Московский институт электроники и
математики им. А.Н. Тихонова

ПОИСК АНОМАЛИЙ В АГРЕГИРОВАННЫХ ЖУРНАЛЬНЫХ ДАННЫХ ОТ СЕРВЕРОВ DNS

Половцев Александр Валерьевич
Егоров Алексей Михайлович

indata
ФОНД РАЗВИТИЯ СЕТЕВЫХ ТЕХНОЛОГИЙ «ИНДАТА»

Проблематика

В системах мониторинга, например в НСДИ, дежурные смены должны быстро принимать решения о вмешательстве в работу систему или эскалации возникших проблем на более высокие уровни в иерархии принятия решений

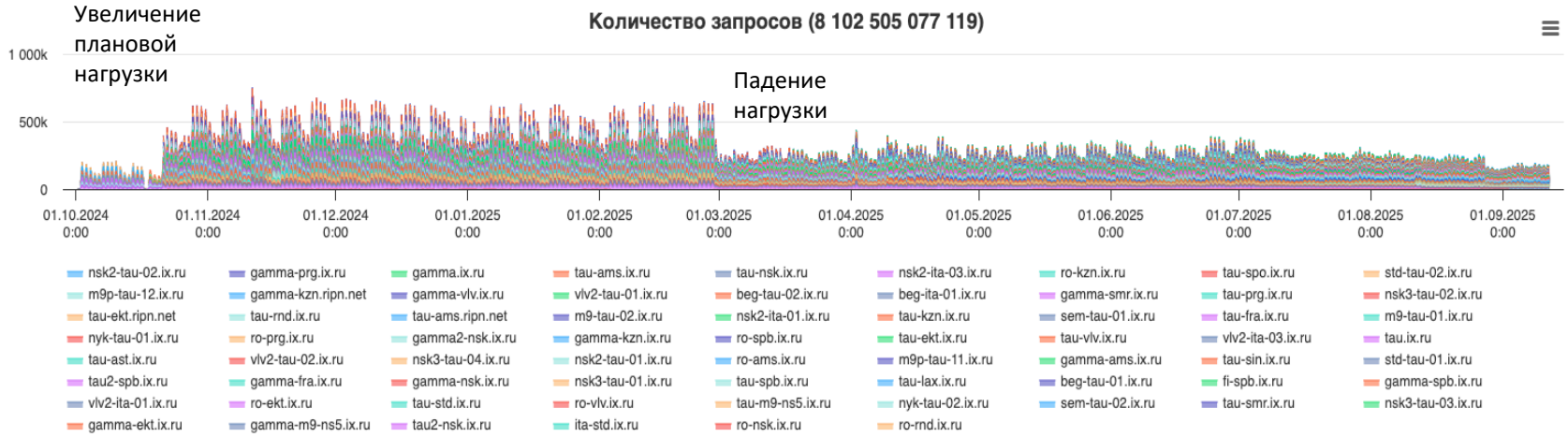
При этом в большинстве систем мониторинга есть несколько состояний, которые описывают контролируемую систему. Каждое состояние определяет набор действий операторов.

Например, в EBERO (Emergency Back-end Registry Operator) выделяют:

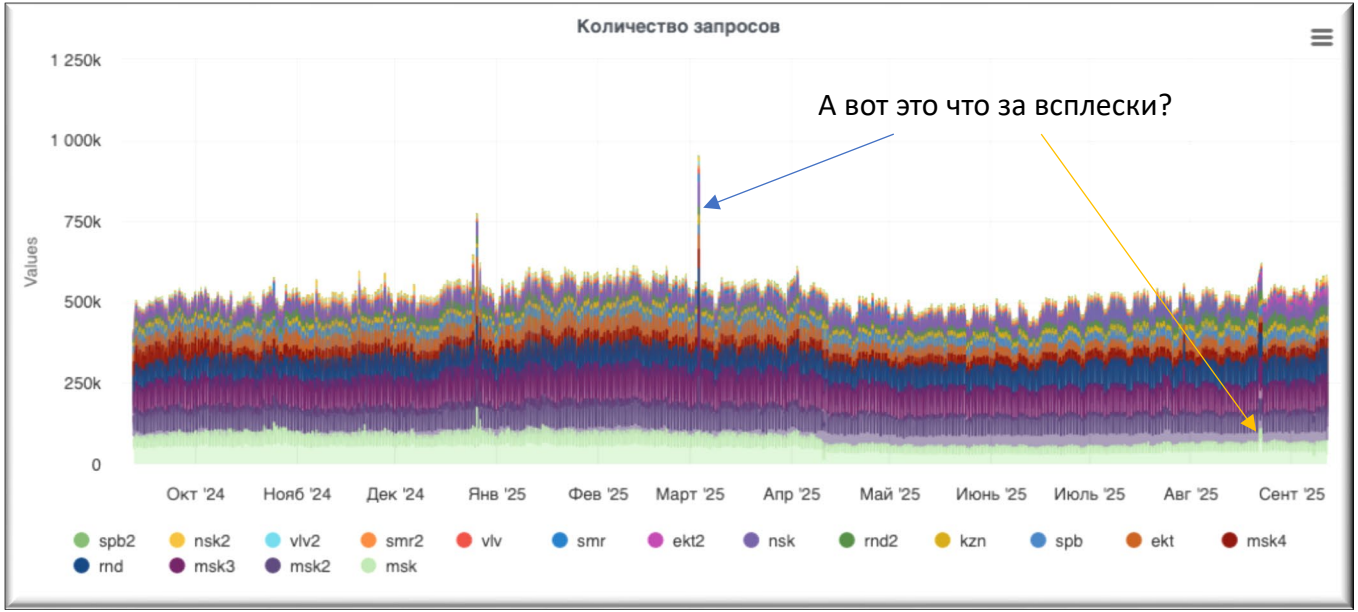
- Режим готовности;
- Режим повышенной готовности;
- Режим объявления о Событии;
- Режим развертывания;
- Режим стабилизации;
- Режим завершения.

Часто переход с режима на режим, из одного состояния в другое может быть связан с изменениями наблюдаемых параметров мониторинга. Необычное изменение значения параметра – это аномалия.

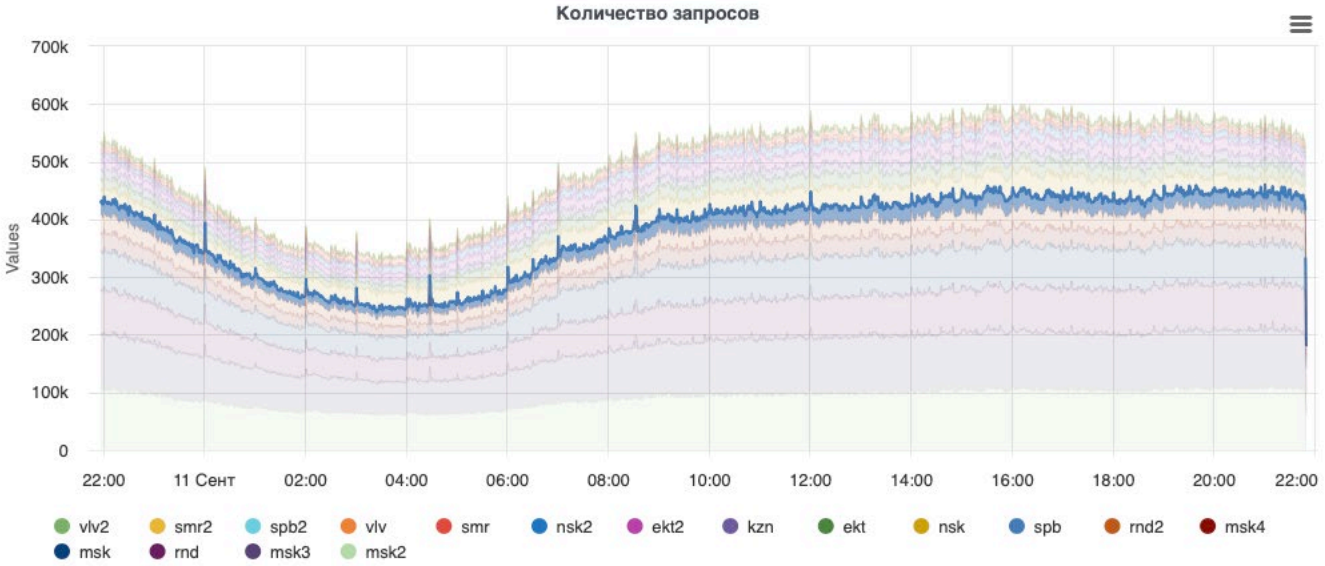
Что такое аномалия



Что такое аномалия



Что такое аномалия



Задача: построить модели обнаружения аномалий для такого сорта графиков

Методы

Старая добрая статистика

- Z-Score
- EWMA (Exponentially Weighted Moving Average)

Прогнозирующие модели (не совсем традиционный анализ временных рядов)

- Prophet

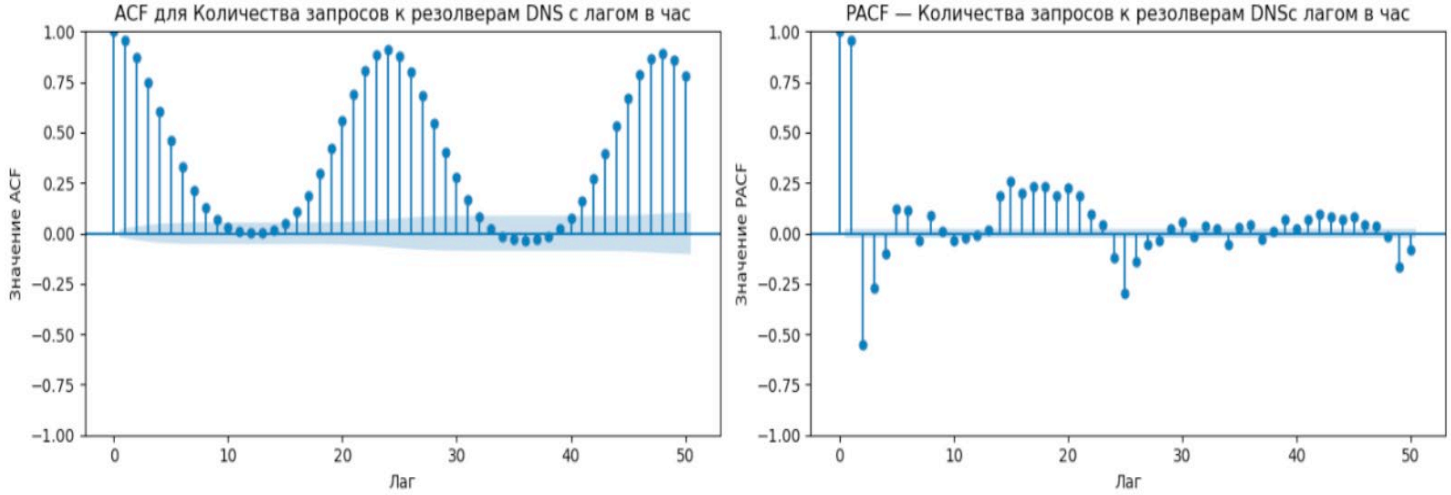
Кластеризация

- Isolation Forest

Рекуррентные нейронные сети

- Long Short-Term Memory (LSTM)

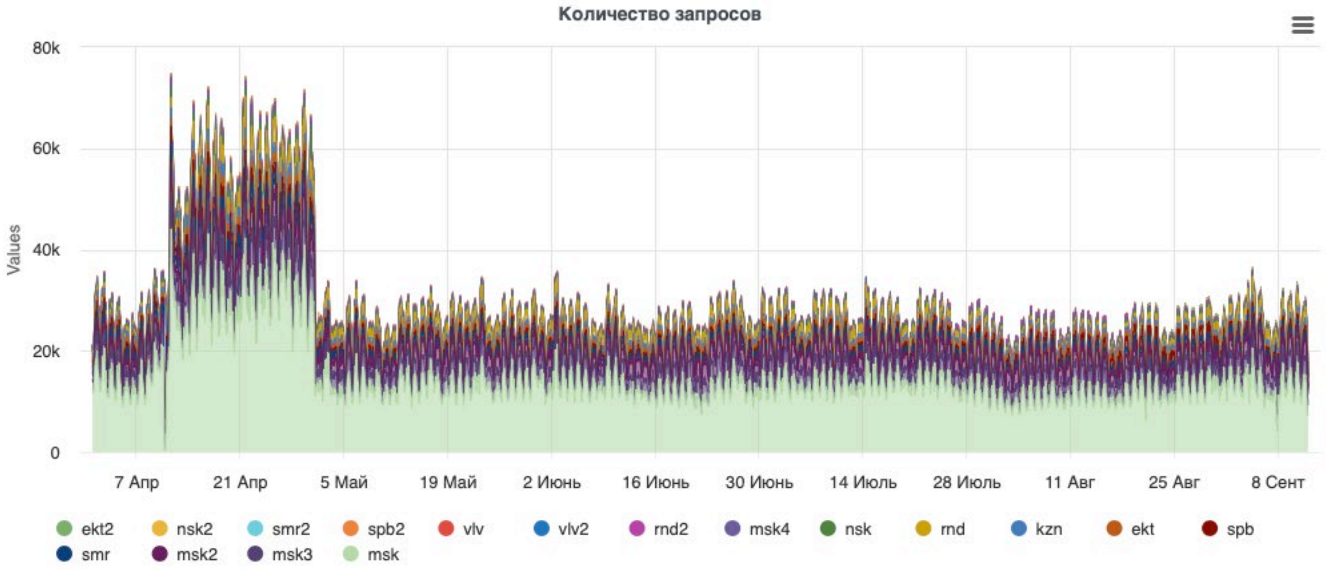
Оценка периодичности (сутки, неделя, месяц, год)
Автокорреляция (ACF) и Частичная автокорреляция (PACF)



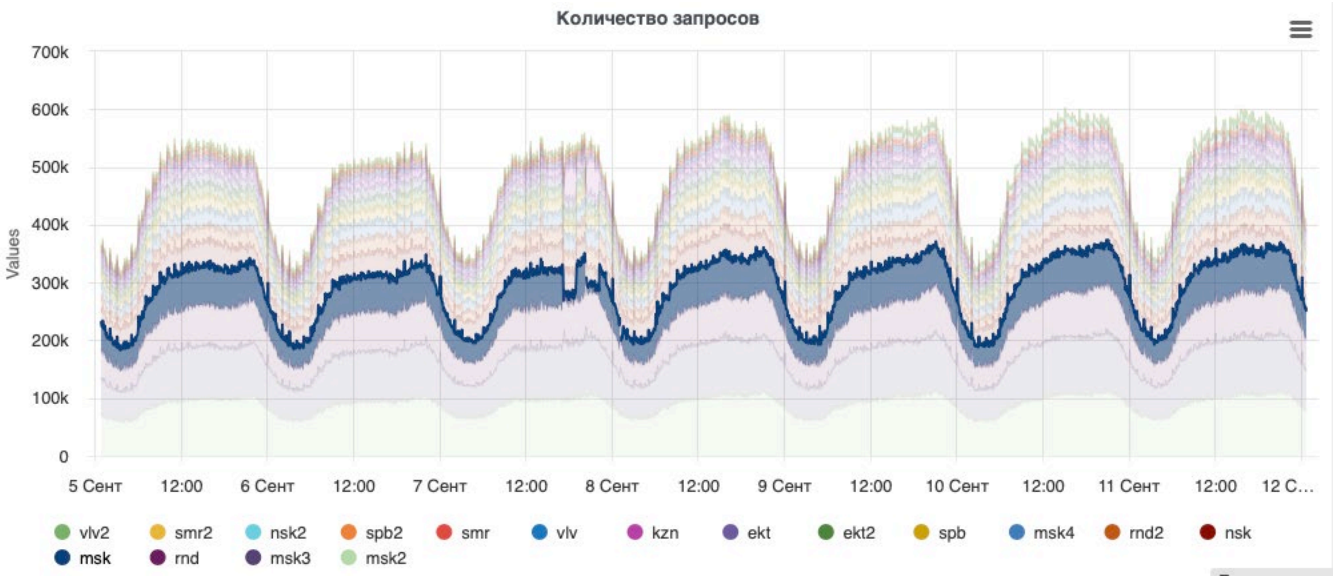
Оценка стационарности ряда*: чем выше интервал, тем ближе к стационарности

* - Тест Дики-Фуллера (ADF)

Z-Score


Количество запросов
80k
60k
40k
20k
0
Values
7 Апр
21 Апр
5 Май
19 Май
2 Июнь
16 Июнь
30 Июнь
14 Июль
28 Июль
11 Авг
25 Авг
8 Сент
ekt2
nsk2
smr2
spb2
vlv
vlv2
rnd2
msk4
nsk
rnd
kzn
ekt
spb
smr
msk2
msk3
msk

AWMA

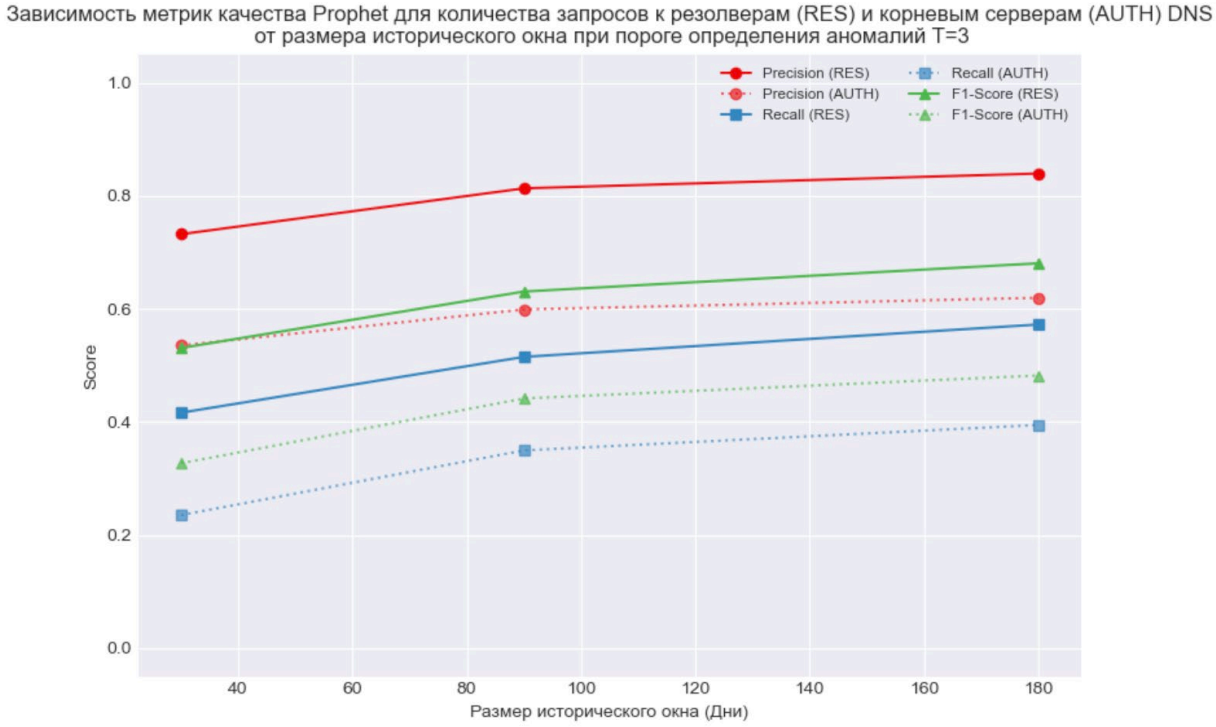

Количество запросов
700k
600k
500k
400k
300k
200k
100k
0
Values
5 Сент
12:00
6 Сент
12:00
7 Сент
12:00
8 Сент
12:00
9 Сент
12:00
10 Сент
12:00
11 Сент
12:00
12 С…
vlv2
smr2
nsk2
spb2
smr
vlv
kzn
ekt
ekt2
spb
msk4
rnd2
nsk
msk
rnd
msk3
msk2

Facebook Prophet (https://facebook.github.io/prophet/)

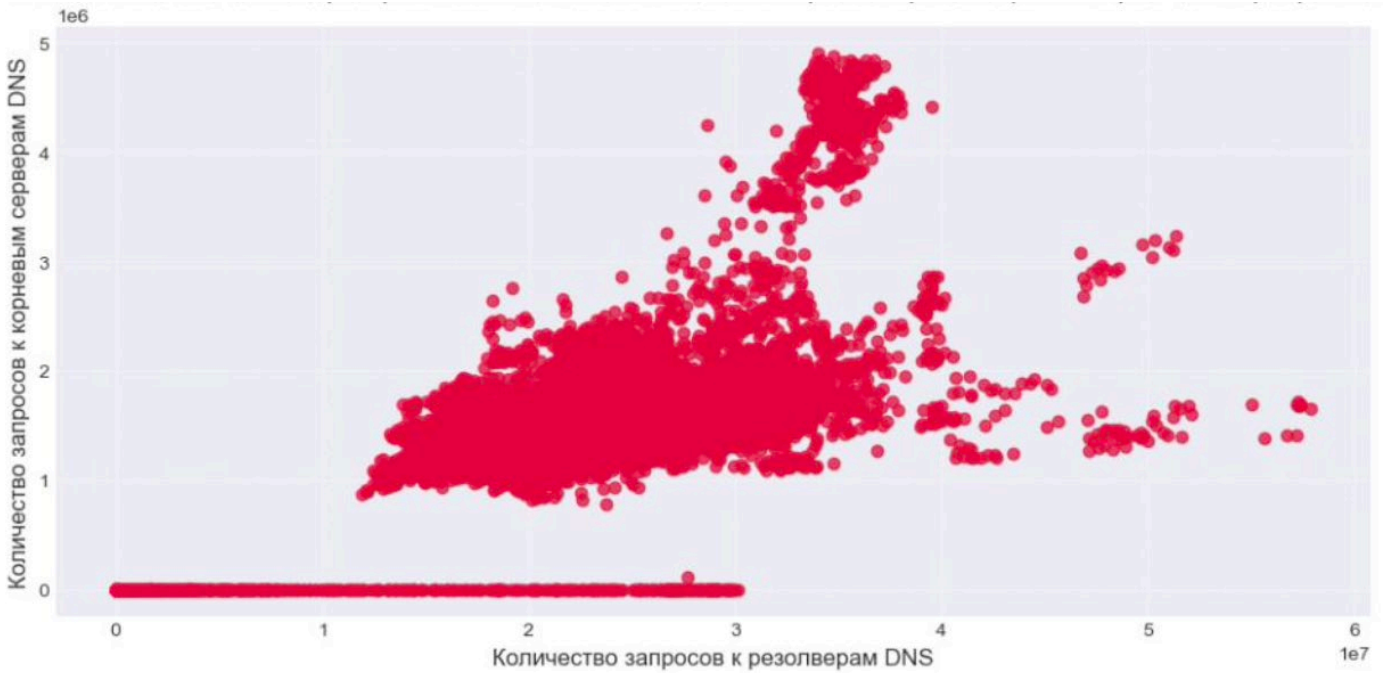
$$y(t)=g(t)+s(t)+h(t)+\varepsilon t$$

$g(t)$ – тренд;
$s(t)$ – периодичность;
$h(t)$ – праздники;
εt – случайное отклонение.

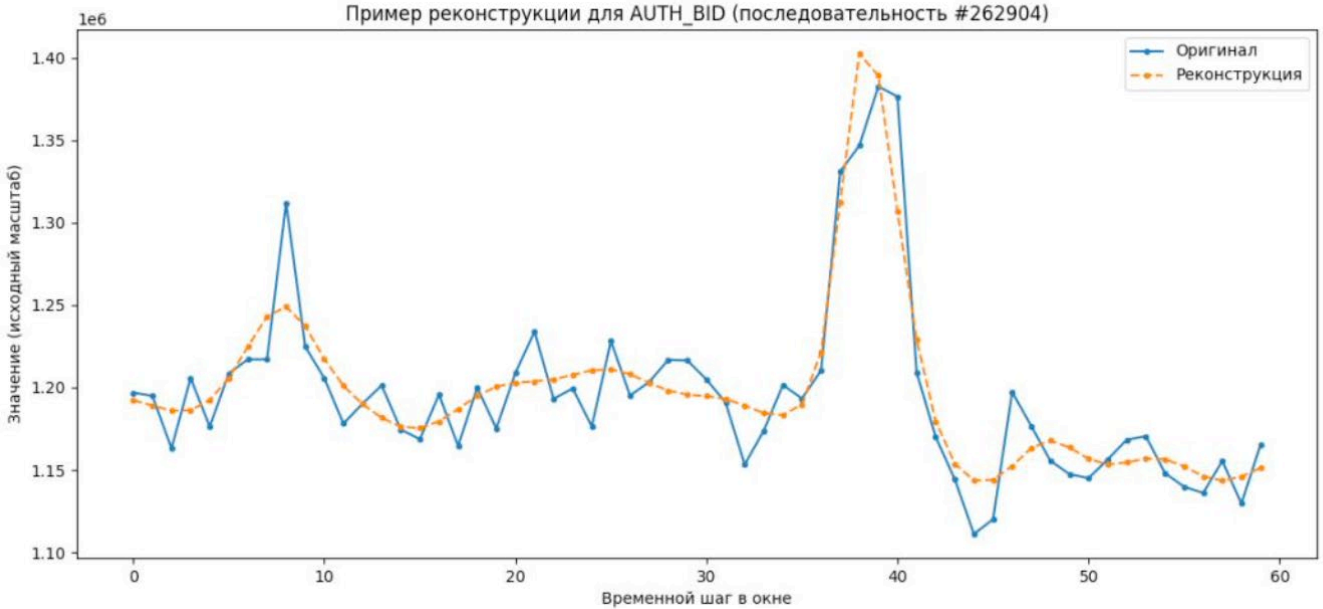
Facebook Prophet (https://facebook.github.io/prophet/)



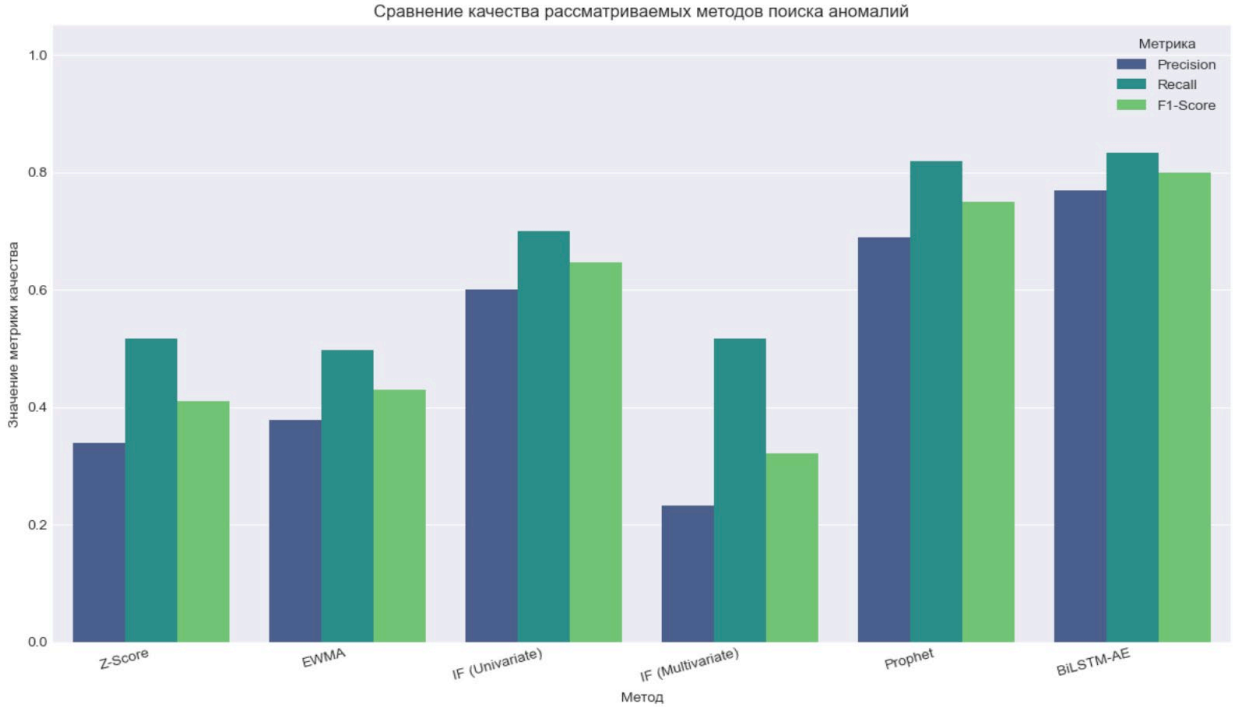
Isolation Forest



LSTM-AE


Пример реконструкции для AUTH_BID (последовательность #262904)
1e6
Оригинал
Реконструкция
Значение (исходный масштаб)
Временной шаг в окне
1.40
1.35
1.30
1.25
1.20
1.15
1.10
0
10
20
30
40
50
60

В итоге


Сравнение качества рассматриваемых методов поиска аномалий
Метрика
Precision
Recall
F1-Score
Значение метрики качества
1.0
0.8
0.6
0.4
0.2
0.0
Z-Score
EWMA
IF (Univariate)
IF (Multivariate)
Prophet
BiLSTM-AE
Метод

В итоге:

Методы, построенные на эмпирическом подборе, дают
лучшие характеристики полноты и точности при поиске аномалий.

Но во многих случаях достаточно и простого z-score

Спасибо за внимание!