KAGGLE CASES, Summer 2019
Project – 1

Name : Dhivya Swaminathan
UID : 2000434729

There were three steps involved in this project:

1. Feature Engineering
2. Modeling
3. Prediction

Feature Engineering:

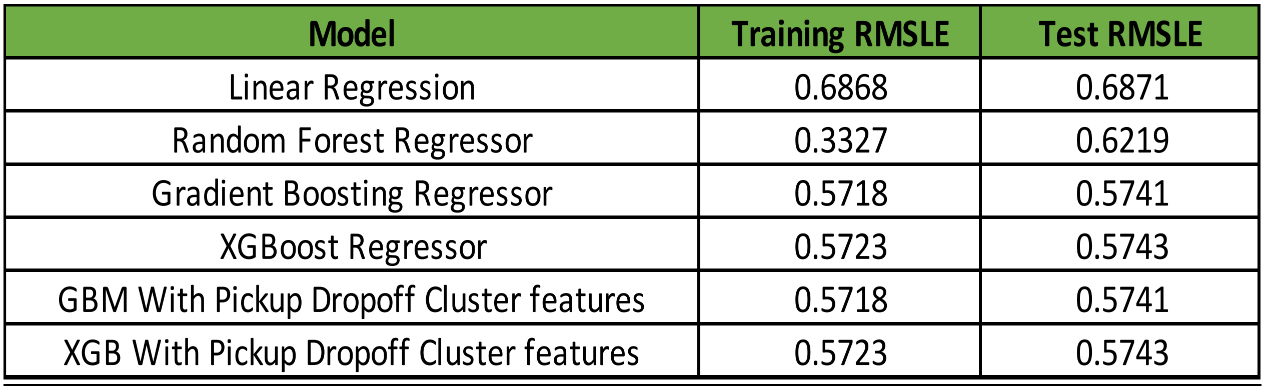
1. The first feature that was created was the **trip_distance**. This was calculated using the haversine distance from the package haversine, using the features pickup_latitude, pickup_longitude, dropoff_latitude, dropoff_longitude.

2. The next set of features that were extracted were the time related features like **hour_of_pickup**, **minute_of_pickup**, **day_of_month**, **month** and **day_of_week**. These features were extracted from the feature *pickup_datetime*.

3. The last set of features that were extracted were the ***pickup_cluster*** and the ***dropoff_cluster***. These were extracted from the features - pickup_latitude, pickup_longitude, dropoff_latitude, dropoff_longitude. This was done by running a Mini batch k means algorithms with 100 clusters.

Modeling:

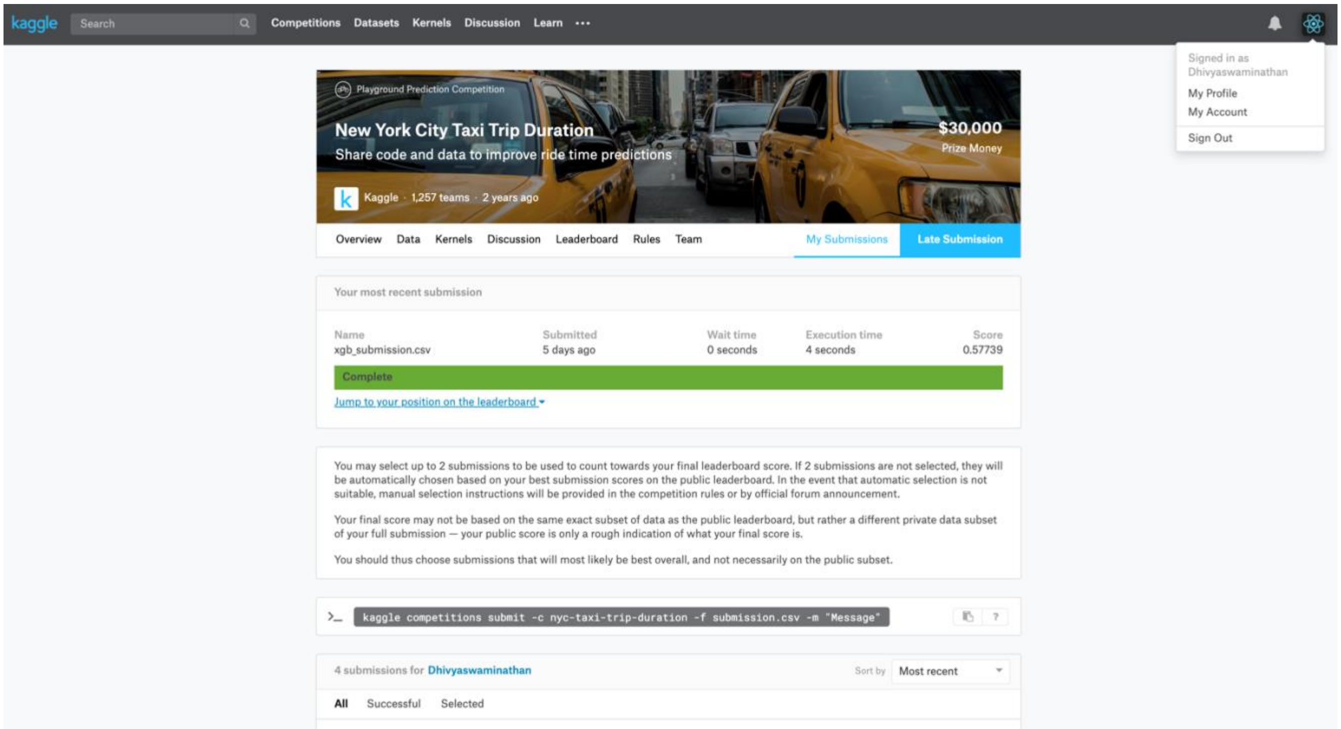
Various models were tried out to identify the best model for the project. The table reporting the models and their accuracies are shown below:



Model	Training RMSLE	Test RMSLE
Linear Regression	0.6868	0.6871
Random Forest Regressor	0.3327	0.6219
Gradient Boosting Regressor	0.5718	0.5741
XGBoost Regressor	0.5723	0.5743
GBM With Pickup Dropoff Cluster features	0.5718	0.5741
XGB With Pickup Dropoff Cluster features	0.5723	0.5743

Prediction:

XGBoost model was decided to be the final model and this was used to predict the test set and the submission was made to the Kaggle competition



New York City Taxi Trip Duration

Share code and data to improve ride time predictions

$30,000 Prize Money

Kaggle · 1,257 teams · 2 years ago

Your most recent submission

Name	Submitted	Wait time	Execution time	Score
xgb_submission.csv	5 days ago	0 seconds	4 seconds	0.57739

Complete

Jump to your position on the leaderboard

You may select up to 2 submissions to be used to count towards your final leaderboard score. If 2 submissions are not selected, they will be automatically chosen based on your best submission scores on the public leaderboard. In the event that automatic selection is not suitable, manual selection instructions will be provided in the competition rules or by official forum announcement.

Your final score may not be based on the same exact subset of data as the public leaderboard, but rather a different private data subset of your full submission — your public score is only a rough indication of what your final score is.

You should thus choose submissions that will most likely be best overall, and not necessarily on the public subset.

```
kaggle competitions submit -c nyc-taxi-trip-duration -f submission.csv -m "Message"
```

4 submissions for Dhivyaswaminathan

Reference:

1. https://www.kaggle.com/gaborfodor/from-eda-to-the-top-lb-0-367#Data-understanding
2. https://www.kaggle.com/karelrv/nyct-from-a-to-z-with-xgboost-tutorial