New York City Taxi Trip Duration

Siyi Yu[1]

[1] Jilin University, China

Introduction

In this competition, Kaggle is challenging us to build a model that predicts the total ride duration of taxi trips in New York City.This is a prediciton problem.The raw datasets mainly two files,whose attributes are shown below.

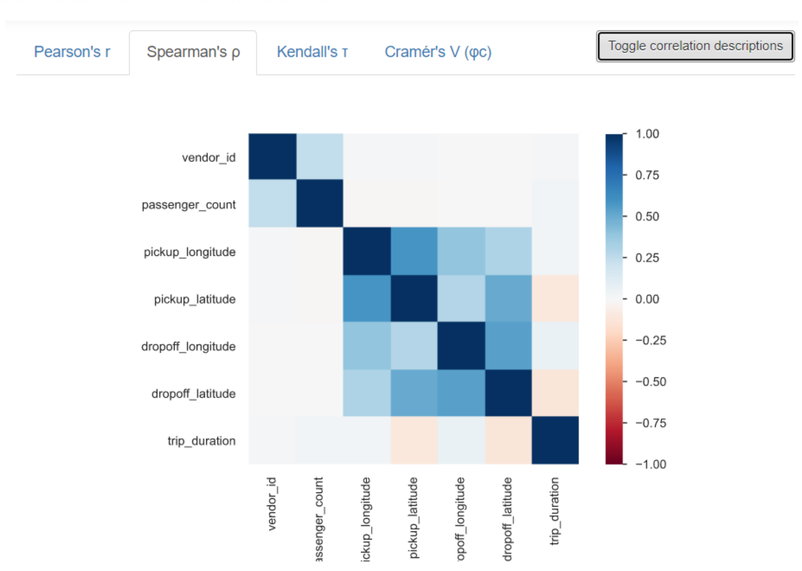
- *train.csv:*
  vendor_id,pickup_datetime,dropoff_datetime,passenger_count
  pickup_longitude,pickup_latitude,dropoff_longitude,dropoff_latitude
  store_and_fwd_flag,trip_duration
- *test.csv:*
  vendor_id,pickup_datetime,passenger_count
  pickup_longitude,pickup_latitude,dropoff_longitude,dropoff_latitude
  store_and_fwd_flag

Data Processing

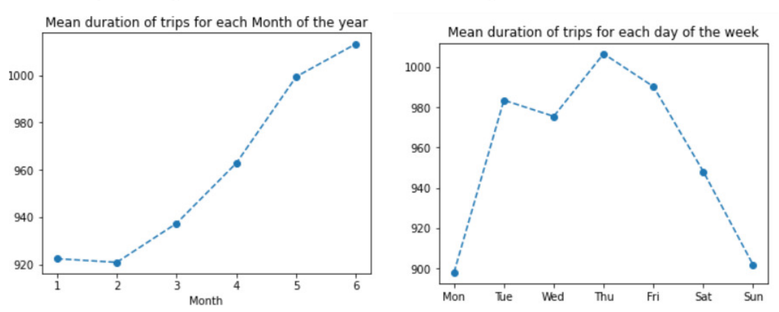
- Remove missing value and NaN value
- Filter outliers and duplicate data
- Process pickup/dropoff_datetime
- Process pick/dropoff_latitude/_longitude
- Process string by one-hot encoding

Data Visualization

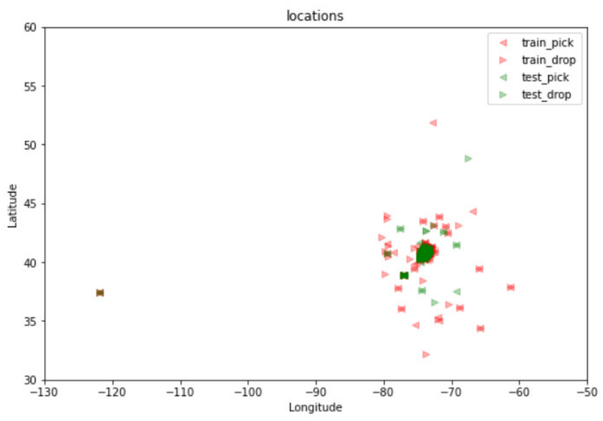
- Spearman Correlation of attributes



- Duration of taxi trips regard to month,weekday



- Pick/Drop locations of taxi trips

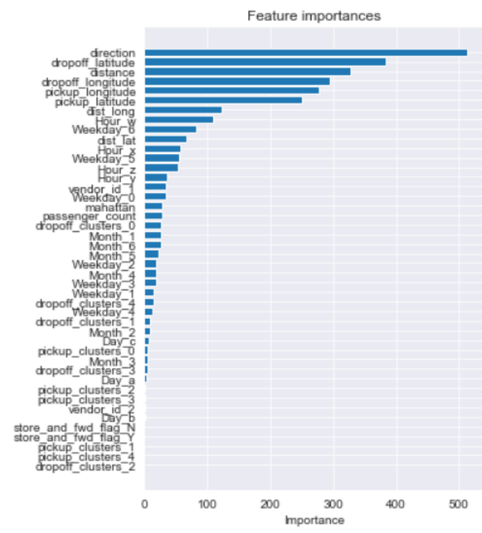


Feature Selection and Feature Importance

- After data Processing,I remove,add and change some attributes.

#	Column	Non-Null Count	Dtype
0	passenger_count	1437158 non-null	float64
1	pickup_longitude	1437158 non-null	float64
2	pickup_latitude	1437158 non-null	float64
3	dropoff_longitude	1437158 non-null	float64
4	dropoff_latitude	1437158 non-null	float64
5	distance	1437158 non-null	float64
6	mahattan	1437158 non-null	float64
7	direction	1437158 non-null	float64
8	dist_long	1437158 non-null	float64
9	dist_lat	1437158 non-null	float64
10	vendor_id_1	1437158 non-null	float64
11	vendor_id_2	1437158 non-null	float64
12	store_and_fwd_flag_N	1437158 non-null	float64
13	store_and_fwd_flag_Y	1437158 non-null	float64
14	Month_1	1437158 non-null	float64
15	Month_2	1437158 non-null	float64
16	Month_3	1437158 non-null	float64
17	Month_4	1437158 non-null	float64
18	Month_5	1437158 non-null	float64
19	Month_6	1437158 non-null	float64
20	Day_a	1437158 non-null	float64
21	Day_b	1437158 non-null	float64
22	Day_c	1437158 non-null	float64
23	Hour_w	1437158 non-null	float64
24	Hour_x	1437158 non-null	float64
25	Hour_y	1437158 non-null	float64
26	Hour_z	1437158 non-null	float64
27	Weekday_0	1437158 non-null	float64
28	Weekday_1	1437158 non-null	float64
29	Weekday_2	1437158 non-null	float64
30	Weekday_3	1437158 non-null	float64
31	Weekday_4	1437158 non-null	float64
32	Weekday_5	1437158 non-null	float64
33	Weekday_6	1437158 non-null	float64
34	pickup_clusters_0	1437158 non-null	float64
35	pickup_clusters_1	1437158 non-null	float64

- Feature Importance



Model and Result

- Model:Lightgbm and Xgboost
- Private score:0.40978
- Public score:0.41166

- Rank:532/1254

Conclusion

Compare to midterm presentation,My score have greatly imporved from 0.59486 to 0.40978. The reason mainly is more features and more models. In the figure of feature importance,we can the attribute direction that I added plays a great role. And xgboost model can fit the train data better, but processing speed of lightgbm model is faster.