Как предсказать количество звонков в экстренные службы в г.Москве?

Что в задаче нам дано
Есть датасет с координатами зон и данные о среднем количестве звонков для некоторых из этих зон.
Всего в датасет 22172 зоны, 4374 отмечены как целевые в трейне.
Для наглядности длина ширина зоны в километрах составляет.., что примерно…

Задача:
«предсказать вызовы экстренных служб для восточной половины» Москвы

Что такое вызов экстренных служб?
Для начала надо разобраться, в каких ситуациях звонят в экстренные службы по номеру 112, и окажется, что номер объединяет 6 служб:
• пожарной охрана (01)
• реагирования в чрезвычайных ситуациях (МЧС)
• полиция (02)
• скорая медицинская помощь (03)
• аварийная служба газовой сети (04)
• «Антитеррор».
Наиболее часто предположительно вызывают 3 службы «скорую», «пожарных» и «полицию». Причем следует заметить, что спектр проблем, по которым звонят в полицию наиболее широкий - дорожно-транспортное происшествие; домашнее насилие; преступления, нарушение общественного порядка и т.д.
С одним концом провода разобрались

Кто звонит?
Перейдем к другому, а именно - звонящему. Человек, проживающий в квартире города М в течение дня может оказаться в довольно различных ролях:

- житель квартиры,
- водитель ТС,
- пешеход,
- пассажир,
- потребитель
- работу делатель,
- родитель,
- вредитель и так далее.

Причем этот субъект предположительно не статичен, а перемещается из зоны в зону в течение дня и дней недели, и при этом в каждой из своих ролей может поучаствовать в ситуации, потребующей помощи специальных служб.

Таким образом, для решения задачи на карту необходимо:
1) нанести дома (жилые и офисные),
2) дороги или некую дорожную характеристику
3) некоторую дополнительную инфраструктуру.

Т.е нужны объекты, которыми пользуется субъект и которые прямо или косвенно влияют частоту возникновения происшествия (например, парки в основном заселены деревьями и белками, но там есть деревья, которые могут упасть)

Про Москву

В Москве есть некоторое количество офисных кластеров – Сити, Павелецкая, Курская, центр города в пределах Бульварного. При таком расположении доля экстренных звонков должна снижаться в выходные в зонах с высокой концентрацией офисных помещений и повышаться в районах, где расположены значительные торговые площади (Мега, Крокус и т.п.).

Поскольку на рабочем месте человек занимает гораздо меньше места (в среднем примерно как двуспальная кровать у него дома), то и офисных помещений в городе требуется не так много, как жилых домов.

Предсказываем плотность населения
Чтобы хорошо предсказывать вызов скорой первое, что приходит в голову – нужно определить плотность населения по зонам, и хорошо бы еще иметь распределение по возрастам, т.к. к пожилым людям скорую вызывают чаще.

В OpenStreetMap есть несколько тегов, которые могли бы при хорошей разметке помочь кластеризовать районы города, например, с помощью таких значений как residential, office, building. Однако в Москве они оказался не очень информативными. Наиболее заполненный тэг по домам (addr:housenumber) присвоен только 5799 записям

tag	count
name	73795
amenity	54400
highway	47306
entrance	37389
barrier	34957
shop	34662
opening_hours	25387
ref	23539
natural	22318
crossing	18459

Хорошими переменными, характеризующими количество людей, являются количество продуктовых магазинов и аптек, так как эти бизнесы постоянно решают задачу «быть ближе к своему клиенту».

Вторым источником можно взять данные по избирателям (см датасеты). Нас будет интересовать в поле ‘size’, которое отражает количество избирателей на участке, и координаты этого участка. Для использования этих данных необходимо соотнести участки с зонами и далее просуммировать поле size.
Если координаты участка попадают в зону то этому участку присваивается номер зоны (zone_id). Далее в качестве переменной используется сумма по всем участкам в одной зоне.

Третий источник – данные data.mos.ru по камерам уличного наблюдения и камерам установленным на домах.
Данные о камерах на домах (дом на карте) могут помочь определить плотность застройки в зоне.

Камеры наружного наблюдения могут характеризовать как активность движения граждан пешеходов и автомобилистов, так и отслеживать потенциально опасные зоны, а, следовательно, должны были бы положительно корелировать с количеством вызовов спецслужб.

Дорожная сеть

Количество остановок общественного транспорта также является прокси для определения густонаселенности зоны.

Многие перекрестки за последние годы стали оборудовать светофорами. Таким образом, нерегулируемые перекрестки, являются потенциально опасным дорожным объектом (node.tags.get('crossing'== 'uncontrolled') and (node.tags.get('highway')=='crossing'))

Данные о ДТП по зонам
Источник данных https://xn--80abhddbmm5bieahtk5n.xn--p1ai/opendata
Больше равно одного ДТП

Метрика Кендалл-тау
В задаче нам нужно отранжировать целевые зоны по количеству звонков.
Это означает, что в качестве таргета можно брать не целевое значение количества звонков, а ранг зоны относительно других зон.
Таким образом, в качестве целевой переменной можно брать, либо исходный таргет, либо попытаться его растянуть, напримеp:

- Ранг зоны по количеству звонков (возрастающий или убывающий порядок)
- Трансформированное значение по Бокс-Коксу (корень из логарифма)
- Ранг трансформированного значения по Бокс-Коксу
- Экспонента таргета

Что можно делать с результатом

В качестве предсказания можно брать

- Единое предсказание по ‘daily’ как в стартере
- среднее предсказаний по будним дням, и среднее предсказаний по выходным
- предсказание каждого дня отдельно
- усреднять предсказание daily и предсказания отдельно дня
- усреднять предсказание daily, рабочего (выходного) и отдельного дня

Нерегулируемые перекрестки

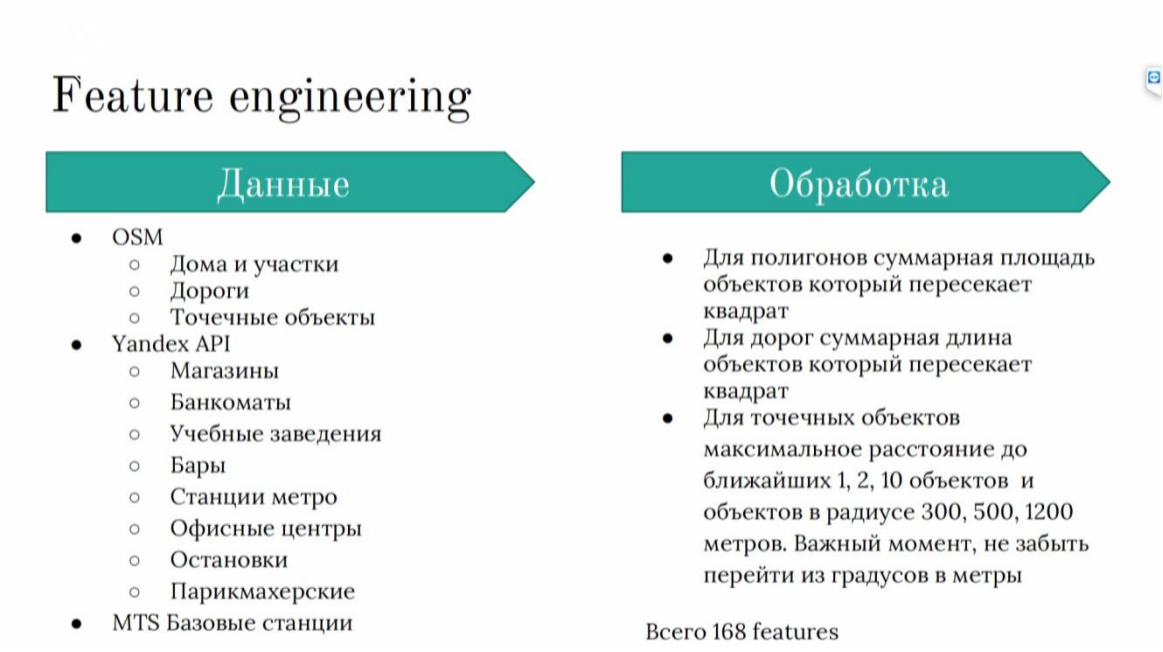
Распределение по районам оказалось следующим

картинка
Данные по камерам наблюдения

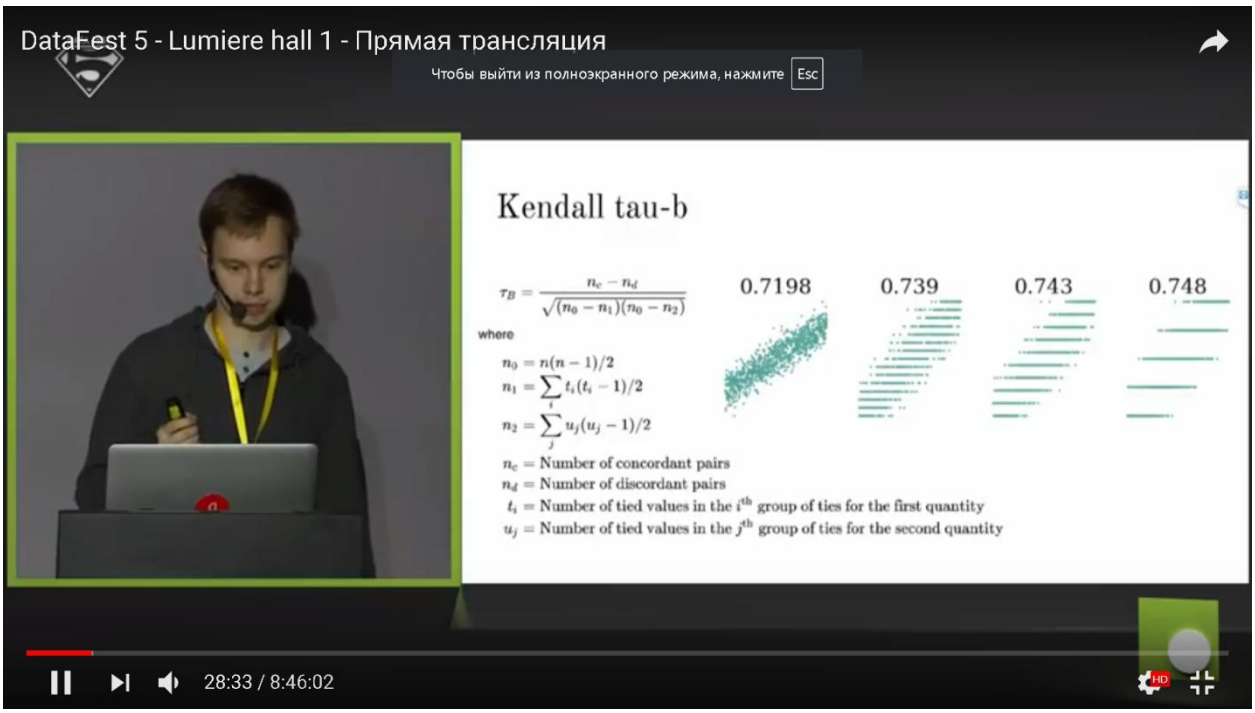
Что можно было еще сделать для результата

- Преобразовать координаты в прямоугольные
- Данные с сайта реформа ЖКХ
- Оценить среднюю этажность домов
- Данные по вышкам МТС в районе
- Нанести на карту крупнейшие ТЦ (Мега, Твой дом, крокус, Оби…)
- Вокзалы и соответственно районы вокруг вокзалов (аэропорты)
- Way – площадь
- Я использовал RandomTreesRegressor, XGBRegessor
- Slonoslon (Андрей Филимонов) – LightGBM – исользовать данные где модель предсказывает хуже чем в реалии для выбора хороших районов для жилья)

Победитель (Павел логачев)



Округление по квантилям (10 штук)



Модель по дням

ExtraTreesRegressor MAE