

Хакатон по метеорологии

Решение команды “Антихайп”

Докладчики:

Дмитрий Юткин

Василий Рязанов

Хакатон по метеорологии

Решение команды “Антихайп”

Докладчики:
Дмитрий Юткин
Василий Рязанов



Команда

Дмитрий Юткин

 ODS: @0x1337

Data Scientist @ Allianz

Студент (4 курс) @ НИУ ВШЭ

<https://www.kaggle.com/yutkin>

diyutkin@edu.hse.ru

Василий Рязанов

 ODS: @ryazanoff

Data Scientist @ Allianz

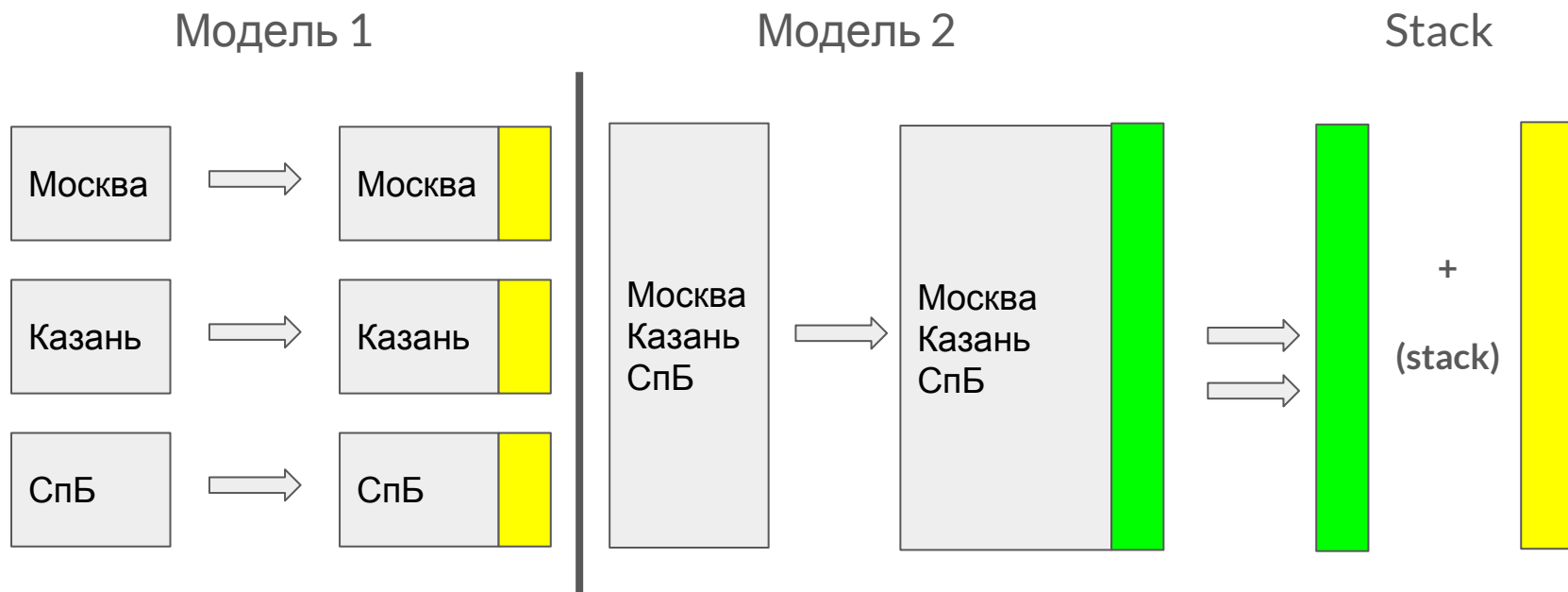
Аспирант, преподаватель @ МФТИ

<https://www.kaggle.com/ryazanoff>

vasily.ryazanov@phystech.edu

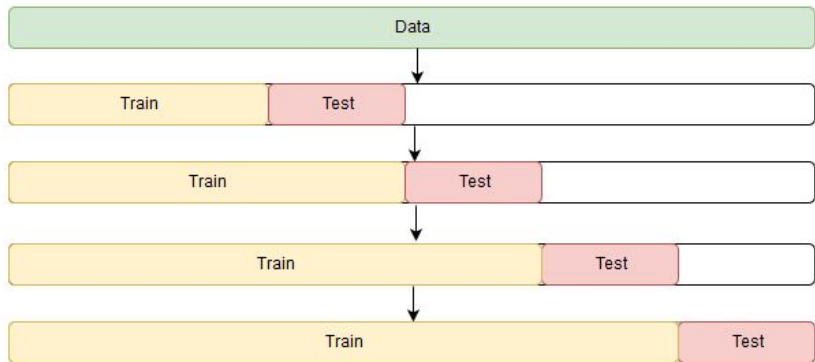
Начало работы

Первая идея: модель на всех городах + модель по каждому городу

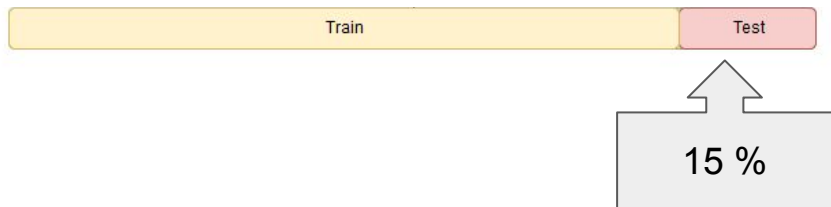


Валидация

Пробовали:



Оставили:



Считалось быстрее, хорошо коррелировало с public (в итоге с private тоже)

На LB учитывалась последняя посылка + разрешалось много посылок (100) => команды включали silent режим (сабмиты с лучшим скором сразу скрывали).

Обзор решения

Software:

Ubuntu 16.04, Python 3.5, SKlearn, CatBoost.

Hardware:

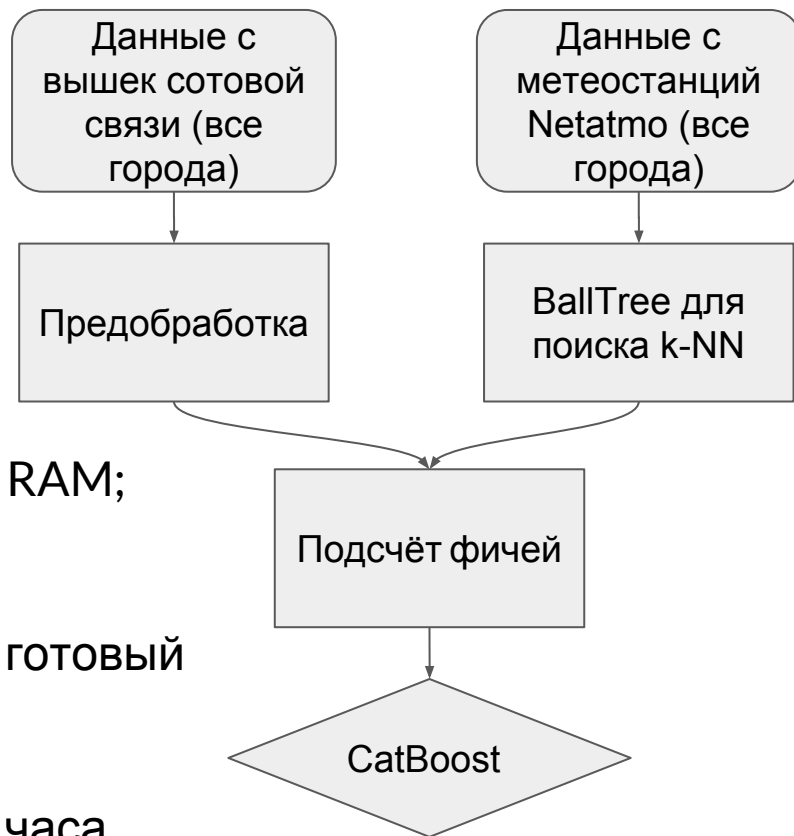
Сервер на AWS: 16 CPU ядер, 128 Gb RAM;

Сервер на Google Cloud: 12 CPU ядер, 64 Gb RAM;

Home Workstation: 6 CPU ядер, 32 Gb RAM.

С нуля pipeline не писали, модифицировали готовый baseline.

Время, затрачиваемое на весь pipeline: ~2.5 часа.



Модель 1

Замена выбросов и “-999” на NaN

Удаление дубликатов (~45% данных)

Признаки:

- Статистики (mean, max, quantile, std, ...) погодных данных по ближайшим станциям netatmo (1, 2, 3, 5, 10, ...)
- Статистики Netatmo по всему городу (если в среднем идет дождь, то скорее всего и в данном квадрате идет дождь)
- Статистики сигнала по всему городу
- Статистики сигнала в срезе по 4 крупным операторам
- Статистики сигнала в срезе по типу сигнала (GSM, LTE, ...)

Погодные данные Netatmo: давление, температура, влажность, скорость ветра и др.

Признаки где много NaN (расстояние до вышки, ...) по возможности игнорировал, так как не было время качественно с ними поработать.

ROC AUC модели: ~0.79

Модель 2: предобработка данных

В признаках “LocationSpeed”, “LocationAltitude”, “range”, “LocationDirection”, “cell_lon”, “cell_lat”, “SignalStrength”, “OperatorID” выбросы и “-999” были заменены на NaN.

На координатах метеостанций Netatmo было построено BallTree для быстрого поиска ближайших станций. Метрика близости – haversine.

Модель 2: признаки

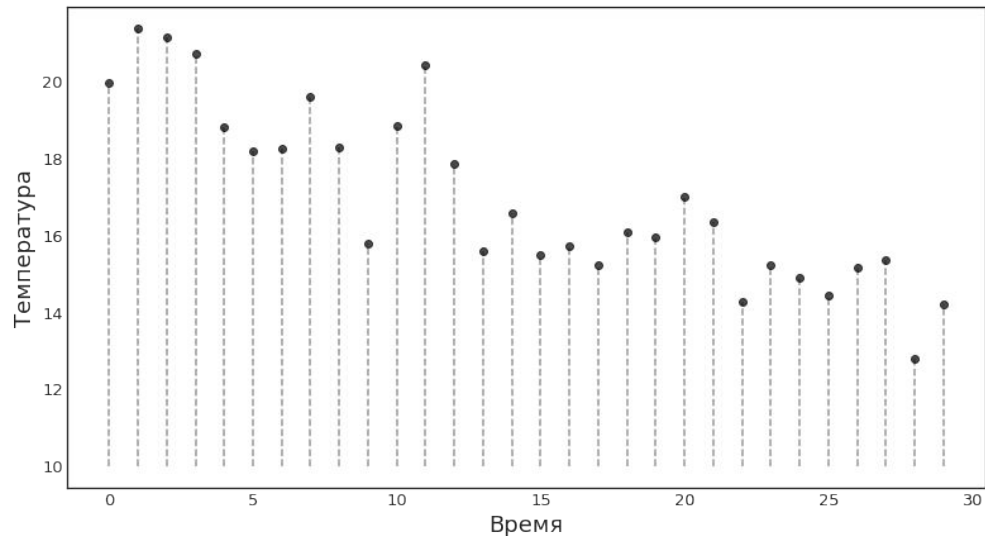
Квадрат: координаты, час дня, доля операторов, число устройств;

Данные с вышек: высота, скорость, расстояние до телефона, мощность сигнала (+ тоже самое с агрегацией по устройствам) - по всем этим величинам считаются различные статистики;

Данные с Netatmo: статистики по погодным данным с соседних станций, расстояние до ближайшей/удаленной вышки, статистики по удаленности до соседних вышек,

+ изменение погодных показателей во времени (тенденции).

Вычисление временных изменений (пример)



Например, есть данные о температуре, отсортированные по времени:
25, 24, 24, 22, 23, 21, 20, 20, 19.

- 1) Вычтем из i -го показания $(i-1)$ -ое:
NaN, -1, 0, -2, 1, -2, -1, 0, -1;
- 2) Посчитаем различные статистики от полученного ряда, например, среднее равно -0.75.

ROC AUC модели: ~0.825.


CatBoost vs. LightGBM

Пробовали LightGBM и CatBoost:

	train	valid	LB
LightGBM	0.91	0.81	0.78 (???)
CatBoost	0.92	0.81	0.82



Что не успели

1. Регрессия: признак `rain == (precipitation >= 0.25)` => можно предсказывать не `rain (0/1)` а `precipitation`;
2. Отдельные модели по городам;
3. HyperOpt, stacking;
4. Сгенерить больше фичей;
5. Нейронные сети. 

Решение: <https://gist.github.com/anonymous/059c4744294f5dbe1e479006c54686c2>