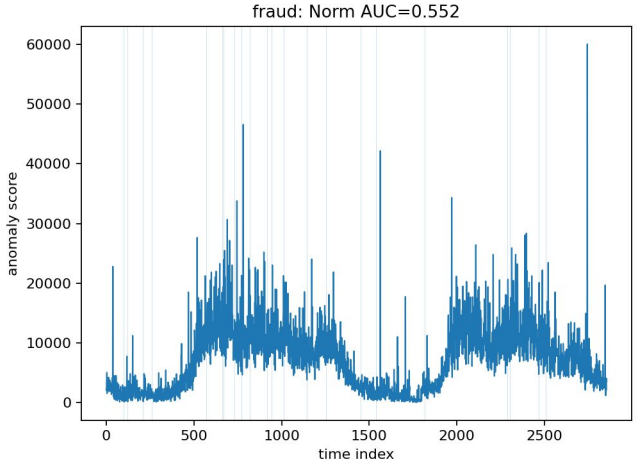
Интерпретация

- Fraud — это редкое событие, а не нарушение динамики временного ряда
- Агрегированные ряды не чувствительны к отдельным транзакциям
- Методы прогнозирования оказываются неэффективными

python detect.py ^
--domain fraud ^
--forecasting linear ^
--lags 25 ^
--fraud-csv data/fraud/raw/creditcard.csv ^
--fraud-agg-sec 60 ^
--out artifacts/fraud_linear

Результаты: Fraud (forecaster - random forest)





Интерпретация

- Fraud — это редкое событие, а не нарушение динамики временного ряда
- Агрегированные ряды не чувствительны к отдельным транзакциям
- Методы прогнозирования оказываются неэффективными

python detect.py ^
--domain fraud ^
--forecasting rf ^
--lags 25 ^
--rf-n 200 ^
--fraud-csv data/fraud/raw/creditcard.csv ^
--fraud-agg-sec 60 ^
--out artifacts/fraud_rf

Выводы

- Один и тот же метод по-разному работает в разных предметных областях
- Эффективность зависит от природы аномалий
- Прогнозно-ориентированные методы подходят для режимных аномалий
- Агрегация теряет информацию => требуются более “умные” модели