1 Краткое описание проекта

Этот учебный проект следует ноутбуку с домашним заданием с курса "Машинное обучение 1" Е. Соколова и соотвествующего соревнования на Kaggle в котором нужно было предсказать длительность поездки на такси по имеющимся данным.

В этом проекте использовалась линейная регрессия $a(x) =< wx >$ и лучший результат получился с L_2 регуляризацией. Основная работа пришлась на доработку признаков которые описаны более детально внизу. В результате оценка качества RMSE после логарифмирования таргета вышла:

$$RMSE_{test} = 0.478,$$

на обучающей выборке:

$$RMSE_{train} = 0.451.$$

Для сравнения RMSE для константного предсказания — среднего значения на тестовой выборке:

$$RMSE_{mean} = 0.797,$$

лучшее RMSLE в соревновании

$$RMSLE_{top} = 0.289.$$

В самом соревновании используется метрика RMSLE.

2 Описание признаков

Здесь кратко описаны наиболее значимые признаки.

2.1 Haversine distance

В датасете были даны координаты (долгота и широта) начала и конца поездки. С помощью этих четырех значений можно вычислить пройденное расстояние между двумя точками как расстояние на сфере. Это самый сильный признак и далее он логарифмируется — $x \longrightarrow \ln(1 + x)$ (в датасете есть данные с нулевым пройденным расстоянием.)

Этот признак на лог. шкале намного лучше коррелирует с таргетом чем без лог. шкалы, 0.57 против 0.75 для коэффициента корреляции Пирсона.

2.2 Начало и конец в аэропорте

На рисунке 1 видно, что есть две области, которые выделяются на карте — это два аэропорта Нью - Йорка. В соответствии с этим наблюдением мы вводим категориальные признаки:

Началась или закончилась ли поездка в одном из аэропортов?

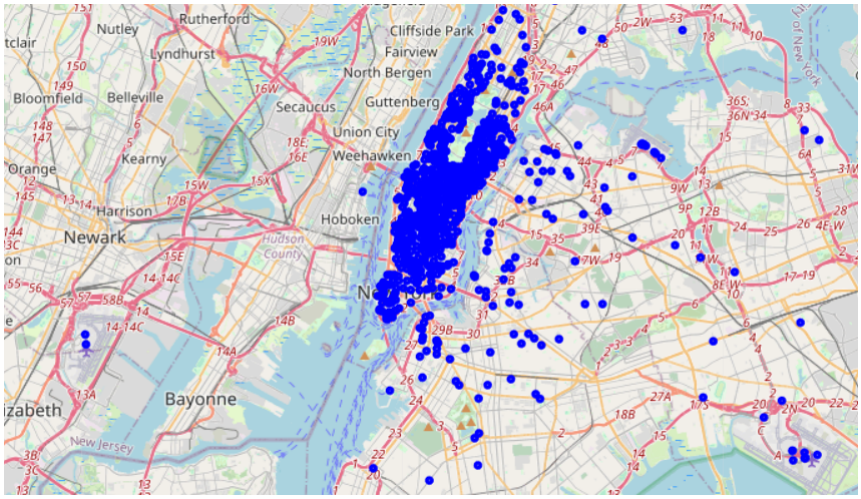

Рис. 1: Точки окончания поездок

Нам даны координаты начала и конца поездки; зная площадь аэропорта и беря точку где-то на территории аэропорта можно оценить с помощью признака расстояния началась ли поездка в аэропорту или нет.

2.3 Была ли поездка в загруженный день? Или наоборот

Мы знаем дату и время начала поездки, подсчитав расстояние на сфере как было выше мы можем оценить среднюю скорость поездки. Что бы оценить загруженность дорог в зависимости от дня недели и часа мы строим график со средней скоростью движения.

С помощью этого графика (Рисунок 2) можно ввести два признака:

Началась ли поездка в час-пик? Или наоборот в не загруженный день?.

2.4 Сетка в области с самым большим количеством поездок

Область на карте с самым большим количеством поездок можно разделить на сегменты, как на рис. 3. Каждый сегмент занумерован номером строки и столбца — (i, j). Можно написать трансформер который принимает две точки на вход — координаты углов прямоугольника и делит его на $m \cdot n$ сегментов. Это еще один признак:

Поездка началась или закончилась в сегменте (i, j).

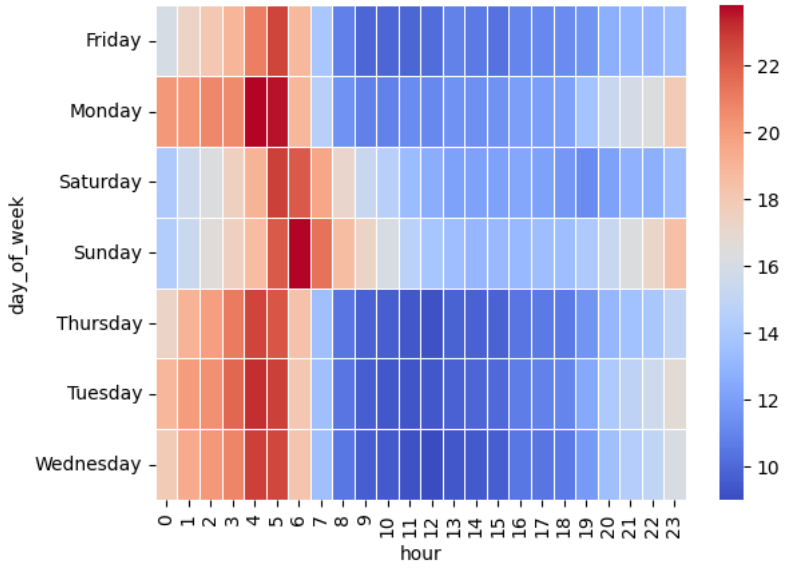

Рис. 2: Средняя скорость в зависимости от часа и дня недели

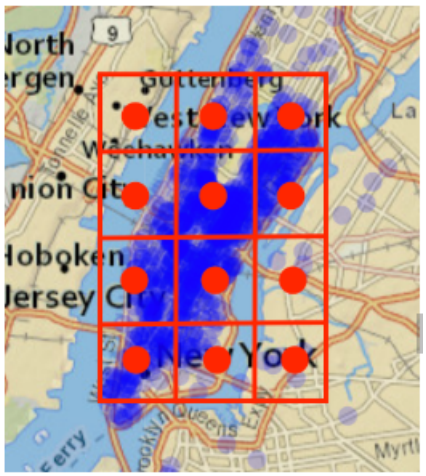

Рис. 3: Пример сетки

3 Что было сделано еще

В основном признаки являлись категориальными и они были закодированы с помощью One-hot encoding. Лучший результат получился с моделью Ridge — L_2 регуляризацией. Оптимальный коэффициент регуляризации можно оценить с помощью GridSearch на валидационной выборке. Он получился:

$$\alpha = 3.684.$$

Аномальные значения с либо слишком низким временем поездки (≈ 0), либо слишком большим ($\approx$ 24 часа) были выброшены из обучающей выборки. Аналогично были обработаны другие численные признаки.

Редкие категории которые встречались ≈ 12 раз или меньше в тренировочной выборке (размера $\approx 10^6$) были объединены в одну категорию. На тестовой было сделано тоже самое.

Все эти действия не дали значительного прироста в показателе качества, примерно в 0.02.