# Cab Fare Prediction

# Navdeep Singh Sidana

3rd, Augest, 2019

Contents

1. Introduction
   1. Problem Statement
   2. Data
2. Methodology
   1. Exploratory Data Analysis
   2. Feature Engineering
      1. Feature Selection
   3. Prediction
      1. Light Gradient Boosting Method
      2. XgBoost
      3. Random Forest Regression
3. Conclusion

Appendix A – Extra Figures

Appendix B – R Code

Appendix C- Daily Bike Rental Prediction and Code Files

1. Introduction

Cabs are the most preferred means of going around the city. In Today’s world Cab is just a Click away from your Smartphones. In today’s world Ola and Uber have made it super Easy and Cost-Effictive solutions of getting a Cab. These Days, Many Traveller’s prefer Cab for Commuting to office everyday, rather than getting their own transport for work. It can also be a tough time for the Companies like OLA & Uber to do effective Fare distribution for their customers.

* 1. Problem Statement: Forecasting the Cab Fare Rental for a Given Date for a Given Time and Distance Travelled, We would have to predict the Fare for the Ride.
  2. Data – Our task is to build different models which can predict the Cab Fare Rental on daily basis depending upon the given Date & Time and also the Distance travelled.

**Table 1.1 Cab fare Sample Data**

-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

fare\_amount pickup\_datetime pickup\_longitude pickup\_latitude dropoff\_longitude dropoff\_latitude passenger\_count

1 45 2009-06-15 17:26:21 -73.84431 40.72132 -73.84161 40.71228 1

2 169 2010-01-05 16:52:16 -74.01605 40.71130 -73.97927 40.78200 1

3 57 2011-08-18 00:35:00 -73.98274 40.76127 -73.99124 40.75056 2

4 77 2012-04-21 04:30:42 -73.98713 40.73314 -73.99157 40.75809 1

5 53 2010-03-09 07:51:00 -73.96810 40.76801 -73.95665 40.78376 1

6 121 2011-01-06 09:50:45 -74.00096 40.73163 -73.97289 40.75823 1

Based upon all the above factors we have to accurately predict the count of Bikes rented on Daily Basis.

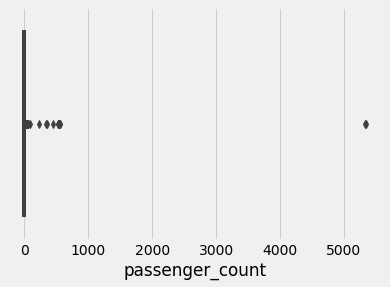
# Chapter 2

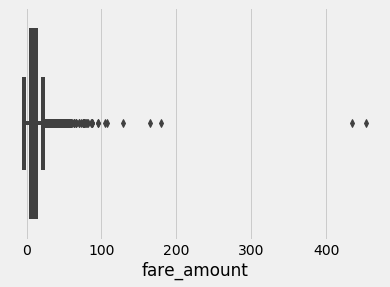
1. **Methodology** 
   1. **Pre-Processing –** Any Predictive modelling requires that we look at the data before we start modelling. However, in data mining terms looking at the data refers to so much that just looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphical representations and plots. This is often called as **Exploratory Data Analysis**. To start this process, we will first try and look at all the probability distribution of the variables. Most analysis like regression, require the data to be normally distributed. We can visualize that in a glance by looking at the probability distributions or probability density functions of the variable.

In Figure 2.1 we have plotted the probability density function of all the Seasonal & Environmental Factors we have in our data as the dependent Cab Rental Count Variable. The blue lines indicate Kernel Density Estimations (KDE) of the variable. While the pointed Blue Line represents the Normal Distribution.

* 1. **Outlier Analysis-** There were some Outliers in the Fare amount and Passenger count. We were able to successfully remove the Outliers. There were many methods which we could follow for removing the outliers.

One of the other steps of pre-processing apart from checking for normality is the presence of outliers. In this case we use a classic approach of removing outliers, Tukey’s method. We visualize the outliers using boxplots.





Other Useful inferences can also be drawn from these plots., if you compare the quality boxplots for each of the predictor variables. We can see that see Fare amount mostly tends to be around 0-20 $. But The maximum fare goes upto little more than 200 $ and the Maximum fare crosses 400$. This could be a useful insight for the business. We might be able to increase the rate by a little margin to enhance profit distribution.

# Feature Engineering

We took the location bounding map for Major Tourist and Travel hub around the New York city such as the Airport, Time Square Museum, LaGuardia Airport.

#jfk

jfk\_lat<-40.6413

jfk\_long<--73.7781

jfk<-c(jfk\_long, jfk\_lat)

#newark

nwk\_lat<-40.6895

nwk\_long<--74.1745

nwk<-c(nwk\_long, nwk\_lat)

#laguardia

lag\_lat<-40.779

lag\_long<--73.8740

lag<-c(lag\_long, lag\_lat)

#MSG

msg\_lat<-40.7505

msg\_long<--73.9934

msg<-c(msg\_long, msg\_lat)

#times square

ts\_lat<-40.7589

ts\_long<--73.9851

ts<-c(ts\_long, ts\_lat)

#freedom tower

freedom\_lat<-40.7127

freedom\_long<--74.0134

freedom<-c(freedom\_long, freedom\_lat)

#empire state building

esb\_lat<-40.7484

esb\_long<--73.9857

esb<-c(esb\_long, esb\_lat)

#grand central

grand\_lat<-40.7527

grand\_long<--73.9772

grand<-c(grand\_long, grand\_lat)

#bronx

bronx\_lat <- (40.837048 \* pi)/180

bronx\_long <- (-73.865433 \* pi)/180

bronx<-c(bronx\_long, bronx\_lat)

nyc<-c(-74.0063889, 40.7141667)

1. We also multiplied the Longitude and the Latitude by pi and then divided by 180 degrees.

**Longitude** is expressed as -**180** degrees (-**pi** radians) to **180** degrees (**pi** radians) with 0 degrees centered at the prime **meridian**. **Latitude** is expressed as -90 degrees (-**pi**/2 radians) to 90 degrees (**pi**/2 radians) with respect to the equator.

We also measured the time of the Day, The Month, The Year in which the Cab is booked.

* Whether, It was a Weekday or Weekend.
* Whether it was Overnight if the Cab was booked.
* Whether, The cab was booked during t5he Day.
* If, The Cab was booked during the evening.
* Weekday or Weekend.
* Hour of the Day.
* Month of Booking.
* Year of Booking.

### **Chapter 3**

**Conclusion**

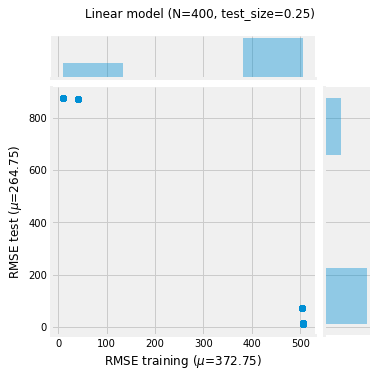
1. **Model Evaluation**

**Now that we have a few models for predicting the target variable, we need to decide which one to choose. There are several criteria that exist for evaluating and comparing models. We can compare the Models using any of the following criteria.**

1. **Predictive Performance**
2. **Interpretability**
3. **Computational Efficiency**

**As we are calculating Cab Rental Count, we need good interpretability. Therefore, we use Predictive Performance as the criteria to compare and Evaluate models.**

**Interpretability can be measured by the Methodology of Understanding a model can provide which can be measure by Metrics such as R Square, Adjusted R square, Mean Absolute Error, etc.**



### Appendix A- Extra Figures

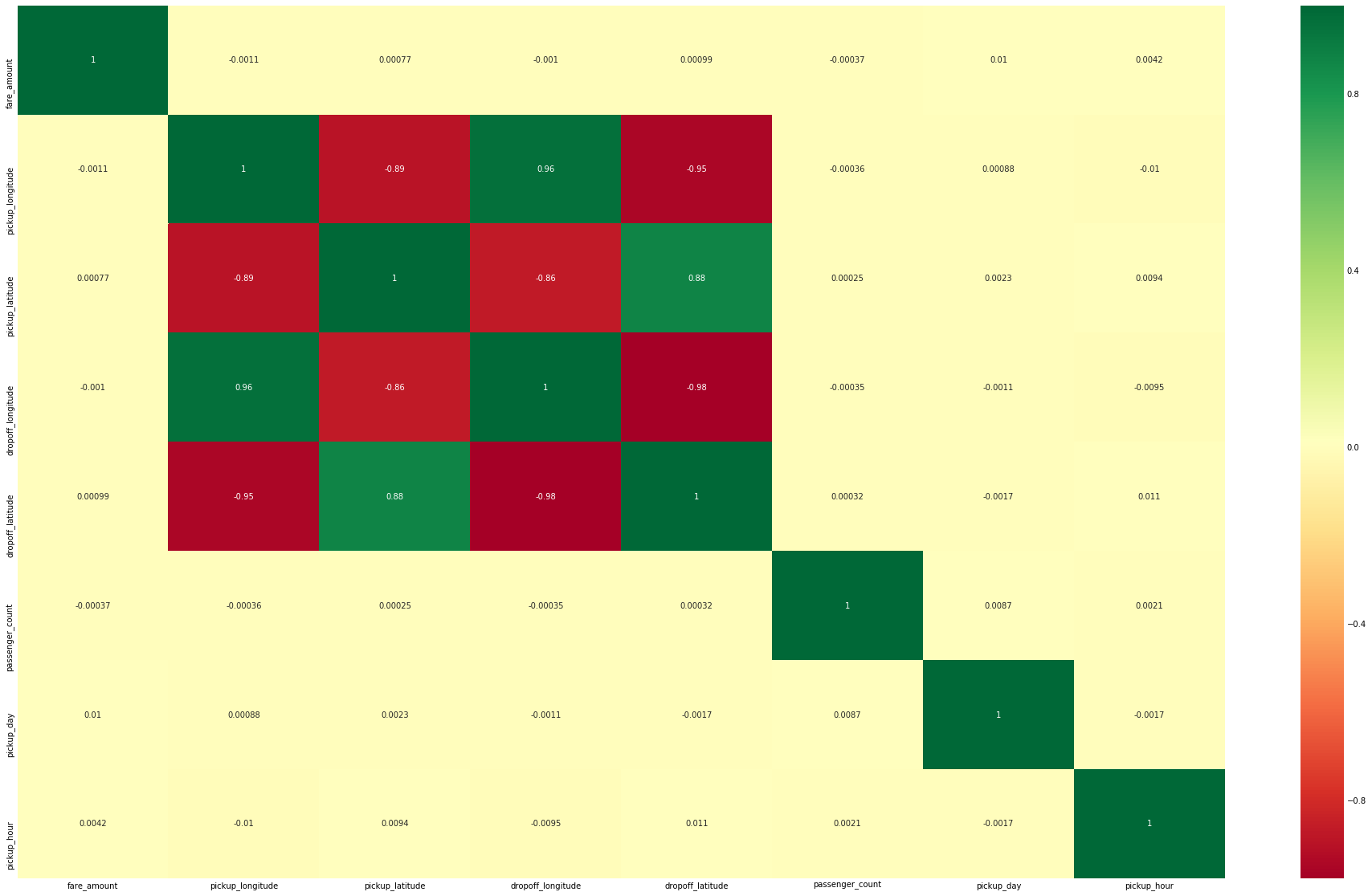


Figure 1.2 HeatMap for our Features

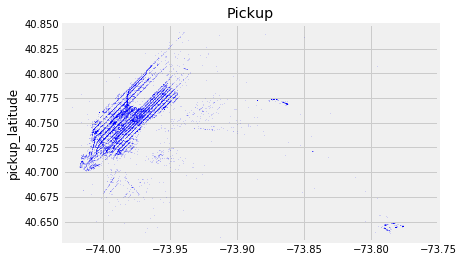


Figure 1.3 Pickup in around NYC

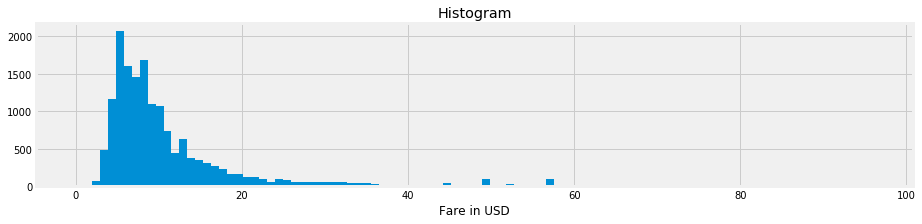


Figure 1.4 Fare Distribution

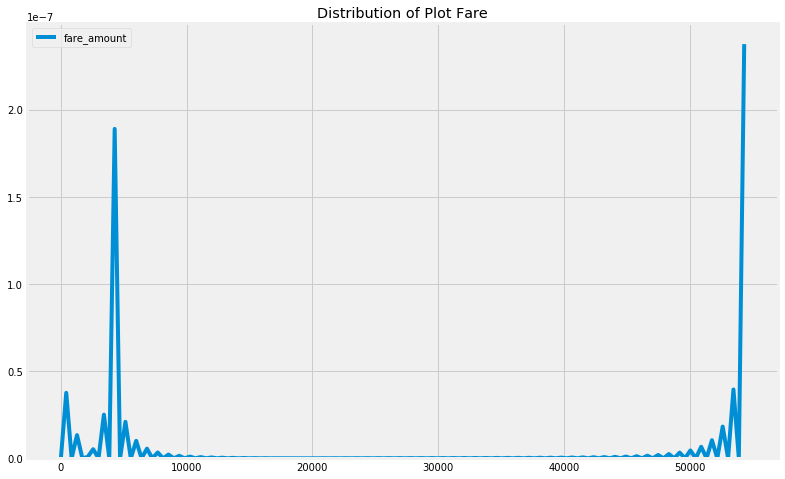


Figure 1.5 Distributiion of Fare

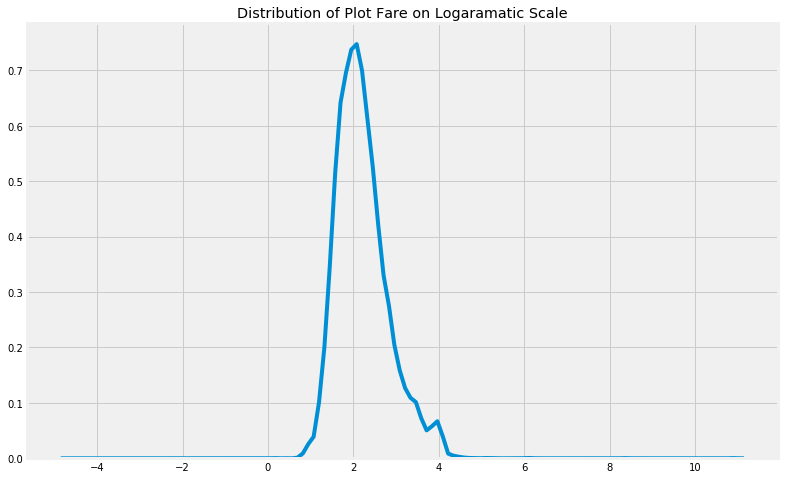


Figure 1.6 Distribution of Plot Fare on Lagrithmic Scale

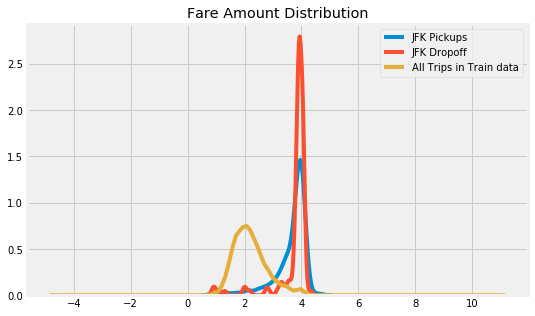


Figure 1.6 Fare Amount Distribution of JFK PUckup and Dropoff with other Trip in the Data

# **Appendix C- Cab Fare Prediction & Code Files**