NYC Taxi Rides: Trip Duration Prediction

Suresh Gopalakrishnan

February 17, 2018

1. INTRODUCTION

New York City taxi rides paint a vibrant picture of life in the city. The millions of rides taken each month can provide insight into traffic patterns, road blockage, or large-scale events that attract many New Yorkers. With ridesharing apps gaining popularity, it is increasingly important for taxi companies to provide visibility to their estimated ride duration, since the competing apps provide these metrics upfront.

Predicting duration of a ride can help passengers decide when is the optimal time to start their commute. This problem was posted by **"NYC Taxi and Limousine Commission"** as competition in Kaggle.com challenging us to build a model that predicts the total ride duration of taxi trips in New York City.

In this report, we discuss three models: Linear Regression, Random Forest, Gradient Boosting. We evaluate these models based on the Root Mean Square Logarithmic Error (RMSLE) and Pseudo R^2. We also discuss the importance of various features in our prediction algorithms.

We achieved lowest RMSLE of 0.3995 and Pseudo R^2 of 0.7452 in duration prediction using Random Forest model .

2. DATA

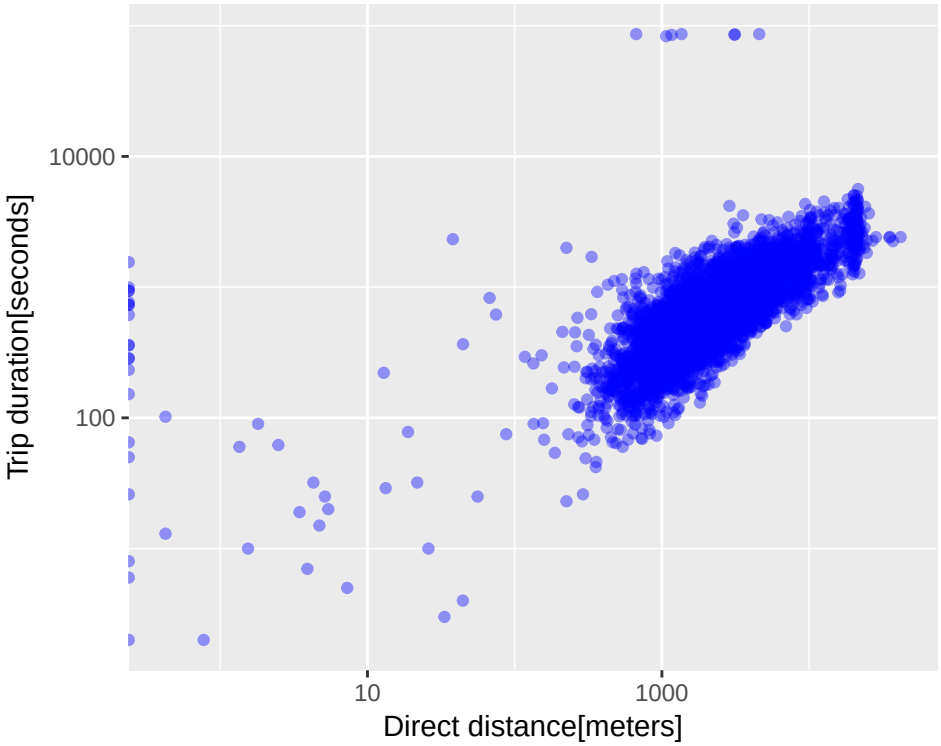
Primary data for this analysis was released by the NYC Taxi and Limousine Commission, which includes pickup time, geo-coordinates, number of passengers, and several other variables. Training dataset has close to 1.5 Million and 630k records in test dataset. Each row contains one taxi trip. Added to Taxi data, we are adding Open Source Routing Machine, OSRM data for each trip. This data is provided by oscarleo and we can download data from here and includes the pickup & dropoff, streets and total distance/duration between these two points together with a sequence of travels steps such as turns or entering a highway. Features from OSRM data like, *total_distance*, *total_travel_time* played an important role in this prediction. After through analysis we have removed trips with zerio miles and trips with more than 24 hours duration as outliers. For modelling we are taking a subset of 400K trips and run our model on the subset data.

Below figure demonstrates the pick-up (blue) and drop-off (red) locations of 5000 trips from our data in New York City.

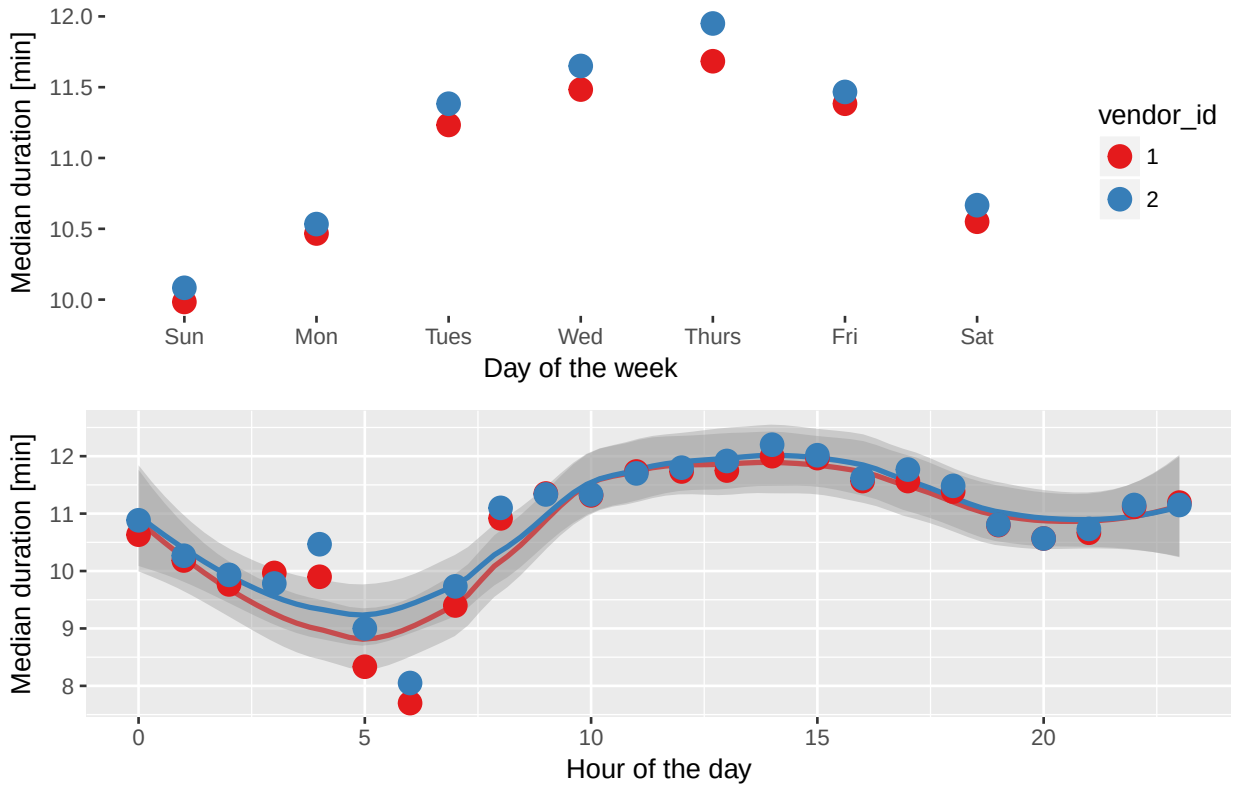
3. EXPLORATORY DATA ANALYSIS

To better understand the problem and the features, we perform exploratory analysis on the data. The purpose of this analysis is to get insights on how the prediction variables behave with various features and how can we leverage these features in our models to achieve best possible results. Below are few plots will provide us better insight of data we are dealing with. However, I have not included complete EDA analysis of this project. Please refer this GitHub Link for complete EDA and modelling document.

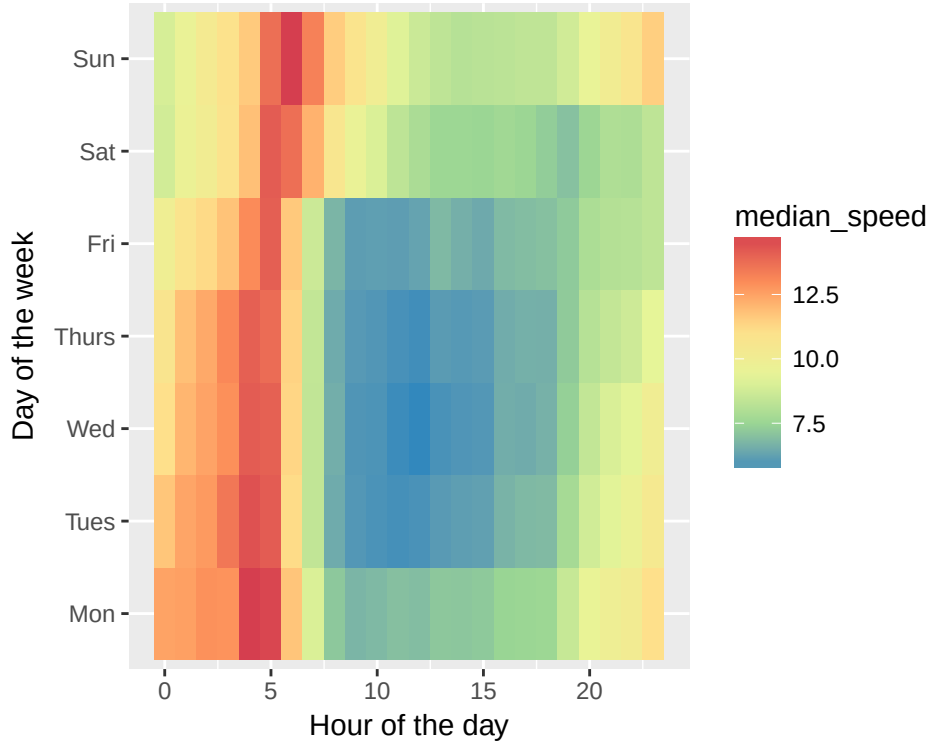
Distanct demonstrates a positive correlation between trip distance and duration. An interesting finding is that the variance of duration increases, as the trip distance increases.



City Rides get busy during day time. Pickup Hour may be an important factor decides trip duration. Lets plot how pick up time impacts duration on weekdays and hours within a day.



Trip Duration increases due to traffic on busy roads. Below heat maps give a insight how speed affects trip duration.



4. PREDICTIVE TASK

We are predicting the duration of a NYC Yellow Taxi ride. Training Data set contains close to 1.4M trips. In order to train the model in less processing time, we have taken subset of 400 thousand records and we are going apply our modelling technique on the data. In order to evaluate the performance of our model, we split the data into a training set (70%), validation set (15%) for cross validation and testing set (15%). Our evaluation metric is **Root Mean Squared Logarithmic Error.** In order to easily simulate the evaluation metric in our model ???tting we replace the *trip_duration* with its logarithm. In comparing the results between our different models, we also report the **pseudo R^2 = 1-MSE/Var(Y)** value in order to evaluate how well the models perform relative to the variance of the data set.

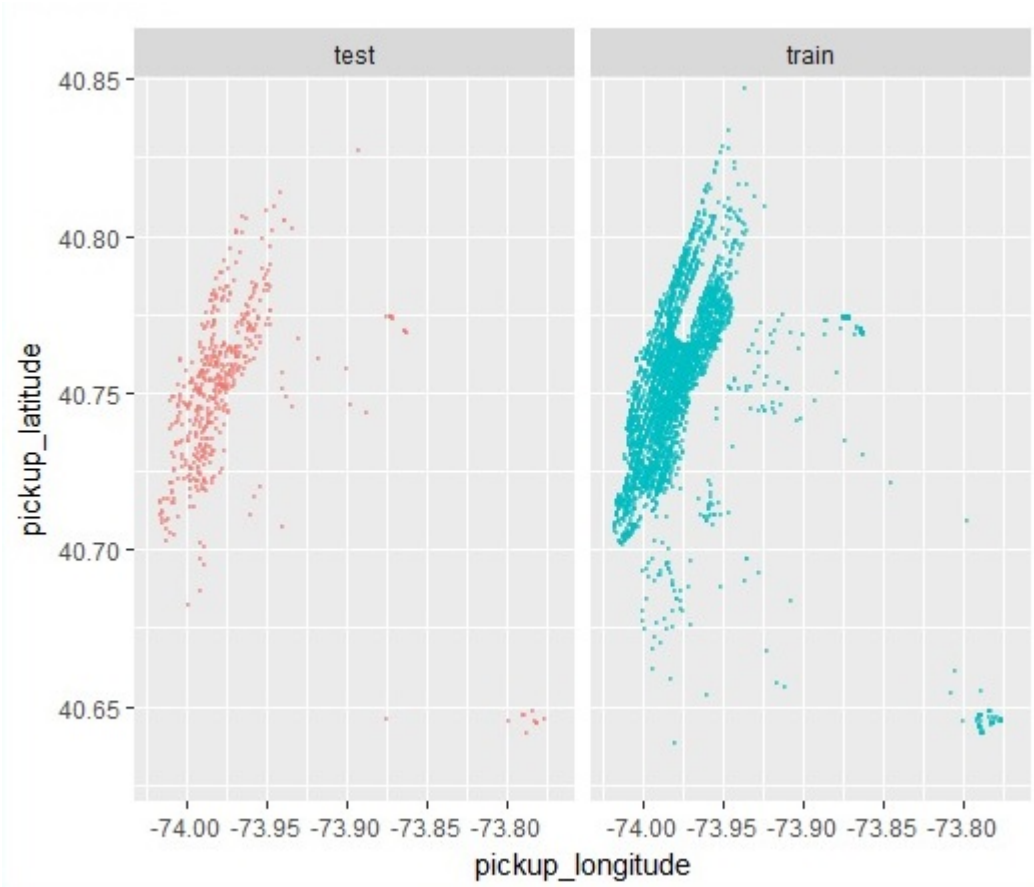
We have plotted trips and summary of Trip Duration to understand training data relevant to test data.

Summary of Trip Duration in Train

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.6931 5.9865 6.4953 6.4665 6.9801 11.3664
```

Summary of Trip Duration in Test

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.099   5.984   6.498   6.467   6.983  11.366
```



5. FEATURES

Feature selection is normally an iterative process where we run an initial model with either few or many features and then step-by-step add or remove some features based on the results of the previous run. Features like *total_travel_time*, *total_distance* , *hour*, *dist*, *wday* are having high corellation and we have used in our models.

6. REGRESSION MODELS

Multiple-linear regression

We are using Linear Regression as our baseline model. The multiple-linear regression model allows us to exploit linear patterns in the data set. This model is an appealing first choice because feature weights are easily interpretable and because it runs efficiently on large datasets.

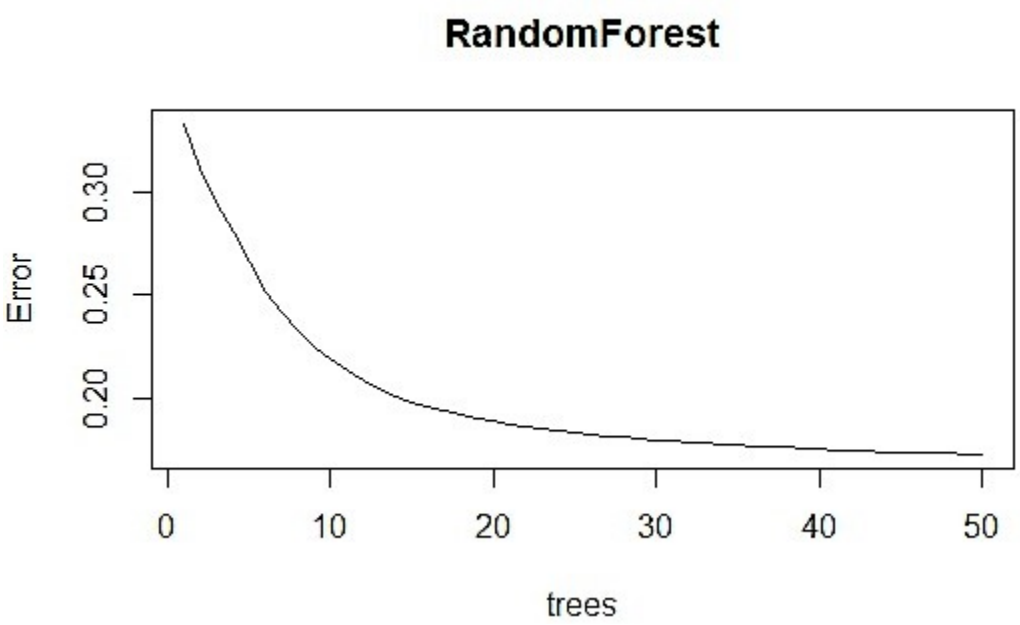
Results -> RMSLE : 0.5321 and Pseudo R^2 : 0.5480

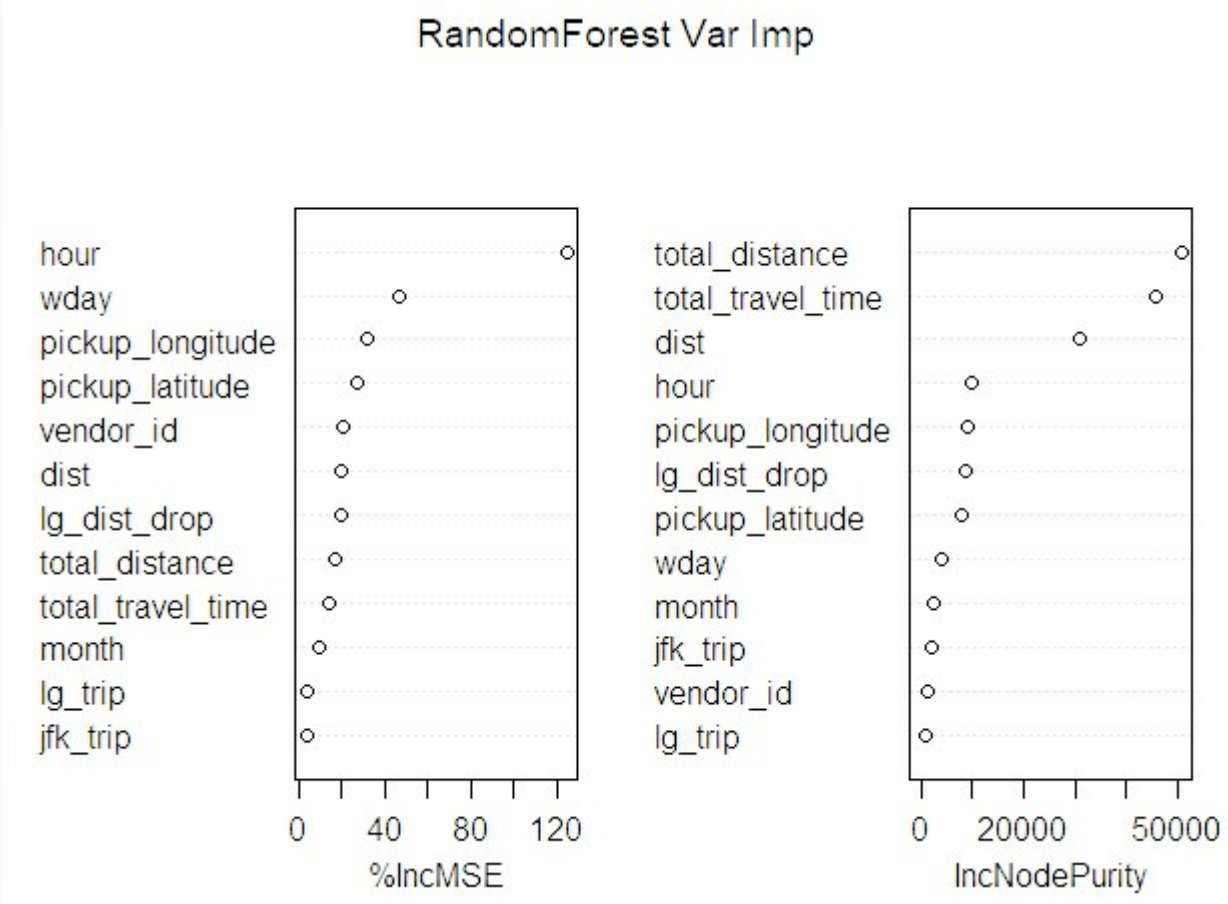
Random Forest

The tree regression model is capable of representing complex decision boundaries, thus complementing our other chosen models. Random Forest is chosen since it prevents overfitting and robust against outliers. Since we have limited computing resource, we have run alogrithm with just 50 trees. Results are based on the same.

Results -> RMSLE : 0.3995 and Pseudo R^2 : 0.7452

Listed below are important features predicted by Random Forest model and error trend as tree grows.





XGBoost

XGBoost is an advanced gradient boosting algorithm. It is a highly sophisticated algorithm, powerful enough to deal with all sorts of irregularities of data. The tool is extremely flexible, which allows users to customize a wide range of hyper-parameters while training the mode, and ultimately to reach the optimal solution.

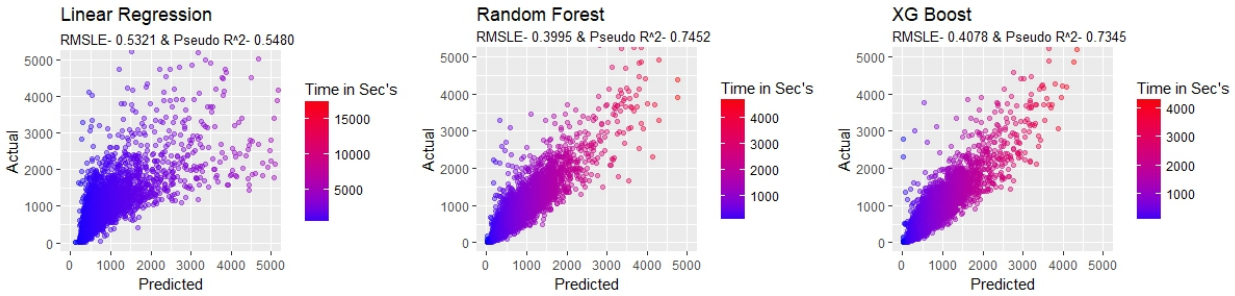
Results -> RMSLE : 0.4078 and Pseudo R^2 : 0.7345

The results of all the models are compared in the below table

Algorithm	RMLSE	Pseudo R^2
Linear Regression	0.5321	0.5480
Random Forest	0.3995	0.7452
XG Boost	0.4078	0.7345

7. MODEL ANALYSIS & CONCLUSION

In order to visualize how well the models (randomforest, xgboost, Linear Regression) perform, we plot the actual versus predicted trip duration in a density plot, where the blue is predicted and red is actual trip duration.



The plots in the figure suggest that both Random Forest and XG Boost models perform well on the test set. Most predictions are almost close to the true values. Random Forest achieved the lowest RMSLE of 0.3995 and Pseudo R^2 of 0.7452 in duration prediction. Though we cannot claim this as best model by looking at the results. But this is my first take on Machine Learning algorithm predictions, so this can still be used as tool to find approximate estimate.

References:

- EDA - NYC TAXI EDA The fast & the curious
- EDA - From EDA to the Top (LB 0.367)
- Mapping techniques
- GGPLOT
- Machine Learning
- XGBOOST Basics