RedLab/Hack

Разработка модели для выявления аномалий
во временном ряду

Команда: **ikanam_chipi_chipi**

Проблематика

Скорость работы

Требуется алгоритм, который **быстро определяет** аномалии в данных

Open-source

Необходимо использовать только **бесплатные, открытые решения**

Сезонность

Внутри дня метрики могут меняться значительно. **Изменения вызванные сезонностью – не аномалии**

Аномалия - это что?

Необходимо определить, что считается **«ненормальным» поведением метрик**

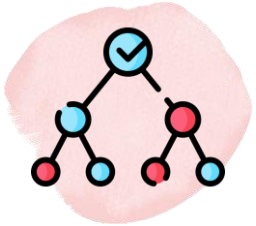
Наше решение

Интерфейс, который позволяет анализировать временной ряд и отмечать выявленные аномалии в данных.

Доступные опции:

- Загрузка .tsv файла
- Выбор временного интервала для анализа
- Интерактивные графики
- Выгрузка файла с отмеченными аномалиями

Стек



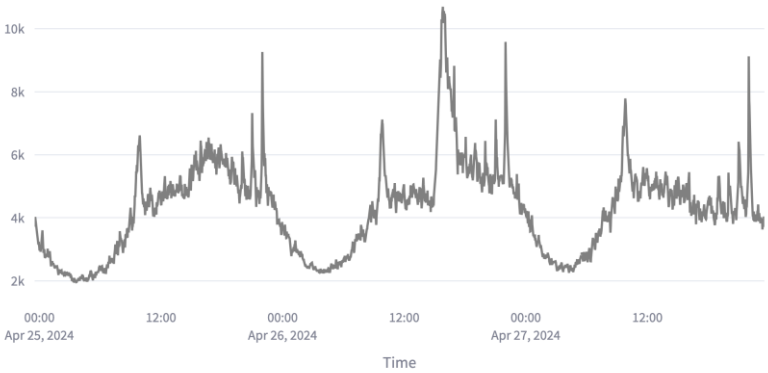
Isolation Forest

Алгоритм **машинного обучения** для поиска аномалий с учётом **нескольких признаков** одновременно

Streamlit

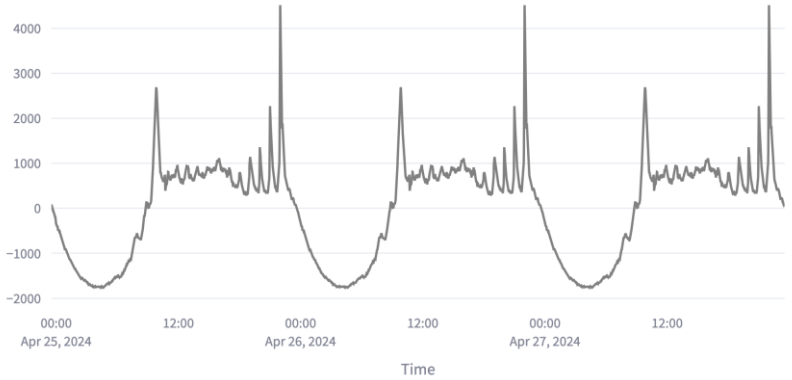
Python библиотека для создания веб-интерфейса с возможностью добавления **интерактивных графиков**

Борьба с внутридневной сезонностью



Метрика до обработки

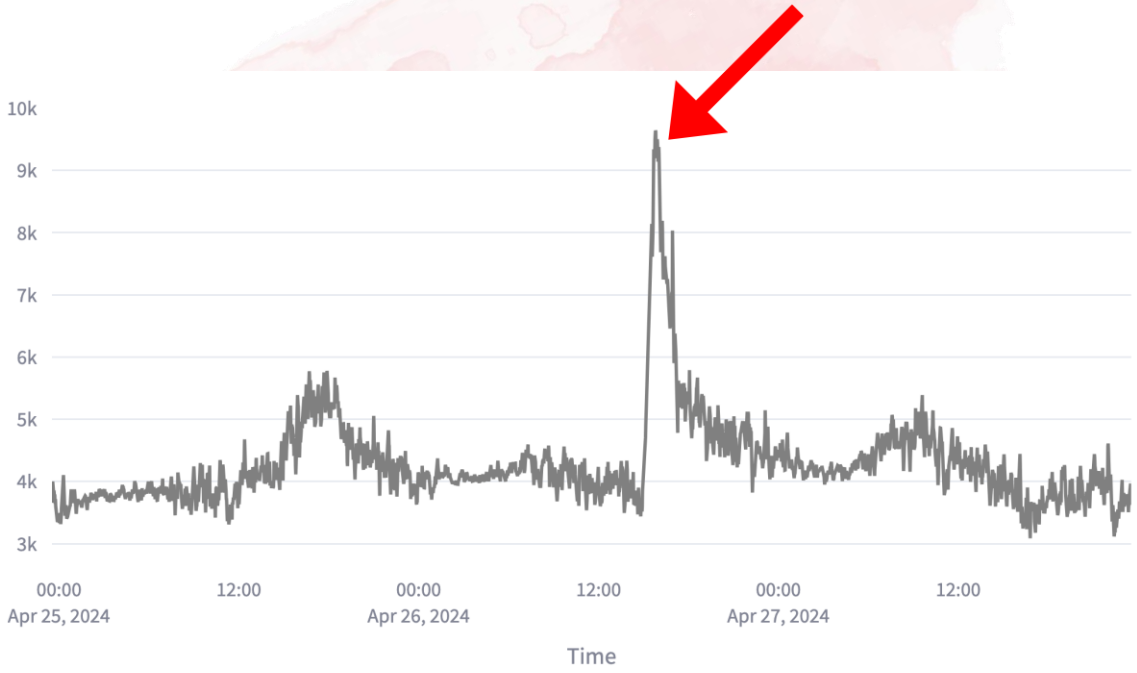
Не всегда понятно – большое значение связано со временем суток или же это аномалия?



Компонента сезонности

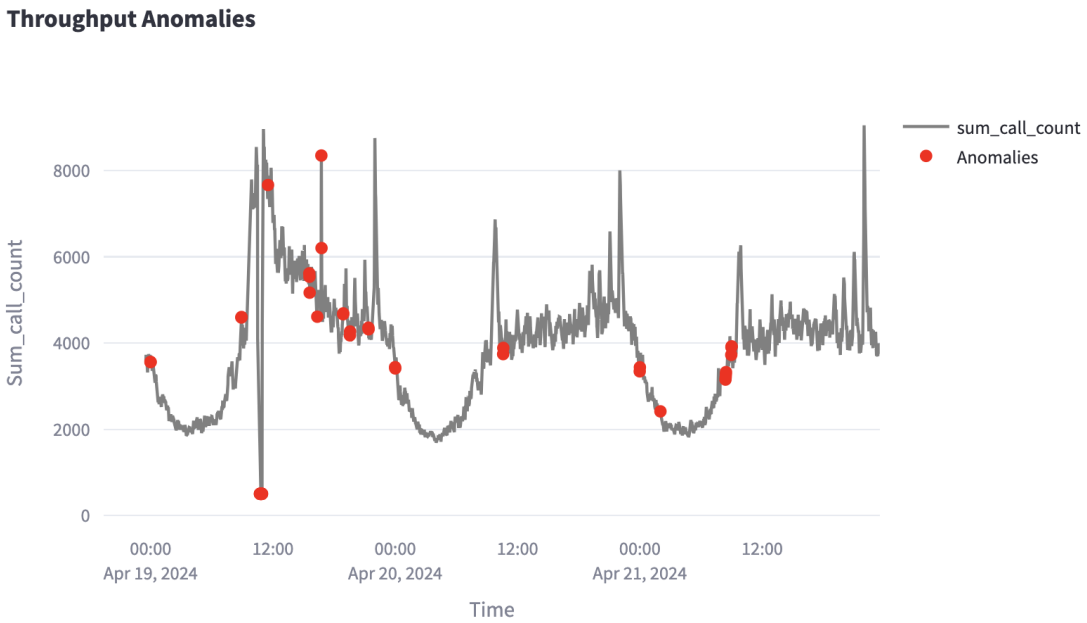
Используя библиотеку statsmodels, находим **компоненту сезонности**

Отчетливо видна аномалия!



Удалили компоненту сезонности

Не аномалии



метрика до обработки

Уникальность решения

Благодаря тому, что мы учли сезонность,
алгоритм удалил экстремальные значения не «в тупую»

Алгоритм формирует **anomaly_score** – вероятность, что объект является аномалией.

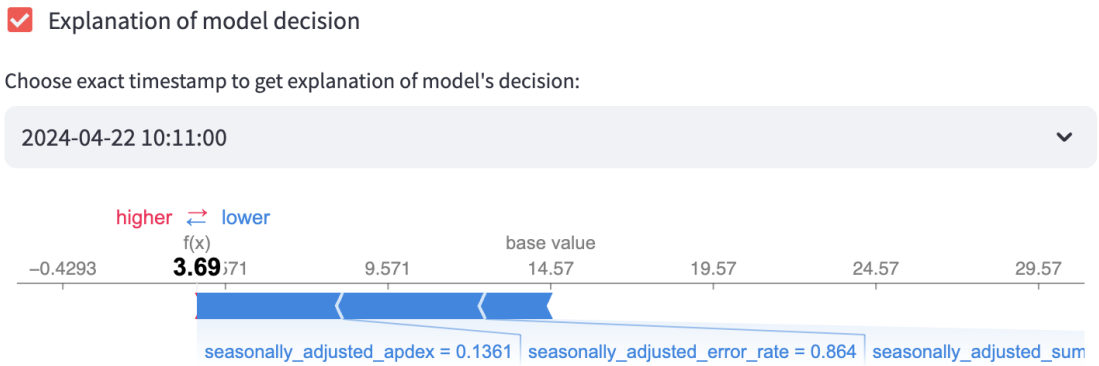
Можно задать свой трешхолд.

	point	seasonally_adjusted_sum_call_count	seasonally_adjusted_web_response	seasonally_adjusted_apdex	seasonally_adjusted_error_rate	anomaly	anomaly_score
19,531	2024-04-29 13:15:00	3,458.9023	0.0146	0.9958	0.002	☐	0.38
19,532	2024-04-29 13:16:00	3,387.0818	0.1488	0.9946	0.0028	☑	0.78
19,533	2024-04-29 13:17:00	3,392.3141	0.1535	0.9946	0.0026	☑	0.78
19,534	2024-04-29 13:18:00	3,405.3919	0.0185	0.9975	0.0003	☐	0.08

Уникальность решения

Если увеличить трешхолд – с наибольшей вероятностью **будут указаны только аномалии.**
Уменьшить – все подозрения на аномалии будут отображены.

Интерпретируемость решения



Есть возможность получить ответ на вопрос:

Руководствуясь какими факторами модель определила объект как аномалию?

Демонстрация решения

https://github.com/maxlyara1/find_anomalies_hackathon/blob/main/%D0%98%D0%BD%D1%82%D0%B5%D1%80%D1%84%D0%B5%D0%B9%D1%81_26_05_2024.mp4

Точность работы модели

TABLE 5
Results on detecting abnormal graphs and anomalies in time series.

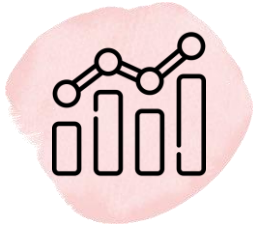
	Data	AUC-ROC					AUC-PR				
		DIF (ours)	EIF	LeSiNN	iForest	eGLocalKD	DIF (ours)	EIF	LeSiNN	iForest	eGLocalKD
Graph	HSE	**0.737**$_{\pm 0.013}$	0.715$_{\pm 0.014}$	0.702$_{\pm 0.001}$	0.697$_{\pm 0.014}$	0.593$_{\pm 0.002}$	**0.094**$_{\pm 0.005}$	0.088$_{\pm 0.004}$	0.084$_{\pm 0.000}$	0.082$_{\pm 0.004}$	0.054$_{\pm 0.000}$
	MMP	**0.715**$_{\pm 0.006}$	0.663$_{\pm 0.012}$	0.666$_{\pm 0.000}$	0.667$_{\pm 0.018}$	0.675$_{\pm 0.001}$	**0.260**$_{\pm 0.006}$	0.216$_{\pm 0.006}$	0.217$_{\pm 0.000}$	0.219$_{\pm 0.011}$	0.233$_{\pm 0.001}$
	p53	**0.680**$_{\pm 0.008}$	0.597$_{\pm 0.017}$	0.606$_{\pm 0.000}$	0.619$_{\pm 0.013}$	0.640$_{\pm 0.001}$	**0.177**$_{\pm 0.006}$	0.138$_{\pm 0.004}$	0.144$_{\pm 0.000}$	0.143$_{\pm 0.004}$	0.150$_{\pm 0.000}$
	PPAR	0.701$_{\pm 0.013}$	0.716$_{\pm 0.005}$	0.711$_{\pm 0.000}$	**0.733**$_{\pm 0.009}$	0.643$_{\pm 0.001}$	0.127$_{\pm 0.008}$	0.173$_{\pm 0.006}$	0.165$_{\pm 0.001}$	**0.208**$_{\pm 0.012}$	0.086$_{\pm 0.000}$
		DIF (ours)	EIF	LeSiNN	iForest	eTranAD	DIF (ours)	EIF	LeSiNN	iForest	eTranAD
TS	Mars	0.952$_{\pm 0.017}$	**0.980**$_{\pm 0.006}$	0.942$_{\pm 0.014}$	0.947$_{\pm 0.015}$	0.947$_{\pm 0.016}$	**0.626**$_{\pm 0.024}$	0.458$_{\pm 0.031}$	0.400$_{\pm 0.009}$	0.390$_{\pm 0.043}$	0.334$_{\pm 0.020}$
	Gait	**0.998**$_{\pm 0.001}$	0.997$_{\pm 0.001}$	0.998$_{\pm 0.000}$	0.997$_{\pm 0.001}$	0.998$_{\pm 0.000}$	**0.835**$_{\pm 0.064}$	0.772$_{\pm 0.048}$	0.829$_{\pm 0.010}$	0.741$_{\pm 0.073}$	0.806$_{\pm 0.010}$
	ECG	**0.997**$_{\pm 0.001}$	0.986$_{\pm 0.001}$	0.987$_{\pm 0.000}$	0.987$_{\pm 0.001}$	0.976$_{\pm 0.001}$	**0.809**$_{\pm 0.031}$	0.705$_{\pm 0.004}$	0.710$_{\pm 0.000}$	0.711$_{\pm 0.002}$	0.692$_{\pm 0.001}$
	ECG-w	**1.000**$_{\pm 0.000}$	0.988$_{\pm 0.001}$	0.985$_{\pm 0.001}$	0.981$_{\pm 0.001}$	0.990$_{\pm 0.000}$	**1.000**$_{\pm 0.000}$	0.255$_{\pm 0.011}$	0.219$_{\pm 0.008}$	0.181$_{\pm 0.005}$	0.297$_{\pm 0.001}$

Статья: https://arxiv.org/abs/2206.06602

Масштабируемость решения

Подключение к БД

Для использования на практике, в реальном времени есть возможность подключиться к базе данных для **прямого доступа к данным**

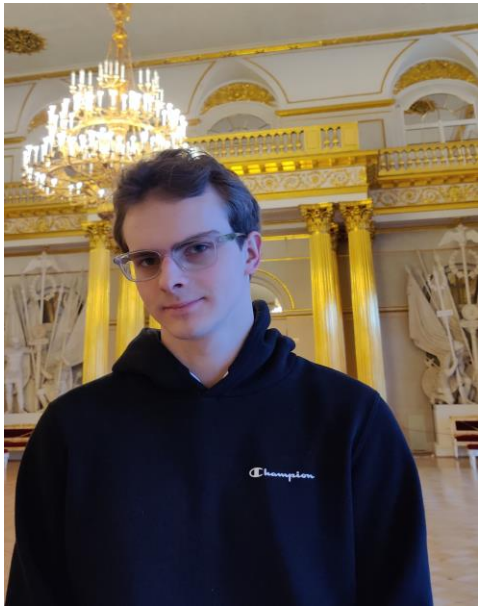


Добавление метрик

Есть возможность дополнить существующие метрики для **увеличения вероятности обнаружения сбоя в работе системы** есть

ikanam_chipi_chipi

Аделя Сабирова
Data engineer, designer



Максим Ляра
Team lead, Data scientist
TG: @maxlyara1

Станислав Палатов
Data scientist