**Cab Fare Prediction**

*Merapureddy Ajay*

*April 7 2019*

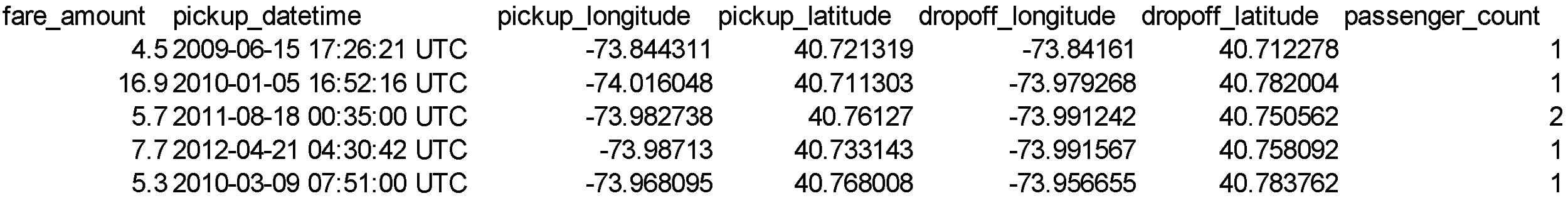
**1.1 Problem Statement** -

The aim of the project is to launch a cab service across the country. I have the historical data from your pilot project and now have a requirement to apply analytics fo fare prediction . I need to design a system that predicts the fare amount for a

cab ride in the city.

# 1.2 Data

Given below a sample of data set that we used to predict fare\_amount.



We will be using given variables to extract relevant features which are used to build models .

In the given dataset target variable is fare\_amount that is to be predicted and rest are independent variables.

2.1 Preprocessing -

Before modeling, we need to explore data.That process often called **Exploratory Data Analysis.** We need to explore the data ,clean the data and make visualizing data through graphs and plots .Most analysis like regression ,require data to be normally distributed. We can visualize that by looking at histogram of the data.

Before proceeding for missing values analysis we look at the structure and summary of the data.

> str(data)

'data.frame' : 16067 obs. of 7 variables:

$ fare\_amount : Factor w/ 469 levels "","-2.5","-2.9",..: 302 59 374 433 371 27 432 57 1 454 ...

$ pickup\_datetime : Factor w/ 16021 levels "2009-01-01 01:31:49 UTC",..: 1115 2509 6550 8252 2919 4986 9743 7479 9838 1606 ...

$ pickup\_longitude : num -73.8 -74 -74 -74 -74 ...

$ pickup\_latitude : num 40.7 40.7 40.8 40.7 40.8 ...

$ dropoff\_longitude: num -73.8 -74 -74 -74 -74 ...

$ dropoff\_latitude : num 40.7 40.8 40.8 40.8 40.8 ...

$ passenger\_count : num 1 1 2 1 1 1 1 1 1 2 ...

**2.1.1 Missing Value Analysis:**

We need to check for missing values in the data.A table is shown below containing two columns named variable and missing\_percent.

|  |  |
| --- | --- |
| variable | missing\_percent |

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 7 | passenger\_count | 0.3423165 |
| 1 | fare\_amount | 0.1555984 |
| 2 | pickup\_datetime | 0.0000000 |
| 3 | pickup\_longitude | 0.0000000 |
| 4 | pickup\_latitude | 0.0000000 |
| 5 | dropoff\_longitude | 0.0000000 |
| 6 | dropoff\_latitude | 0.0000000 |

Based on the missing percentage of the data , we either drop the variable or impute the missing values.If the missing percentage is greater than 30 percent ,then we remove that variable from the data. From the above table it is clear that there are two variables containing

missing values whose percentage is 0.3 % and 0.15%.

As the missing\_ percent is less than 1 percent ,we can safely remove those observations containing missing values .

The shape of the data is (16067,7).

So, after removing the observations containing the missing values ,the shape of the data is

(15987,7)

**2.1.2 Outlier analysis:**

Outliers are observations which are inconsistent with the rest of the data.It is one of the step of preprocessing ,before modeling data.We remove or impute outliers based on the count .

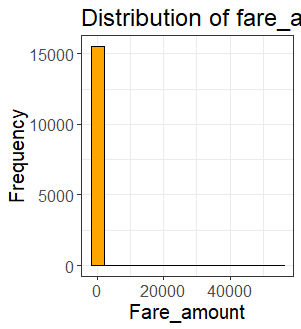
Outliers can be found using boxplot method.Boxplot is a graphical tool for data visualization

Firstly, lets clean the data then we can perform the outlier analysis.

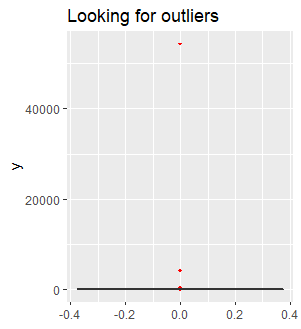
From the given dataset ,we have pickup and dropoff latitude and longitude . Latitude ranges between -90 and +90 .Longitude ranges between -180 and +180.So, we remove observations which are falling outside this range.

We remove observations whose source and destination are same , then we proceed with outlier analysis.

# Histogram of Fare\_amount

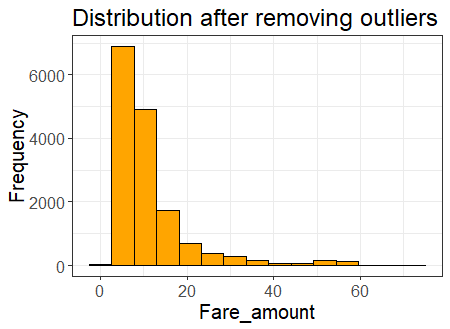


# Boxplot of fare\_amount



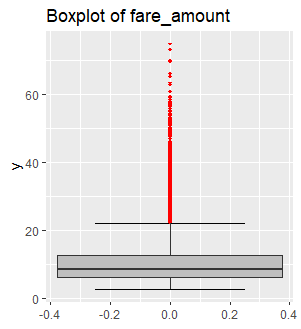
Let’s visualize the histogram and boxplot of fare\_amount after removing the observations containing outliers.

#Histogram of fare\_amount after removing outliers.



We can see the improvement in the distribution of fare\_amount after removing outliers.

# Boxplot of Fare\_amount



Let’s look at the summary of the fare\_amount.

Min. 1st Qu. Median Mean 3rd Qu. Max.

2.50 6.00 8.50 11.23 12.50 75.00

Distance limit : 50 km

Base fare\_amount : 1.5 $ per km

Eventhough there are some outliers in the data , we consider them as fare\_amount can be charged upto 75 dollars.

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.000 1.000 1.000 2.639 2.000 5345.000

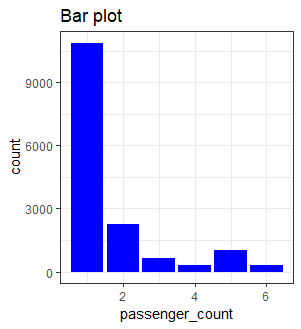
Passenger\_count in a cab can be in between 1 and 6. So , pre-process passenger\_count variable.

Table of passenger\_count after pre-processing

1 2 3 4 5 6

10880 2258 656 317 1018 294

#Bar\_plot of passenger\_count



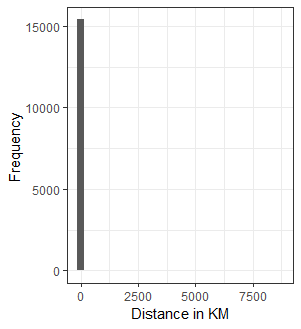
**2.1.3 Feature extraction :**

From the given variable pickup\_datatime ,we can extract various features like hour ,weekday

,day,month and year.And we can extract a new feature distance with the use of pickup and drop-off latitude and longitude using distHaversine() function or by creating a generic function.

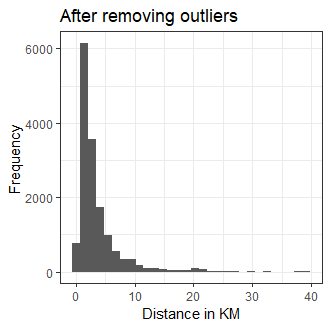
Let’s look at the histogram of distance

# Histogram of distance



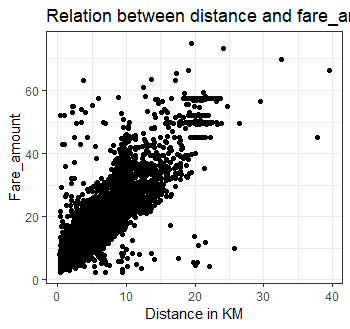
# No one chooses to travel in cab for more than 50km and for less than 300 meters

After processing it , the histogram of distance is shown below



Now , let’s visualize and understand the relation between fare\_amount and newly added features.

#Scatter Plot between distance and fare\_amount



Fare\_amount increases with increase in distance .We check the correlation between distance and fare\_amount using pearson's corelation test.

Pearson's product-moment correlation

data: data$fare\_amount and data$distance

t = 264.2, df = 15183, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.9033964 0.9090809

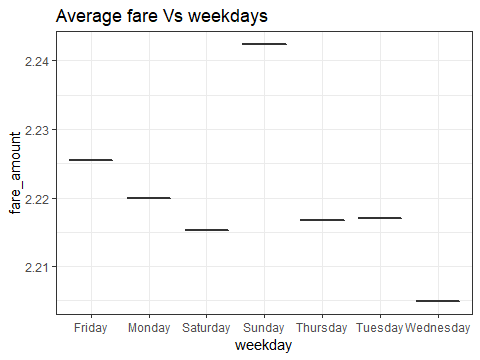
sample estimates:

cor

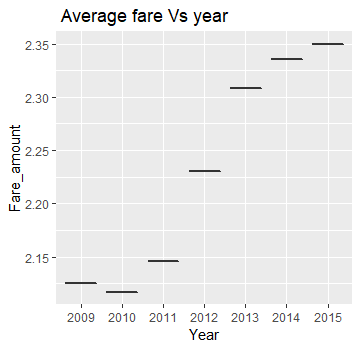
0.9062796

The result shows that there is a is a strong strength of association between fare\_amount and distance.

Now we explore the relation between fare\_amount and features extracted from timestamp variable visually.

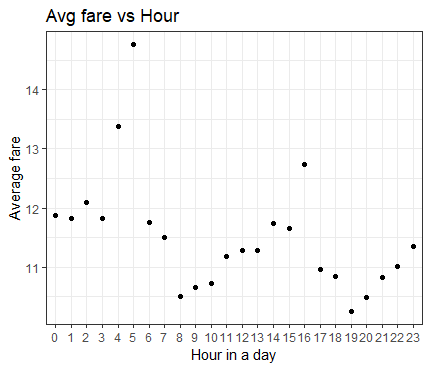


Above plot shows us the average fare\_amount versus the weekdays .Average fare is high on friday and sunday, as many go home at weekends .And pretty uniform on weekdays.



Plot shows us that there is increase in avg fare in some amount every year.

#Scatter plot between average fare and hour in a day



Average fare is high at midnight and early morning and low at 8 am and there is gradual increase in avg fare with time till 4 pm .Hence hour explains about fare\_amount and considered as a feature.

Fare amount can be dynamic with time so , we also consider day and month as features in model building.

**2.2 Modelling :**

Now we have the cleaned data which can be used to bulid models.

We start building models from simple to complex and check the prediction performance of each model .

The first step before model buliding is fixing the size of train and test data.

2.2.1 Starting of with **Multiple Linear Regression:**

After passing train data to lm() function ,it calculates and assign weights to each predictor variable .Look at the part of summary of lr\_model

Residual standard error: 3.581 on 12066 degrees of freedom

Multiple R-squared: 0.843, Adjusted R-squared: 0.8419

F-statistic: 790 on 82 and 12066 DF, p-value: < 2.2e-16

As you can see the Adj. R^2 value, we can explain 84.1% of the data using multiple linear regression.Day and month predictors contribute least for reduction of fitting error of the model.However removing these variables also didnot change the predictive power of our regression model.Therefore , this is the maximum accuracy that we can get from this model.

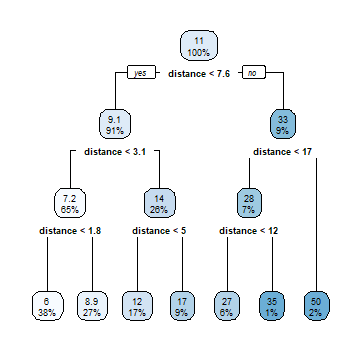
We choose rmse as our error metrics as squared nature of this metric helps to deliver more robust results which prevents the cancelling the positive and negative error values.

**Model evaluation** : Metrics used is rmse (Root Mean Square Error)

Rmse : 3.49

2.2.2 **Regression Trees:**

We will use regression tree to predict the values of our target variable.



The above figure shows the decision tree of the model build based on the input data.

Performance measure : rmse is 4

2.2.3 **Random Forests:**

Random forest is one the supervised machine learning algorithm , which works on the concept of bagging .The result of this algo would be the average of the predicted values of n random decision trees.

The model performed well using categorical variables with one hot encoding.

A one hot encoding allows the representation of categorical data to be more expressive.

Performance measure:

a) with integer encoding of categorical predictors : rmse is 4.1

b) with one hot encoding of categorical predictors : rmse is 3.41

**2.2.4 Stepwise regression:**

Stepwise regression is a method of fitting regression models in which the choice of predictive variables is carried out bu an automatic procedure. Here, after one hot encoding ,we bulid a linear regression model and pass that model to step() function in r , to perform stepwise regression.

Performance Measure: rmse is 3.48

**2.3 Model Selection:**

After looking at the different performance measures of different models , we fix random forest model that we build using categorical variables with one hot encoding .