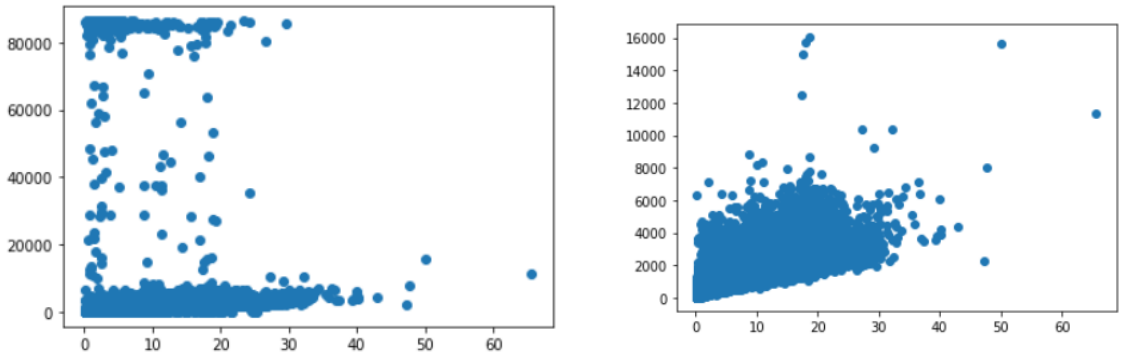
NYC Taxi Travel Duration Prediction Competition Report – Changjin Liu

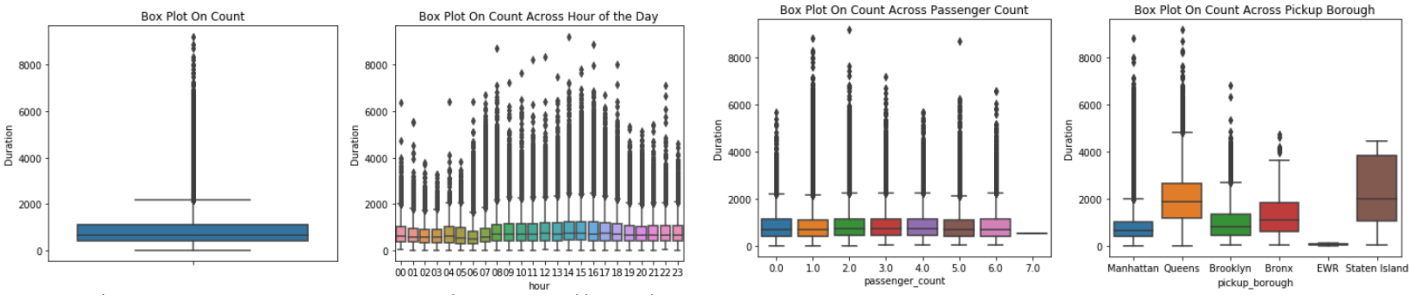
My work is divided into four parts: data exploration, feature engineering, train one simple model and ensemble learning with several regressors.

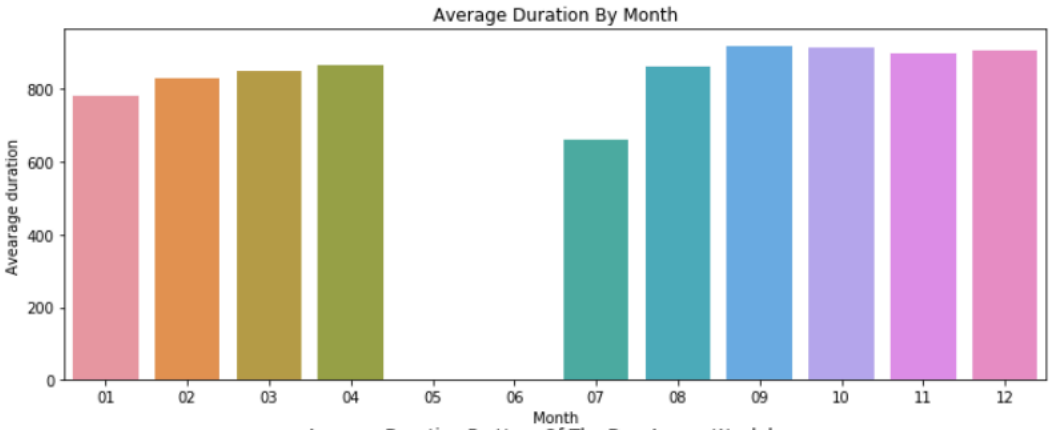
1. Data exploration

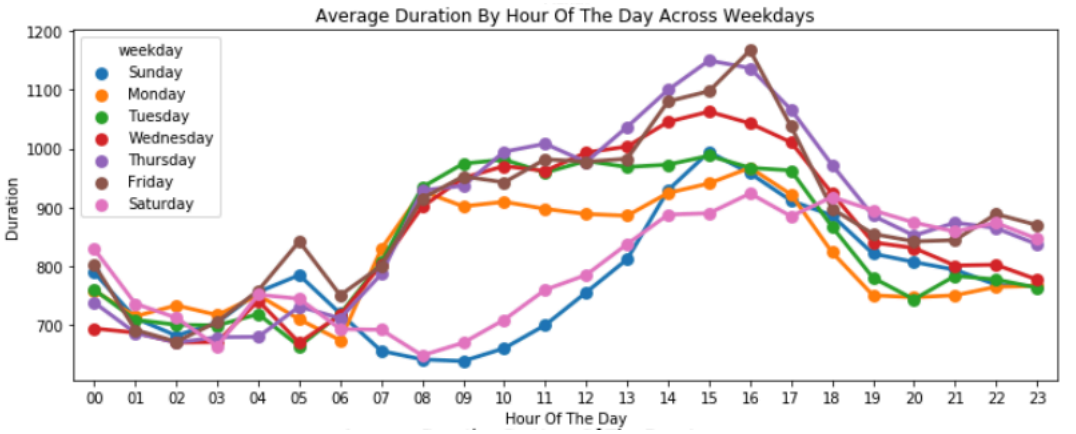
Intuitively, duration may have a very large correlation with trip distance, so I drew one scatter with those two variables, as shown on the left figure (x=trip_distance, y=duration).

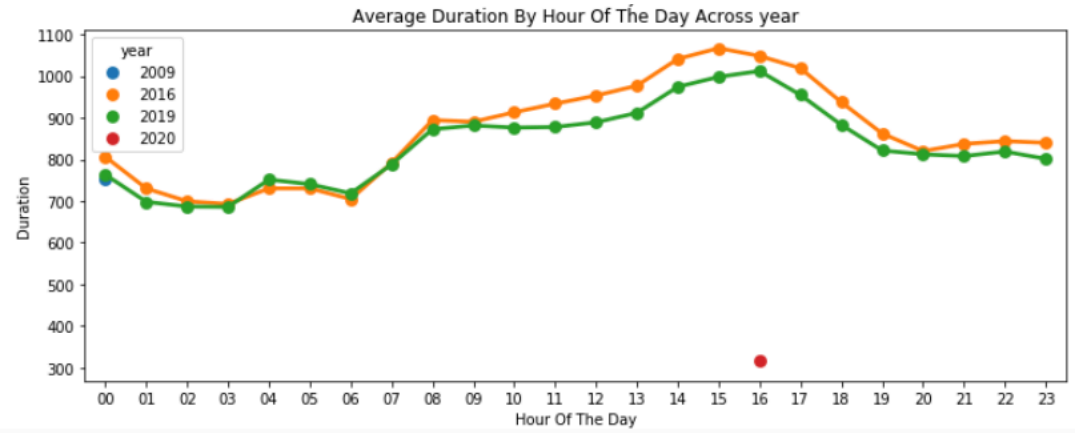


We can see that when the trip distance is less than 35 miles, there are many abnormal points on the top. It makes no sense that one can spend more than 22 hours (79200s) taking taxi. Therefore, we set some limitations for the speed (trip_distance / duration). The upper bound is 100 mph (the highest speed for a normal car) and the lower bound is 3 mph (speed on foot). However, there are also some exceptions. When the trip distance is small, it is also possible that the speed for a taxi is smaller than that on foot because of the traffic jam, time for getting on etc. Therefore, we also include that part of data and draw the right scatter. We also remove the outliers above 10000. We filled the nulls of VendorID and passenger_count with the corresponding mode, the nulls of pickup_zone and dropoff_zone with NV. We select the corresponding year, month, day, hour from timestamp. Then we can draw the relationship with duration and some features.





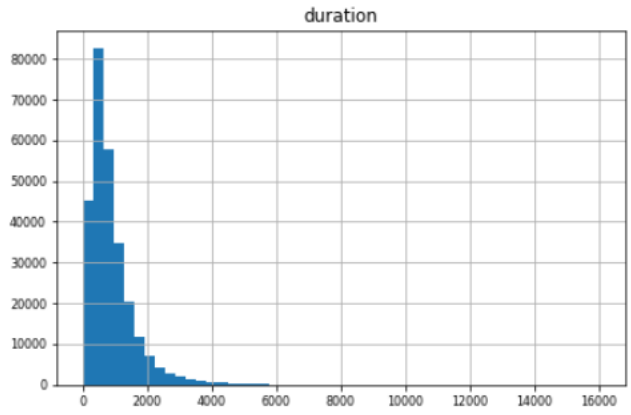


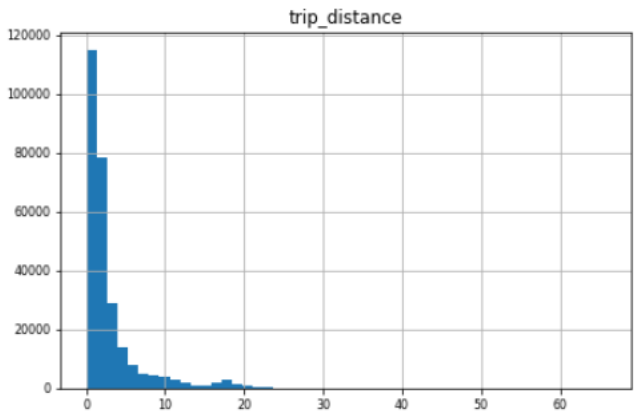


We can see that hour, pickup_borough, month and hour-weedkay have much variance, however there is almost no difference between different passenger_count. So we select trip_distance, pickup_borough, pickup_zone, dropoff_borough, dropoff_zone, **year**, **month**, **hour**, **weekday** as the training features from the original data.

2. Feature engineering

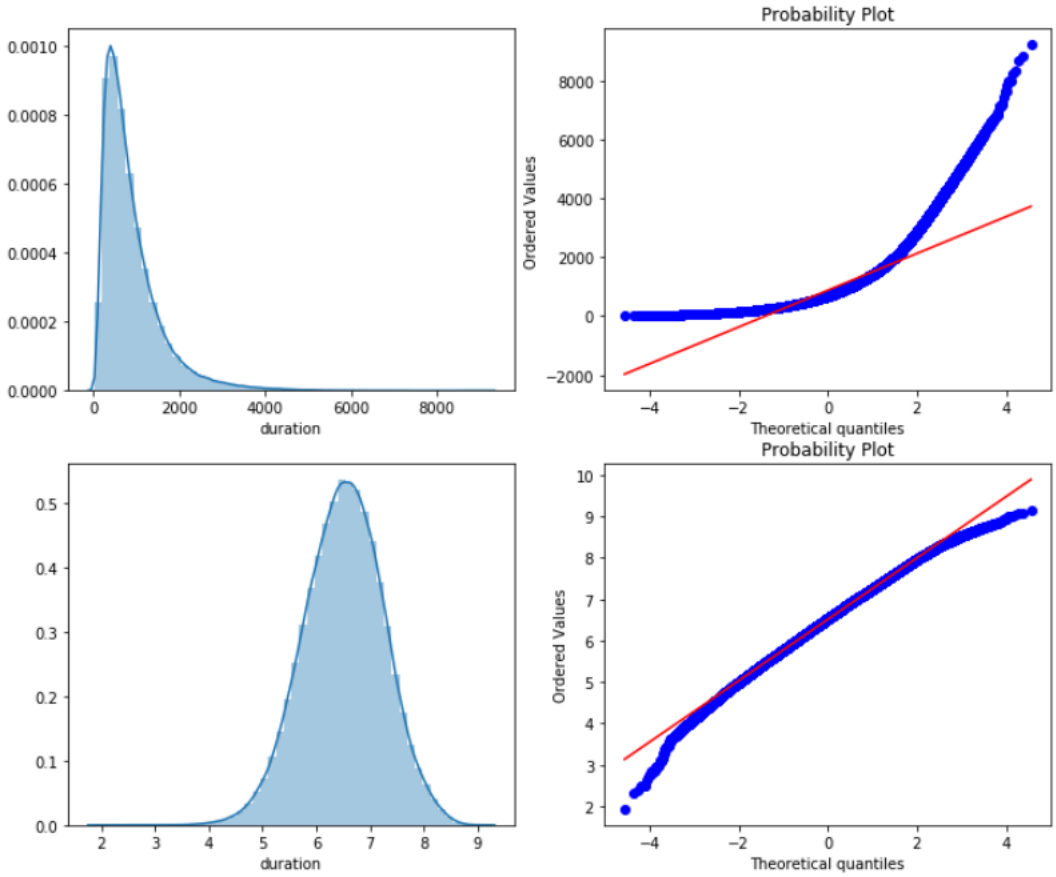
We plot the duration and trip_distance and duration, and see that the patterns are similar. And further we convert





duration into log(duration+1) and see that it comes to the Gaussian distribution. So we did the same thing to **trip_distance** and get the new feature **trip_distance_log**.

Further, we concat (hour, weekday) and (pickup_borough, dropoff_borough), calculate the average duration for (hour, weekday) and (pickup_borough, dropoff_borough) pair, hour, and got **hour_weekday_mean**, **hour_mean**. Also we can see that whether the day is workday matters a lot. So we added 0/1 variable **workday** to judge whether one weekday is a workday. We convert categorical variables into hash codes including **workday, month, hour, weekday, pickup_borough, pickup_zone, dropoff_borough, dropoff_zone, hour_weekday, borough_path** and delete variables which are too concentrated.



3. Train the Gradient Boosting Regressor

We then did the 5-ford cross validation to test the gradient boosting regressor. It gave 290.365 RMSE offline with parameters n_estimators=3000 and alpha=0.05. The score on Kaggle is 287.61 (5th).

4. Ensemble Learning

We then did the model fusion, using ridge, lasso, elasticNet, gradient boosing regressor, lightGBM and XGBoost. The scores (RMSLE) of those models are as follows.

Ridge	Lasso	ElasticNet	GBR	LightGBM	XGBoost
0.3205	0.3208	0.3212	0.3014	0.3190	0.3138

Then we train the ensemble model of those 6 models, and combined the results of ensemble models and those 6 models together with different weights (Ridge=0.1, Lasso=0.1, ElasticNet=0.05, GBR=0.2, LightGBM=0.1, XGBoost=0.1, Ensemble=0.35). However, we abandon that solution because it takes so much time and the result is not better than that of Gradient Boosting Regressor.

Tips:

1. My Kaggle name is ChangjinLiiu123, and the final rank is 5th.
2. My codes were divided into 3 files (data_exploration, model_training and ensemble_learning), and I uploaded on Github. You guys can just click and check.

Github address: https://github.com/EvanLiu123/NYC-Taxi-Travel-Duration-Prediction-Competition

Kaggle Leaderboard address: https://www.kaggle.com/c/ieor242hw4/leaderboard