Хакатон по метеорологии

Алексей Харламов
Павел Остяков

Яндекс

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
ВЫСШАЯ ШКОЛА ЭКОНОМИКИ

#	Участник	Хакатон Баллы
1	10011000 (pavelostaaa, MrAxcel)	84489992
2	Антихайп (cszdr, vasyarv)	82909074
3	Classics(ivanicki-i, Amir14111)	80423665
4	kutcu	79701913
5	sal.vios	78750097
6	alexisonfireyar	75905679
7	Ясонов Евгений	75734019
8	Антон Патрикеев	75606736
9	Aiaiai (karfly, illusionww)	74950292
10	Iviconun(FireSonics, mrk.andreev)	74761301

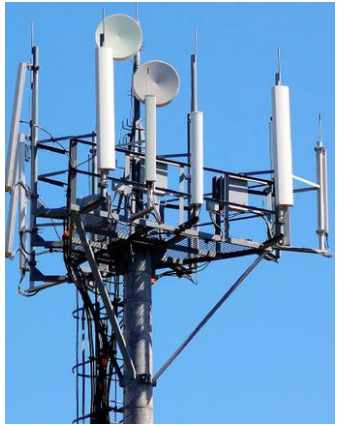
Задача

- Даны три города: Москва, Санкт-Петербург и Казань
- Каждый город разбит на квадраты
- Есть данные с любительских метеостанций и данные о качестве связи у пользователей
- Для каждого квадрата и часа нужно восстановить, шёл дождь или нет
- Тренировочные данные – начало июля 2017. Тестовые – июль-август 2017.

Что нужно было предсказать?

- Для каждого квадрата было известно precipitation - количество мм осадков, выпавших на конец часа
- Таргет: precipitation > 0.25
- Бинарная классификация
- Метрика - ROC AUC

Данные

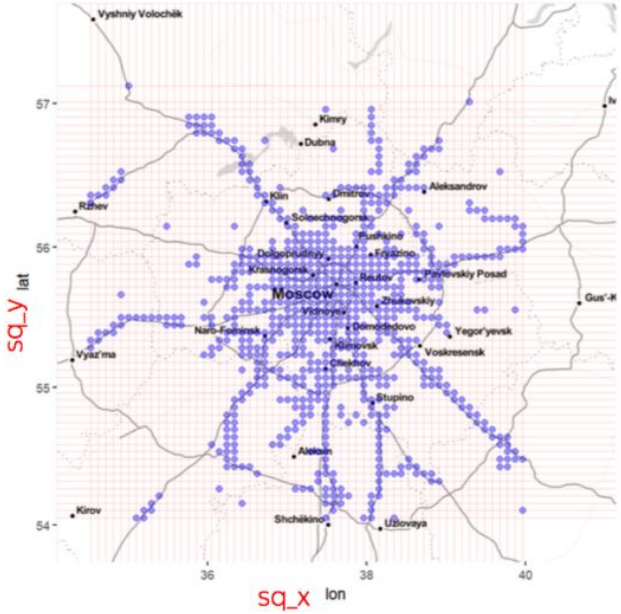


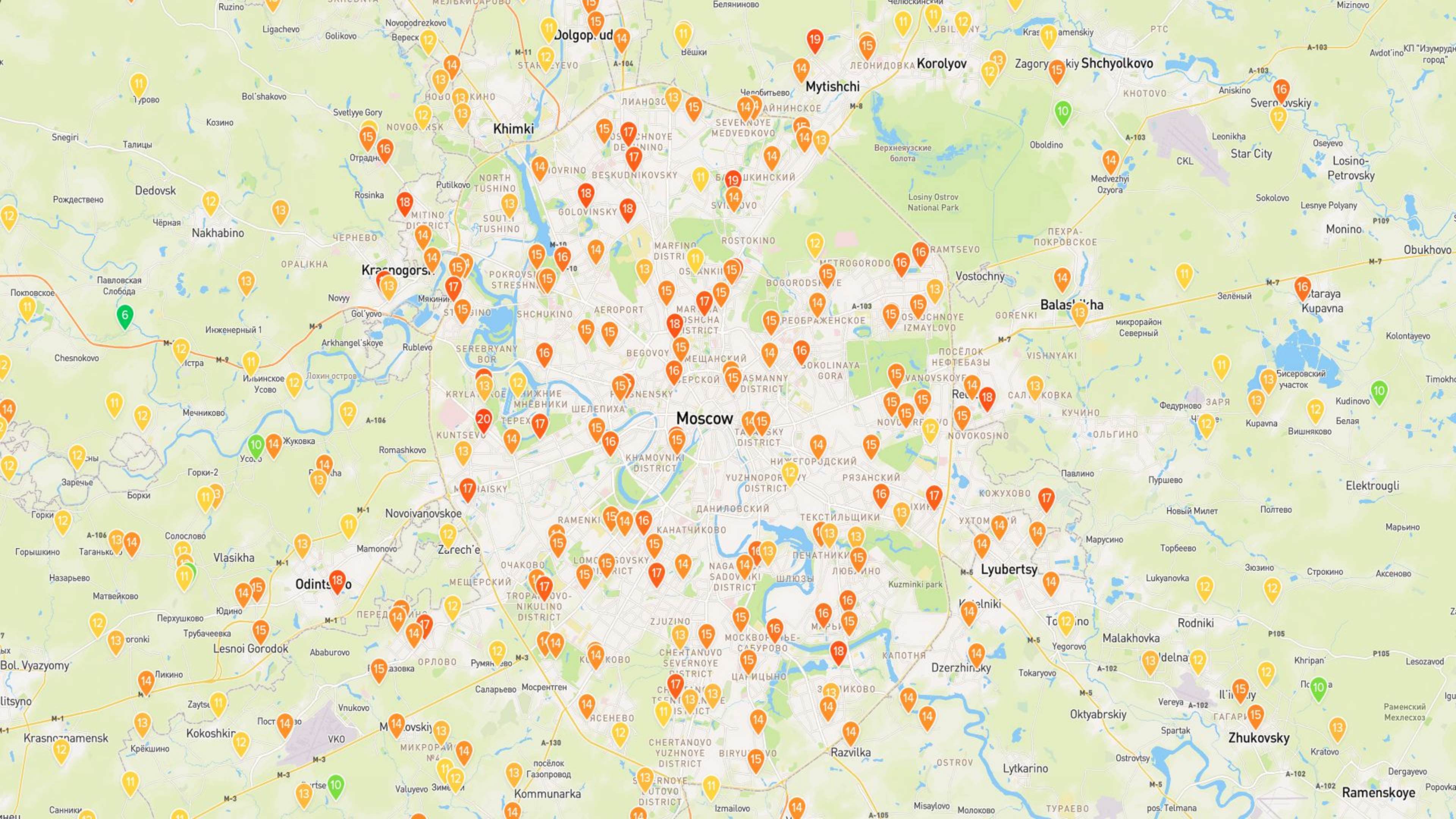
- Порядка 30ГБ
- 2 типа: Пользовательские и Netatmo
- Поделены по времени(час)

Семплированные (по Москве 0.1% пользователей), обфусцированные данные из Аппметрики + практически сырые данные любительских метеостанций в Москве, Санкт-Петербурге и Казани.

Проблемы

- Квадраты распределены неравномерно
- Данные от пользователей очень шумные
- Много пропусков
- Данных мало




Ruzino
Ligachevo
Golikovo
Novopodrezkovo
Вереск
Dolgoprud
Белянинове
Mizinovo
Туровo
Bol'shakovo
Svetlyye Gory
Козино
Snegiri
Талицы
Novogorsk
Khimki
Отрадное
Rosinka
Putilkovo
North Tushino
South Tushino
Mitino District
Dedovsk
Рождествено
Чёрная
Nakhabino
Чернево
Opalikha
Krasnogorsk
Павловская Слобода
Покровское
Novyy
Gol'yovo
Мякинино
Инженерный 1
Arkhangel'skoye
Rublevo
Chesnokovo
Истра
Ильинское Усово
Лохин остров
Serebryany Bor
Мечниково
Жуковка
Усово
Romashkovo
Горки-2
Заречье
Борки
Горки
Горышкино
Таганьково
Солослово
Vlasikha
Назарьево
Матвейково
Odintsovo
Novoivanovskoe
Mamonovo
Zarech'e
Юдино
Перхушково
Трубачеевка
Lesnoi Gorodok
Ababurovo
Bol. Vyazyomy
Ликино
Zaytsevo
Vnukovo
Kokoshkino
Krasnoznamensk
Крёкшино
VKO
Valuyevo
Санники
Mytishchi
Korolyov
Челобитьево
Леонидовка
Лианозово
Медведково
Северное Medvedkovo
Beskudnikovsky
Golovinsky
Свиблово
Верхнеяузские болота
Losiny Ostrov National Park
Rostokino
Marfino District
Ostankino
Pokrovskoye Streshnevo
Shchukino
Aeroport
Maryina Roshcha District
Begovoy
Мещанский
Преображенское
Sokolinaya Gora
Bogorodskoye
Metrogorodok
Гольяново
Vostochny
Krylatskoe
Нижние Мнёвники
Шелепиха
Presnensky
Krasnoselsky
Basmanny District
Moscow
Taganskiy District
Kuntsevo
Khamovniki District
Нижегородский
Yuzhnoporty District
Рязанский
Даниловский
Текстильщики
Ramenki
Канатчиково
Lomonosovsky District
Nagatino-Sadovniki District
Шлюзы
Печатники
Люблино
Kuzminki park
Очаково
Мещерский
Troparevo-Nikulino District
Zjuzino
Москворечье-Сабурово
Chertanovo Severnoye District
Царицыно
Марьино
Капотня
Ясенево
Chertanovo Yuzhnoye District
Biryulyovo
Razvilka
Ostrov
Мосрентген
Саларьево
Румянцево
Орлово
Переделкино
Moskovskiy
Микрорайон №4
Зименки
посёлок Газопровод
Kommunarka
Северное Бутово District
Izmailovo
Misaylovo
Молоково
Zagoryanskiy
Shchyolkovo
Yubileyny
Krasnoznamenskiy
Khotovo
Oboldino
PTC
Aniskino
Sverdlovskiy
Avdot'ino
КП "Изумрудный город"
Leonikha
Star City
CKL
Oseyevo
Losino-Petrovsky
Medvezhyi Ozyora
Sokolovo
Lesnye Polyany
Monino
Пехра-Покровское
Obukhovo
Balashikha
Зелёный
Staraya Kupavna
Gorenki
микрорайон Северный
Kolontayevo
Посёлок Нефтебазы
Vishnyaki
Бисеровский участок
Timokhovo
Ivanovskoye
Reutov
Салтыковка
Федурново
Заря
Кучино
Kudinovo
Novogireevo
Novokosino
Ольгино
Kupavna
Вишняково
Белая
Павлино
Пуршево
Elektrougli
Кожухово
Новый Милет
Полтево
Ухтомский
Марусино
Торбеево
Марьино
Lyubertsy
Зюзино
Строкино
Аксеново
Lukyanovka
Kotelniki
Tomilino
Yegorovo
Rodniki
Malakhovka
Dzerzhinsky
Tokaryovo
Udelnaya
Khripan'
Lesozavod
Il'inskiy
Vereya
Гагарина
Раменский Мехлесхоз
Oktyabrskiy
Spartak
Zhukovsky
Kratovo
Ostrovtsy
Lytkarino
Dergayevo
Ramenskoye
Popovka
Тураево
pos. Telmana
M-1
M-3
M-5
M-7
M-8
M-9
M-10
M-11
A-102
A-103
A-104
A-105
A-106
A-130
P105
P109

Пользовательские факторы

- Данные о сотовом сигнале (сила сигнала, тип, оператор)
- GPS данные (высота, скорость движения, точность определения)
- Координаты пользователя
- Модель телефона

Данные с любительских метеостанций

Метеоданные с шагом в 20 минут

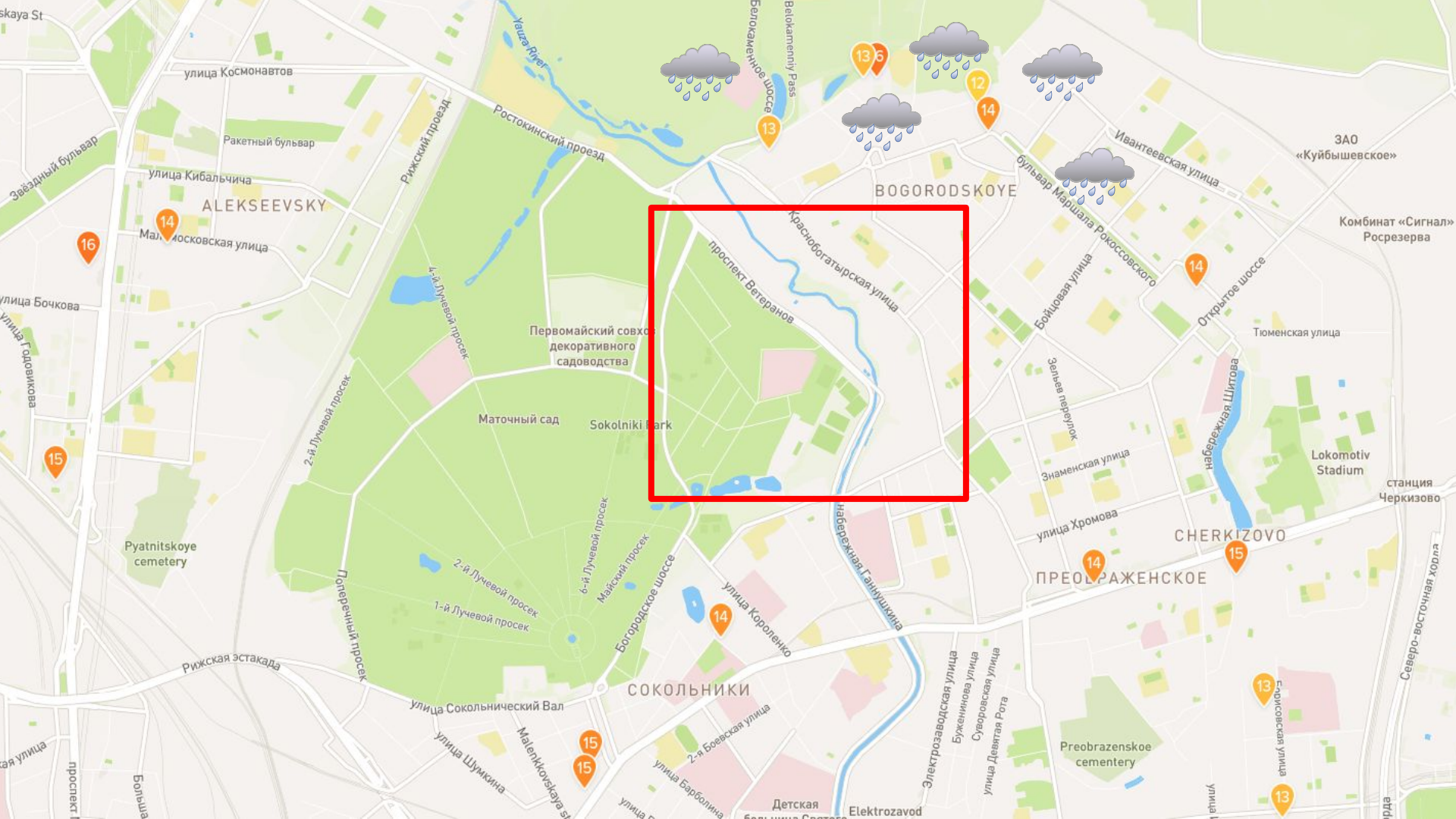
- Координаты метеостанции
- Температура
- Скорость и направление ветра
- Влажность
- Количество осадков за последний час

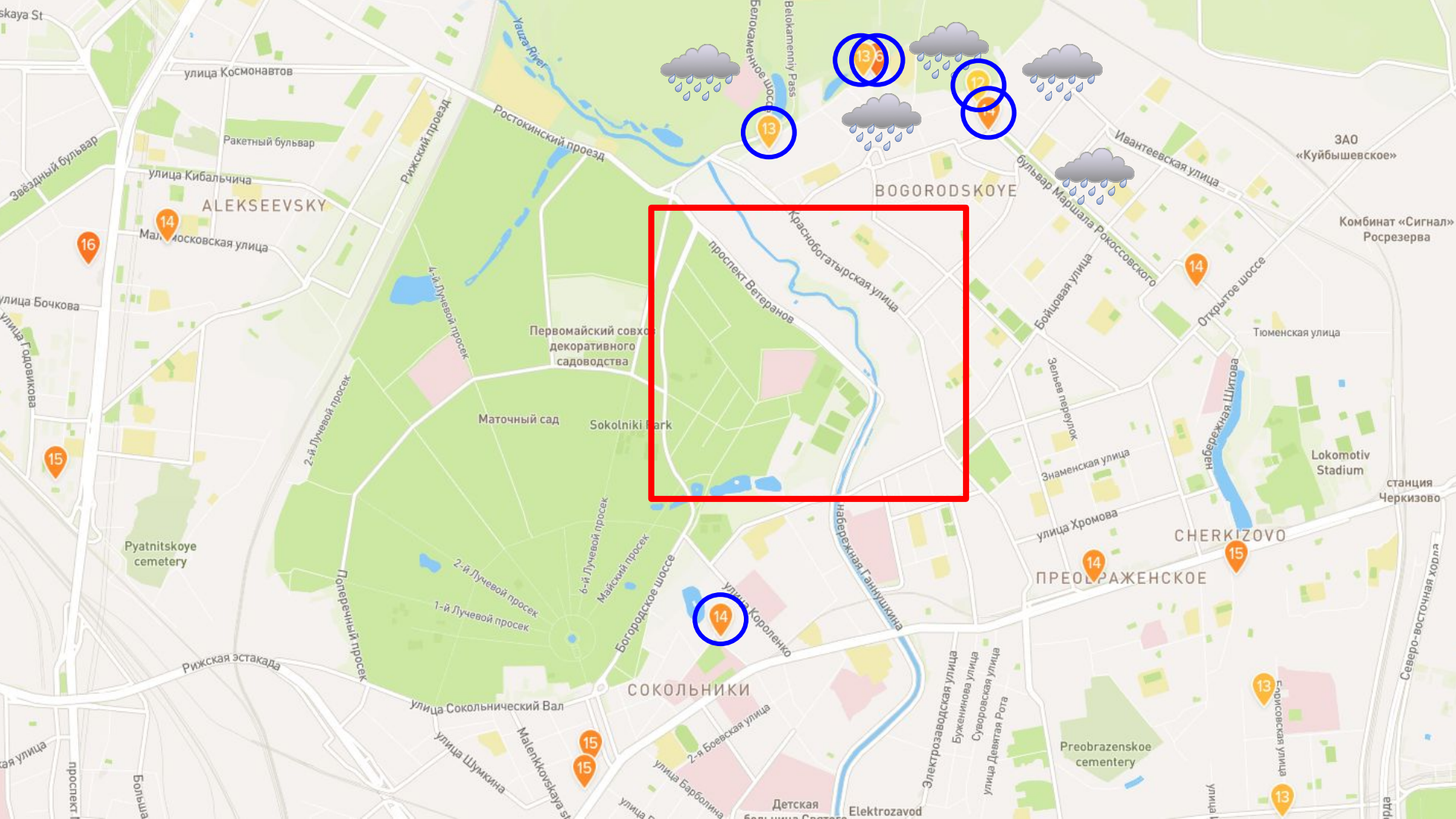
Что сделали: факторы

- Базовые статистики для признаков(mean, var, min, max, etc.)
- 10 квантилей (0.1, 0.2, …)
- Статистики для различных подвыборок ближайших метеостанций.
- Всего около 3000 факторов

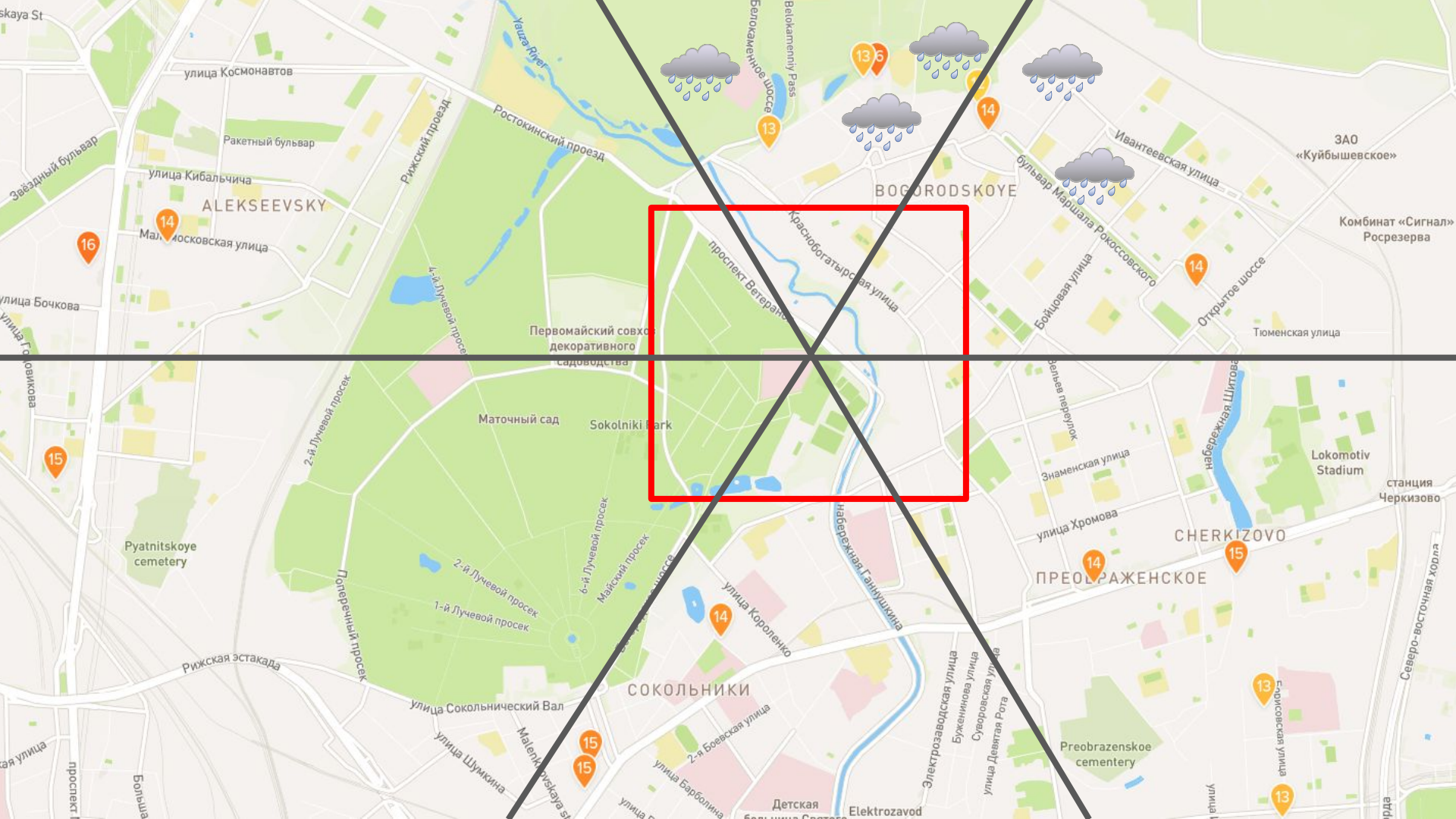
Что сделали: модель

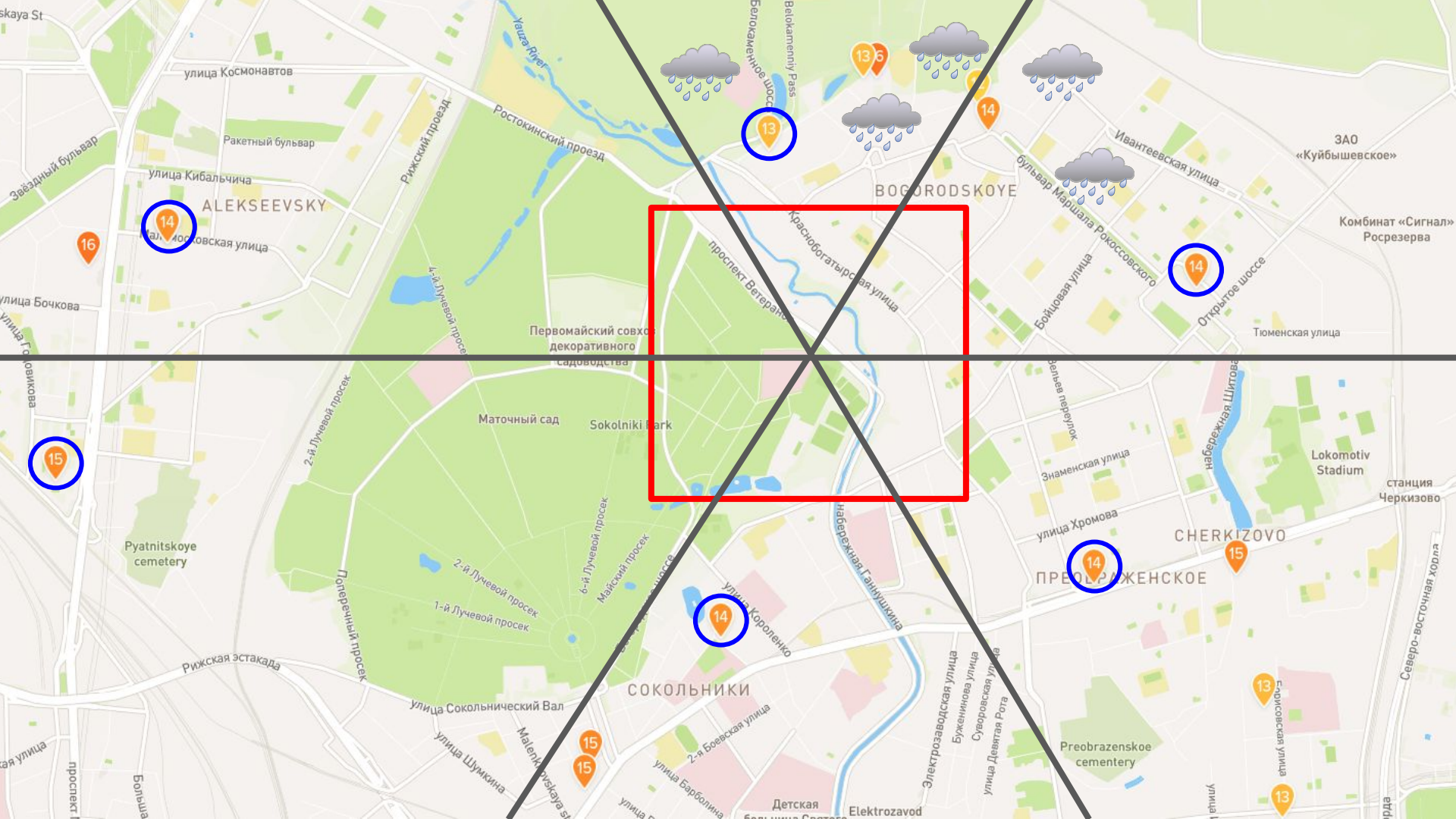
- Одна модель для всех городов.
- Catboost на дефолтных параметрах
- Получили 0.82 на лидерборде


Yauza River
Белокаменное шоссе
Belokamennyy Pass
улица Космонавтов
Ростокинский проезд
Рижский проезд
Звёздный бульвар
Ракетный бульвар
улица Кибальчича
ALEKSEEVSKY
Малая Московская улица
улица Бочкова
улица Годовикова
Первомайский совхоз декоративного садоводства
4-й Лучевой просек
2-й Лучевой просек
Маточный сад
Sokolniki Park
Pyatnitskoye cemetery
Поперечный просек
2-й Лучевой просек
1-й Лучевой просек
6-й Лучевой просек
Майский просек
Богородское шоссе
Рижская эстакада
улица Сокольнический Вал
улица Шумкина
Malenkovskaya st
СОКОЛЬНИКИ
улица Барболина
2-я Боевая улица
Детская
Elektrozavod
BOGORODSKOYE
проспект Ветеранов
Красноборатырская улица
Ивантеевская улица
бульвар Маршала Рокоссовского
Бойцовая улица
Открытое шоссе
ЗАО «Куйбышевское»
Комбинат «Сигнал» Росрезерва
Тюменская улица
Зельев переулок
Знаменская улица
улица Хромова
набережная Шитова
набережная Гаянушкина
улица Короленко
Lokomotiv Stadium
станция Черкизово
CHERKIZOVO
ПРЕОБРАЖЕНСКОЕ
Электрозаводская улица
Буженинова улица
Суворовская улица
улица Девятая Рота
Preobrazenskoe cemetery
Барсиковская улица
Северо-восточная хорда


Yauza River
Белокаменное шоссе
Belokamennyy Pass
улица Космонавтов
Звёздный бульвар
Ракетный бульвар
улица Кибальчича
ALEKSEEVSKY
Малая Московская улица
Рижский проезд
Ростокинский проезд
BOGORODSKOYE
Ивантеевская улица
бульвар Маршала Рокоссовского
ЗАО «Куйбышевское»
Комбинат «Сигнал» Росрезерва
Открытое шоссе
Бойцовая улица
Красноватырская улица
проспект Ветеранов
Тюменская улица
Зельев переулок
Набережная Шитова
Знаменская улица
Lokomotiv Stadium
станция Черкизово
улица Хромова
CHERKIZOVO
ПРЕОБРАЖЕНСКОЕ
Первомайский совхоз декоративного садоводства
Маточный сад
Sokolniki Park
4-й Лучевой просек
2-й Лучевой просек
1-й Лучевой просек
6-й Лучевой просек
Майский просек
Поперечный просек
Богородское шоссе
Набережная Гаянушкина
улица Короленко
Pyatnitskoye cemetery
Рижская эстакада
улица Сокольнический Вал
СОКОЛЬНИКИ
улица Шумкина
Malenkovskaya st
2-я Боевая улица
улица Барболина
Детская
Elektrozavod
Электрозаводская улица
Буженинова улица
Суворовская улица
улица Девятая Рота
Preobrazenskoe cemetery
Барсовская улица
Северо-восточная хорда
13
14
15
16

Можно добавлять не просто блажайшие, а брать равномерно вокруг


улица Космонавтов
Ракетный бульвар
улица Кибальчича
ALEKSEEVSKY
Малая Московская улица
Звёздный бульвар
улица Бочкова
Рижский проезд
Ростокинский проезд
Yauza River
Белокаменное шоссе
Belokamennyy Pass
BOGORODSKOYE
Ивантеевская улица
бульвар Маршала Рокоссовского
ЗАО «Куйбышевское»
Комбинат «Сигнал» Росрезерва
Открытое шоссе
Бойцовая улица
Красноботатырская улица
проспект Ветеранов
Тюменская улица
Первомайский совхоз декоративного садоводства
4-й Лучевой просек
2-й Лучевой просек
Маточный сад
Sokolniki Park
Зельев переулок
Набережная Шитова
Знаменская улица
Lokomotiv Stadium
станция Черкизово
улица Хромова
CHERKIZOVO
ПРЕОБРАЖЕНСКОЕ
Pyatnitskoye cemetery
Поперечный просек
2-й Лучевой просек
1-й Лучевой просек
6-й Лучевой просек
Майский просек
набережная Гаянушкина
улица Короленко
Рижская эстакада
улица Сокольнический Вал
СОКОЛЬНИКИ
улица Шумкина
2-я Боевая улица
улица Барболина
Электрозаводская улица
Буженинова улица
Суворовская улица
улица Девятая Рота
Preobrazenskoe cemetery
Елисеевская улица
Северо-восточная хорда
Детская
Elektrozavod


Yauza River
Белокаменное шоссе
Belokamennyy Pass
улица Космонавтов
Ракетный бульвар
Звёздный бульвар
улица Кибальчича
Ростокинский проезд
Рижский проезд
ALEKSEEVSKY
Ярославское шоссе
улица Бочкова
BOGORODSKOYE
Ивантеевская улица
бульвар Маршала Рокоссовского
ЗАО «Куйбышевское»
Комбинат «Сигнал» Росрезерва
Открытое шоссе
Бойцовая улица
Красноборатырская улица
проспект Ветеранов
Тюменская улица
Первомайский совхоз декоративного садоводства
Маточный сад
Sokolniki Park
2-й Лучевой просек
4-й Лучевой просек
6-й Лучевой просек
Майский просек
1-й Лучевой просек
Поперечный просек
Pyatnitskoye cemetery
Рижская эстакада
улица Сокольнический Вал
улица Шумкина
Malenkovskaya st
2-я Боевая улица
улица Барболина
Детская больница Святого
Elektrozavod
СОКОЛЬНИКИ
набережная Ганнушкина
улица Короленко
Электрозаводская улица
Буженинова улица
Суворовская улица
улица Девятая Рота
Зельев переулок
Знаменская улица
набережная Шитова
Lokomotiv Stadium
станция Черкизово
улица Хромова
CHERKIZOVO
ПРЕОБРАЖЕНСКОЕ
Преображенское кладбище
Preobrazenskoe cemetery
Гиляровская улица
Северо-восточная хорда

Снова считаем различные статистики, получаем ещё 4000 фичей и 0.842 на лидерборде

Какие модели пробовали обучать?

- Линейные
- Метрические
- Нейронные сети
- Деревья
- Градиентный бустинг над решающими деревьями

Как и всегда, градиентный бустинг сработал, а всё остальное нет

Yandex
CatBoost

dmlc
XGBoost

Microsoft
LightGBM

Yandex
CatBoost

dmlc
XGBoost

Microsoft
LightGBM

Yandex
CatBoost

dmlc
XGBoost

Microsoft
LightGBM

Признаки внесшие наибольший вклад

- time_of_day
- netatmo_pressure_mbar
- netatmo_sum_rain_1h
- netatmo_sum_rain_24h
- netatmo_humidity_percent
- netatmo_temperature_c
- netatmo_wind_direction
- LocationSpeed

На чем обучали?

- Два процессора Intel xeon по 14 ядер
- 256 ГБ ОЗУ
- Время генерации фичей - 2 часа
- Время обучения catboost - 10 минут

Возможные улучшения

- Чистка данных
- Оптимизация ROC AUC directly
- Более явный учет информации с ближайших квадратов
- Больший упор на пользовательские признаки
- Явная дифференциация информации по операторам сотовой связи
- Стекинг, блендинг…
- Смена таргета (переход к задаче регрессии)

Ссылки

Воспроизводимый код - github.com/gamers5a/YandexMeteumSolution

Алексей Харламов - axcel@me.com

Павел Остяков - pavelosta@gmail.com