MODELLING AND ANALYSIS OF DEMAND FOR YELLOW TAXI IN NEW YORK CITY

Submitted by

GROUP 10

BHAGUTHARIVALAN NATARAJAN MUTHUKKANNU

HARISH KANNAN VENKATARAMANAN

PRAVEEN MOHAN

RAJARAJESWARAN CHANDRAMOHAN

IE 515 TRANSPORTATION ANALYTICS

13 MAY 2020

ABSTRACT

The fast-paced and connected world is run by many services and the subtle but vital one is taxi services. The yellow taxi is the heartbeat of economic capital (New York City) and being a part of its culture gives us enough reason to analyse them deeply. The Yellow Taxicab Co. was incorporated in New York on April 4, 1912. Yellow taxi in New York city had roughly around 14000 cabs permitted to operate in the city as per 2014. The taxi takes people from one location to other location within NYC and its demand is influenced by many factors like duration, trip distance, number of passengers, pickup locations, etc. The factors discussed above, and other uncertain factors has become a very important aspect to be discussed and visualized upon. Also, considering external atmospheric conditions as an important factor could lead us to a better predictive model. As a result, it becomes essential to analyse underlying factors and subtle parameters to enhance the existing system to have a better overview on the demand of taxis.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

RESPONSE TO REVIEWS

QUESTION 1

What is the demand to be predicted? per day?

Response: Hourly Demand is predicted in our project.

QUESTION 2 (Group 6)

How to mix classification model and prediction model?

Response: Classification and Regression Tree (CART), Random Forest and Extreme Gradient Boosting (XGBoost) can be used as a Regression Model as well. Specifically, for XGBoost the input needs to be numerical values. For this we have transformed the data into a suitable form by creating dummy variables for all the categorical predictors and converted this data in to a DMatrix form with the function called *xgb.DMatrix()* for training the XGBoost model.

QUESTION 3 (Reza)

RMSE and R_2 are for test or train?

Response: The R_2 and RMSE values presented in the ppt are for the test data for all the models.

QUESTION 4

Enhance the findings and interpretations of the results.

Response: Findings and interpretation of the results have been properly elaborated and enhanced in this report.

CHAPTER 1

INTRODUCTION

The data used in this study were subsets of New York City Taxi and Limousine Commission's trip data, which contains a total observation of around 72 million taxi rides in New York City in the year 2019. This project is focused only on yellow taxis, which operates the most in NYC. For analysis of this study, the data for yellow taxi ride during the month of January to December 2019 were used, although the models were validated on additional data. Since each month consists of about 6 million observations, and there were computational limitations, the invalid and wrong entries were removed, and separate relative data were used for validation. The Weather data is retrieved from Dark-sky (API) and are apprehend to their respective time with the taxi trip data set. Predicting demand for taxis can benefit companies in being prepared for upcoming volatility thus being more profitable.

1.1 PROBLEM STATEMENT

The current prediction of demand for taxis includes many factors but some external factors like atmospheric conditions are not explored enough. Due to which there is a mismatch between demand and supply of taxis which results in cabs running empty at times and other times where there were shortages. This uncertainty in allocating taxis, threatens the foreseeable future.

1.2 RESEARCH OBJECTIVE

The objective of this project is to model and analyse the demand for yellow taxi in NYC taking in account of various parameters. Exploring the multiple parameters for the demand might give a greater insight to make calculated decisions. This visibility will give insight to tackle surge in demand during times where external atmospheric factors come in play.

1.3 SCOPE AND EXPECTED CONTRIBUTION

There are many number of locations (Pickup & Dropoff Zones) in NYC and the demand can be predicted for these locations on a particular day for a particular hour. This project revolves to contribute in matching the supply and demand to provide hassle free communication in Yellow cabs.

CHAPTER 2

LITERATURE REVIEW

Imbalance of availability of taxi is big problem in Big cities. Sometimes, customers need to wait too long for the taxis or sometimes taxi drivers can't find the customers easily. Such problems can be resolved by informed driving, which is a key feature for increasing sustainability for taxi companies [4].

Identifying the factors that influence taxi demand is very important for understanding where and when people use taxis. For effective planning and management of the taxi fleet, understanding what factors drive taxi demand, how taxi use is related to the availability of public transit, and how these patterns vary over space and time is necessary [1].

The work by Schaller (2005) developed multiple regression models to estimate taxi demand for 118 US cities. The study found that the number of workers commuting by subway, the number of households with no vehicles available, and the number of airport taxi trips were strongly correlated and statistically significant in estimating the number of taxi cabs in a given US city. Next, the analysis by Gonzales et al. (2014) modelled taxi pick-ups and drop-offs for specific hours of the day in NYC and Manhattan. The results indicated that population, age, education, income, total jobs were significant in predicting taxi demand in NYC and Manhattan [2].

A particular paper that analysed over 14 million yellow and Uber taxi pick-up samples in NYC found that there is a high predictability of taxi demand (up to 83% in average). In the areas with low predictability, the NN predictor can reach high accuracy by capturing additional features (weather, etc) [5]. A paper discussed about the idea of predicting future demand by considering weather and atmospheric conditions in small districts in an urban area to reduce the empty time in taxi services is combined with improving the spatial-temporal time series models. This idea can be used for reducing traffic congestion [3].

Forecasting and prediction based on zones, along with boroughs will make it very easy for the drivers to be present at the location at any given point of time [6]. One of the papers researched about the changes in weather which affects the number of rides, and customers care about the fare amount along with the availability of the ride [4].

Another paper had a detailed analysis based on the general belief that when it rains or when it is too hot to walk down, people tend to go for a cab rather than walking or going through a subway. Demand-supply levels in taxi services have been studied in the mentioned paper.

Identification and modelling of passenger hot spots for rerouting taxi drivers is also a widely researched area. Some of the commonly employed modelling techniques were the Auto Regressive Integrated Moving Average model (ARIMA) and its variants, Exponential Weighted Average (EWA) models, Nearest Neighbour clustering, and Neural Networks (NN) [7].

Various atmospheric conditions have got a major role to play in predicting the taxi demand. During rainfall, it is generally considered to be undesirable to be outside. As such, individuals who would generally walk or wait for public transit may wish to utilize taxis. Also, during these periods of increased demand, leads to surge pricing (prime time) [8]. Each observation reports the atmospheric conditions of the time period. Possible conditions include “Clear”, “Heavy Rain”, “Rain”, “Light Rain” or “Partly Cloudy” [9]. Surge pricing and prime time pricing schemes are respectively implemented, passengers pay a higher rate for the ride during times of high demand; this higher pricing scheme gives incentives to drivers to provide rides in inclement conditions [10]. Very few studies tried to figure out the taxi demand by considering factors like weather and atmospheric condition into consideration. So, we are trying to determine an influential parameter that determines the taxi demand and providing an alternative perspective with analysis of response (demand) corresponding to individual factors as well.

CHAPTER 3

DATA DESCRIPTION

The raw data comprises of trip records from January 2019 to December 2019 and are explored and visualised by performing Exploratory Data Analysis to gain insight.

3.1 DATA SOURCE

NYC yellow cab data - https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

Weather data - https://darksky.net/dev

3.2 DATA DESCRIPTION

Location ID	ID of the pickup location zone
Month	Number of the month in a year (1 to 12)
Day	Day name of a week
Hour	Hour in a day (00-23)
Demand (**response**)	Number of taxis requested in a given hour
Duration.mins	Trip duration in minutes.
Passengercount	Passenger count excluding driver
Tripdistance.Miles	Distance covered per trip
Amount	Cost of the trip in USD
Temperature.F	Temperature on a particular day in Fahrenheit
Windspeed.MPH	Wind speed in miles per hour
Windgust.MPH	Sudden change in increase of wind speed in miles per hour
Windbearing	Direction of wind in degrees
Visibility.Miles	Distance visible to a human eye in miles

Table 3.2.1 Data Description

3.3 SUMMARY STATISTICS

Fig.3.3.1 describes the raw data summary of January 2019. The raw data consists NA's and some wrong entries. Fore example, the column "total_amount" had negative values and some "trip_distance" values were in thousands. It is insensible for the amount to be in negative and the distance to be in thousands. *Fig. 3.3.2* describes the summary of the refined data set for the whole year (January 2019 to December 2019).

```
> summary(data1)
    VendorID        tpep_pickup_datetime        tpep_dropoff_datetime passenger_count trip_distance      RatecodeID     store_and_fwd_flag  PULocationID    DOLocationID
 Min.   :1.000   2019-01-11 15:15:27:     47   2019-01-20 00:00:00:     69   Min.   :0.000   Min.   :  0.000   Min.   : 1.000   N:7630142          Min.   :  1.0   Min.   :  1.0
 1st Qu.:1.000   2019-01-08 15:15:01:     45   2019-01-30 00:00:00:     67   1st Qu.:1.000   1st Qu.:  0.900   1st Qu.: 1.000   Y:  37650          1st Qu.:130.0   1st Qu.:113.0
 Median :2.000   2019-01-09 15:15:31:     44   2019-01-02 00:00:00:     66   Median :1.000   Median :  1.530   Median : 1.000                      Median :162.0   Median :162.0
 Mean   :1.637   2019-01-10 15:15:08:     44   2019-01-26 00:00:00:     65   Mean   :1.567   Mean   :  2.801   Mean   : 1.058                      Mean   :165.5   Mean   :163.8
 3rd Qu.:2.000   2019-01-12 15:15:33:     37   2019-01-12 00:00:00:     64   3rd Qu.:2.000   3rd Qu.:  2.800   3rd Qu.: 1.000                      3rd Qu.:234.0   3rd Qu.:234.0
 Max.   :4.000   2019-01-12 15:15:34:     35   2019-01-07 00:00:00:     63   Max.   :9.000   Max.   :831.800   Max.   :99.000                      Max.   :265.0   Max.   :265.0
                 (Other)           :7667540   (Other)           :7667398
  payment_type    fare_amount          extra            mta_tax          tip_amount       tolls_amount      improvement_surcharge  total_amount      congestion_surcharge
 Min.   :1.000   Min.   :  -362.0   Min.   :-60.000   Min.   :-0.5000   Min.   :-63.500   Min.   : -70.000   Min.   :-0.3000       Min.   :  -362.8   Min.   :0
 1st Qu.:1.000   1st Qu.:     6.0   1st Qu.:  0.000   1st Qu.: 0.5000   1st Qu.:  0.000   1st Qu.:   0.000   1st Qu.: 0.3000       1st Qu.:     8.2   1st Qu.:0
 Median :1.000   Median :     8.5   Median :  0.000   Median : 0.5000   Median :  1.430   Median :   0.000   Median : 0.3000       Median :    11.3   Median :0
 Mean   :1.292   Mean   :    12.4   Mean   :  0.328   Mean   : 0.4969   Mean   :  1.827   Mean   :   0.317   Mean   : 0.2993       Mean   :    15.7   Mean   :0
 3rd Qu.:2.000   3rd Qu.:    13.5   3rd Qu.:  0.500   3rd Qu.: 0.5000   3rd Qu.:  2.330   3rd Qu.:   0.000   3rd Qu.: 0.3000       3rd Qu.:    16.6   3rd Qu.:0
 Max.   :4.000   Max.   :623259.9   Max.   :535.380   Max.   :60.8000   Max.   :787.250   Max.   :3288.000   Max.   : 0.6000       Max.   :623261.7   Max.   :2
                                                                                                                                                       NA's   :4855978
```

Fig. 3.3.1 Raw Data Summary

```
> summary(data2)
   LocationID        Month              Day            Hour          Demand       Duration.mins.   Passengercount  Tripdistance.Miles.    Amount...      Temperature.F.   Windspeed.MPH.
 Min.   : 48.0   Min.   : 1.000   Friday   :23467   Min.   : 0.00   Min.   :   1.0   Min.   : 1.00   Min.   :1.000   Min.   : 1.030      Min.   :  6.30   Min.   : 2.33   Min.   : 0.47
 1st Qu.:132.0   1st Qu.: 4.000   Monday   :23909   1st Qu.: 6.00   1st Qu.:  89.0   1st Qu.:12.66   1st Qu.:1.510   1st Qu.: 2.550      1st Qu.: 17.71   1st Qu.:41.78   1st Qu.: 4.53
 Median :163.0   Median : 7.000   Saturday :23458   Median :12.00   Median : 187.0   Median :14.94   Median :1.600   Median : 2.920      Median : 19.12   Median :57.73   Median : 6.72
 Mean   :164.4   Mean   : 6.843   Sunday   :23508   Mean   :11.53   Mean   : 201.6   Mean   :16.67   Mean   :1.598   Mean   : 3.969      Mean   : 21.96   Mean   :56.61   Mean   : 7.30
 3rd Qu.:234.0   3rd Qu.:10.000   Thursday :23466   3rd Qu.:18.00   3rd Qu.: 284.0   3rd Qu.:18.29   3rd Qu.:1.680   3rd Qu.: 3.710      3rd Qu.: 21.42   3rd Qu.:71.54   3rd Qu.: 9.51
 Max.   :249.0   Max.   :12.000   Tuesday  :23935   Max.   :23.00   Max.   :1459.0   Max.   :97.50   Max.   :6.000   Max.   :24.300      Max.   : 73.70   Max.   :98.37   Max.   :28.58
                                  Wednesday:23456
 Windgust.MPH.    Windbearing    Visibility.Miles.
 Min.   : 0.82   Min.   :  0.0   Min.   : 0.502
 1st Qu.: 6.38   1st Qu.: 98.0   1st Qu.:10.000
 Median : 9.82   Median :207.0   Median :10.000
 Mean   :11.07   Mean   :193.1   Mean   : 9.460
 3rd Qu.:14.45   3rd Qu.:286.0   3rd Qu.:10.000
 Max.   :55.68   Max.   :359.0   Max.   :10.000
```

Fig. 3.3.2 Final Data Summary

3.4 EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis is a process of conducting initial investigations on data set to identify patterns, anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. Before removing the outliers from the dataset the distribution of predictors are highly skewed which is evident from Fig. 3.4.1. After removing outliers, the distribution of parameters turns out to be normalized (Fig. 3.4.2). Making the predictor more normalized gives us the better model.

The insights that can be derived from the Scatter Matrix Plot are:

- Most of the trips recorded had the passenger count to be in two's and three's, this shows that most of the passengers who opted for the yellow taxi tends to travel in two's and three's the most.
- Most number of trips recorded had passengers travelling less than 30 miles, which makes sense as yellow taxis are commonly used for short trips within NYC.
- Similarly average trip duration lies within fifty minutes.

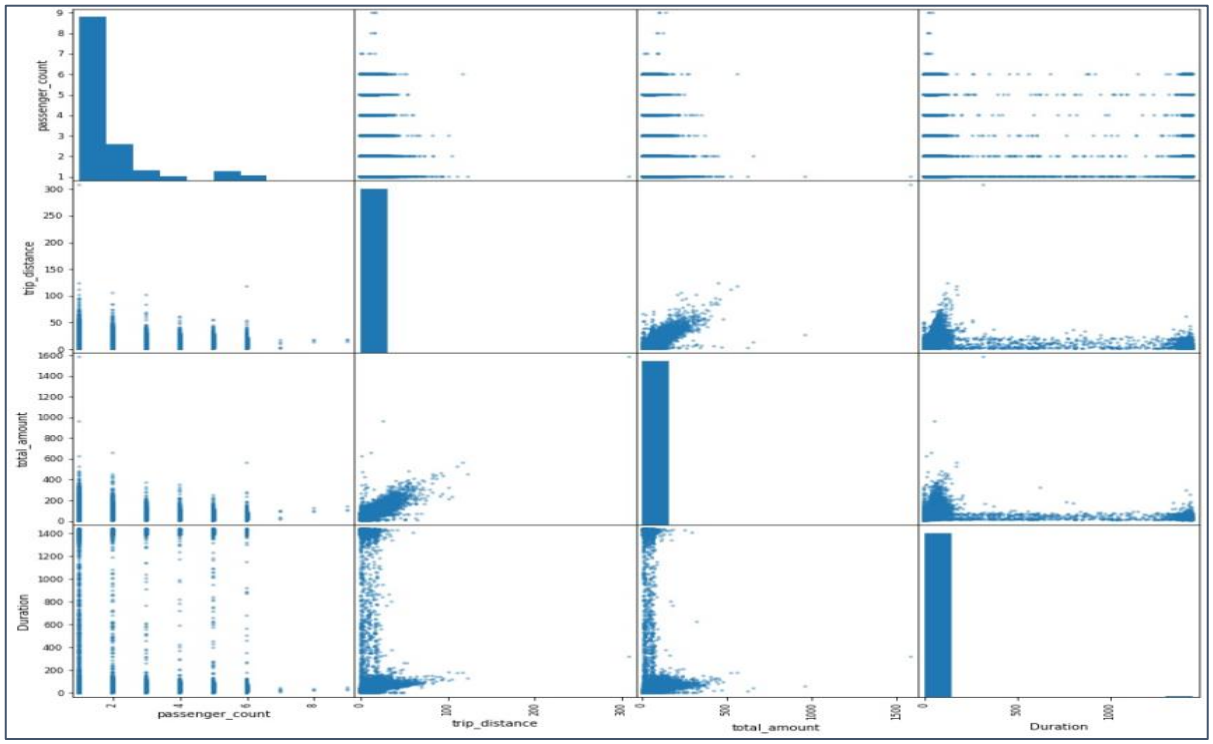


Fig. 3.4.1 Scatter Matrix Plot (Raw Data)

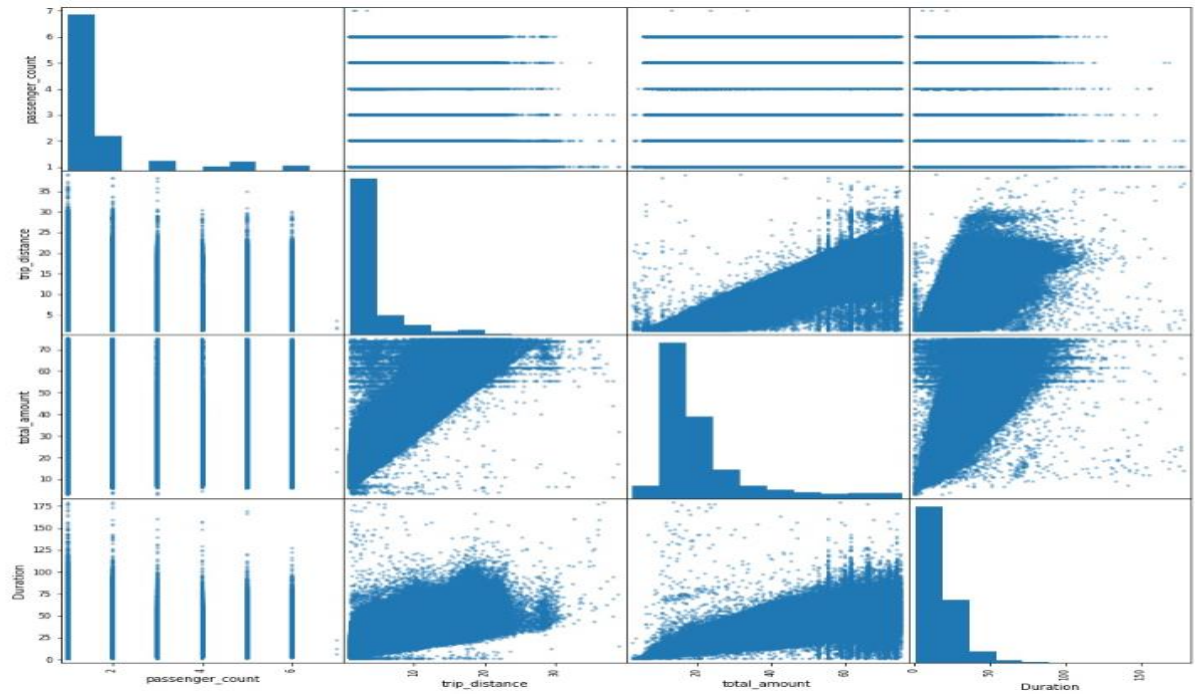


Fig. 3.4.2 Scatter matrix Plot (Final Data)

The line graph (Fig. 3.4.3) shows the monthly fluctuation of demands for the top 5 locations (location with most number of rides in the year 2019) and can be inferred that the demand for yellow taxi was at the peak during the month of March at Madison square.

Minimum demand was observed during the month of January and December. This might be due to the impact of winter vacation. Also, the similar downward trend was observed during the summer vacation.

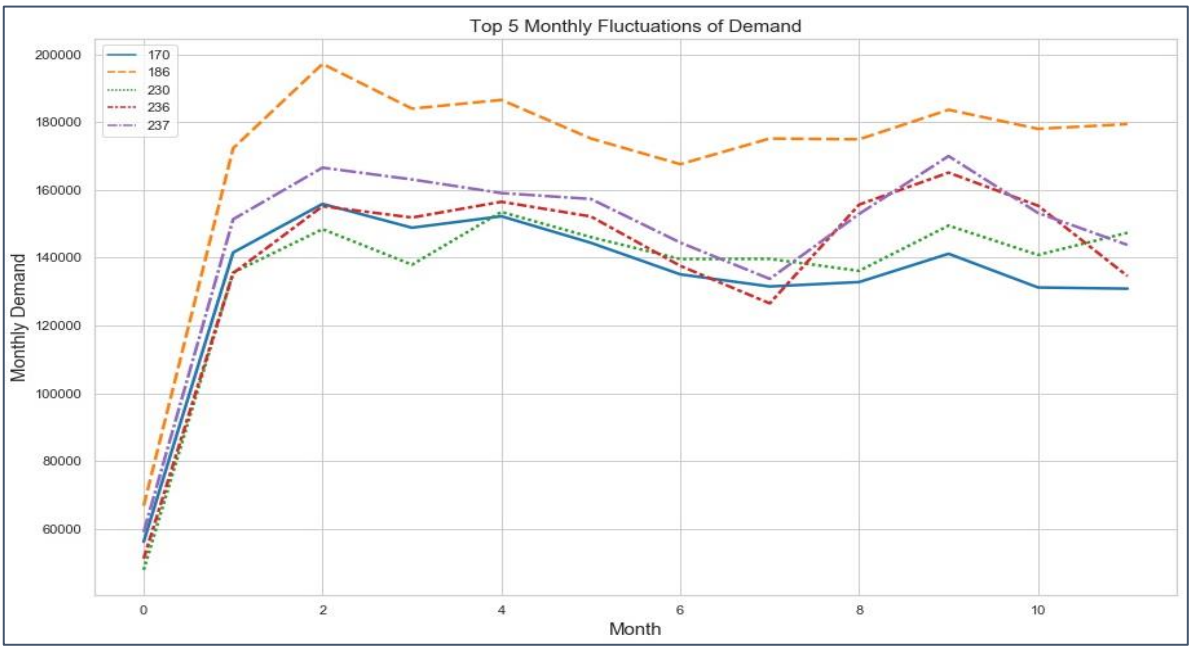


Fig. 3.4.3 Monthly Fluctuation of Demand

The following Heat Map (Fig. 3.4.4) compares weekly vs hourly taxi demand and it is evident that the demand was maximum during weekdays. Also, it can be inferred that demand was at its peak from evening to late in night. It is understood that people prefer to take yellow taxi during night time rather than the public transport.

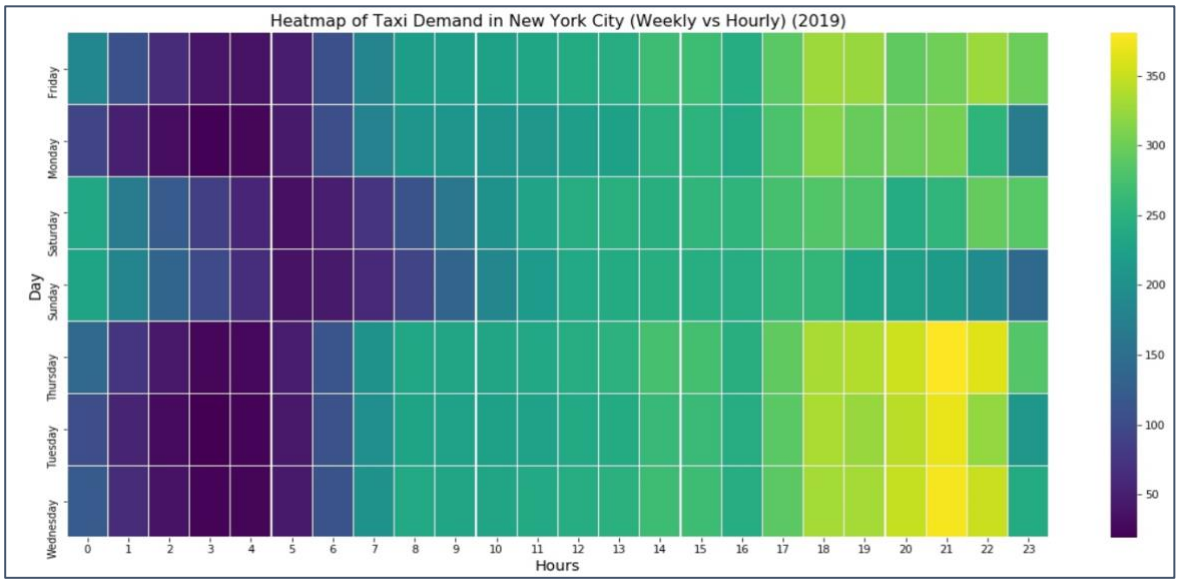


Fig. 3.4.4 Heat Map (Weekly vs Hourly)

From the Temperature vs Demand plot (Fig. 3.4.5) we can see that there is an increase in trend for demand of taxis when the temperature is very low as well as the temperature is very high. This might be because people may decide their mode of transportation depending upon the atmospheric condition.

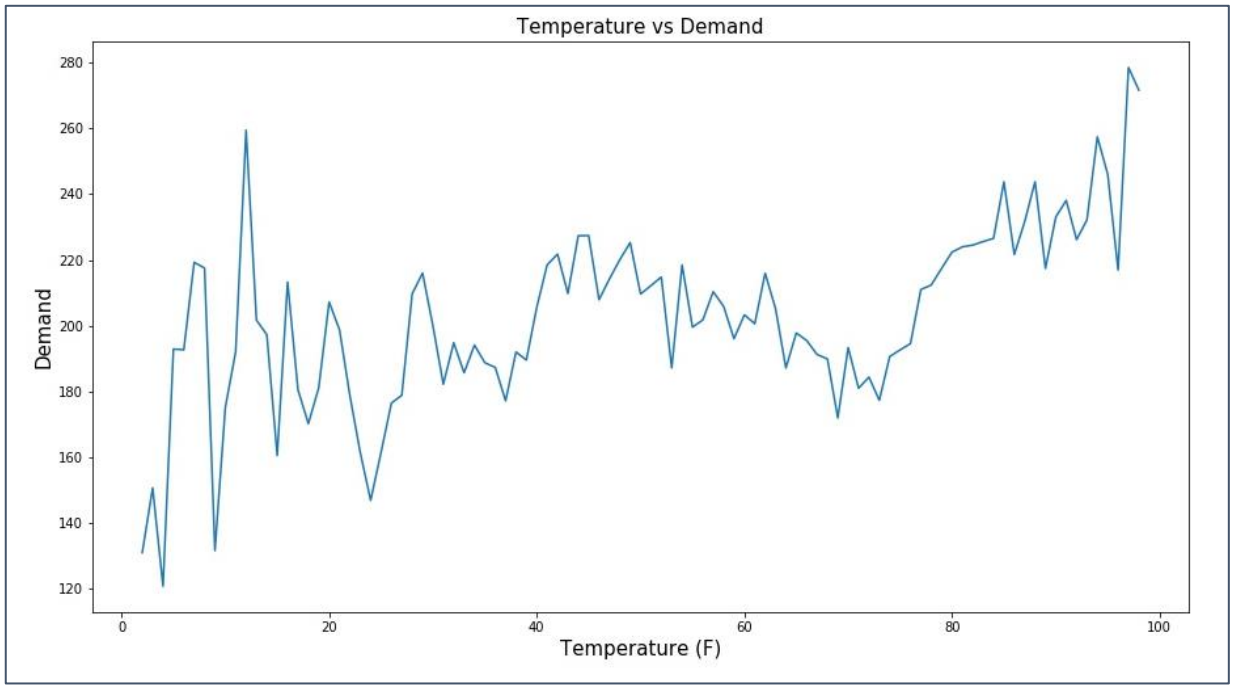


Fig. 3.4.5 Temperature vs Demand

.

CHAPTER 4

MODELLING METHODOLOGIES

4.1 LINEAR REGRESSION

Linear Regression is used for predicting the response variable by the given input variable. This is based on linear relationship between response variable(y) and other variables. This regression is primarily used to establish correlation between multiple variables and response variable to gain deeper insight. The ordinary least square method is used widely in linear regression models and it plots response variable and other variables by minimizing the sum of squared errors.

$$\mathbf{Y = \beta_0 + \beta_1 * X_1 + €}$$

where,

Y – Response Variable

X_1 – Explanatory Variables,

β_0 – Y-Intercept,

€ – Error Term (Residuals),

β_1 – Slope

4.2 CLASSIFICATION AND REGRESSION TREE (CART)

The Classification and Regression tree model is an machine - learning method for predicting data models. It falls under supervised machine learning technique. CART is nothing but if-else statements that are used to produce results based on the available data. It is primarily based on binary tree structure with roots partitions. The root shows input variables (x). The leaf nodes of the tree show an output variable (y) which is used to make a prediction. As our response variable is numeric, we use regression trees of CART to predict the model. During each partition, weighted mean square error is calculated. Based on this weighted mean square error value the child nodes are generated. On continuous generation of these child nodes, the weighted mean square value is reduced and the leaf nodes are generated and the average of this is taken.

4.3 RANDOM FOREST

Random forest is a multi-tree classifier that operates as an ensemble. It works on the various subsets using averaging to improve the accuracy and can control overfitting. Random forest is a supervised learning algorithm and these models are best to beat at performance. Generally random forest has both the hyper-parameters as a decision tree or a bagging classifier. We can deal with the regression by using he algorithm regressor. As the trees grows, Random forest will add randomness to model. Here in random tree, instead of searching for the most important feature while splitting the node, it looks for best feature among the random subset features. The main advantage of Random forest is it can classify large number of data with better prediction accuracy. As overfitting is a big problem in machine learning, this model can control overfitting. The main drawback of Random forest would be, as the number of trees increases the algorithm becomes too slow and runtime is affected. Finally, random tree is simple, and perfect tool with some limitations.

4.4 EXTREME GRADIENT BOOSTING

Extreme gradient boosting works in such a way that the individual trees are built by using more weights on instance with wrong predictions and high errors. This model learns from its past mistakes. A special case of boosting where errors are minimized by gradient descent algorithm that is it removes less qualified. XGBoost and Gradient Boost both ensemble tree methods that apply the principle of boosting weak learners (CART generally) using the gradient descent architecture. XGBoost algorithm works fast even for large and complex datasets and highly efficient in compare to other tree models such as Random Forest, CART, etc. However, XGBoost improves upon the base framework through systems optimization and algorithmic enhancements.

CHAPTER 5

MODEL FRAMEWORK

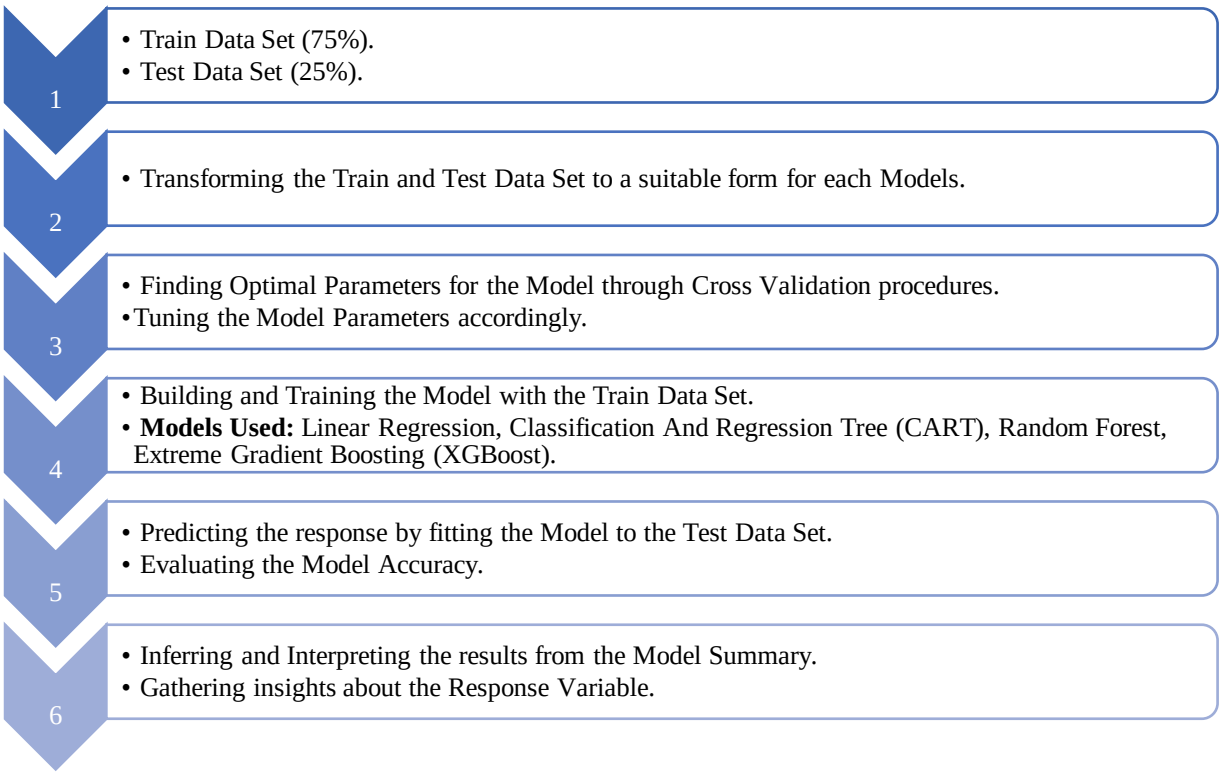


Fig. 5.1 Model Framework

Initially the data set is randomly split into train and test data set. Train data set has 75% of data and test has 25% of data. Depending upon each model requirements, the data set is transformed to a suitable form so that the train data can fit into the model. The models used are Linear Regression, CART, Random Forest and Extreme Gradient Boosting.

Cross-validation procedures and hyper parameter tuning were performed for each of the models and optimal parameter values were found. Later the train data set is fitted into the model and the models were tested with the test data set. Then the corresponding R_2 and RMSE values were calculated in evaluating the performance and accuracy of each model.

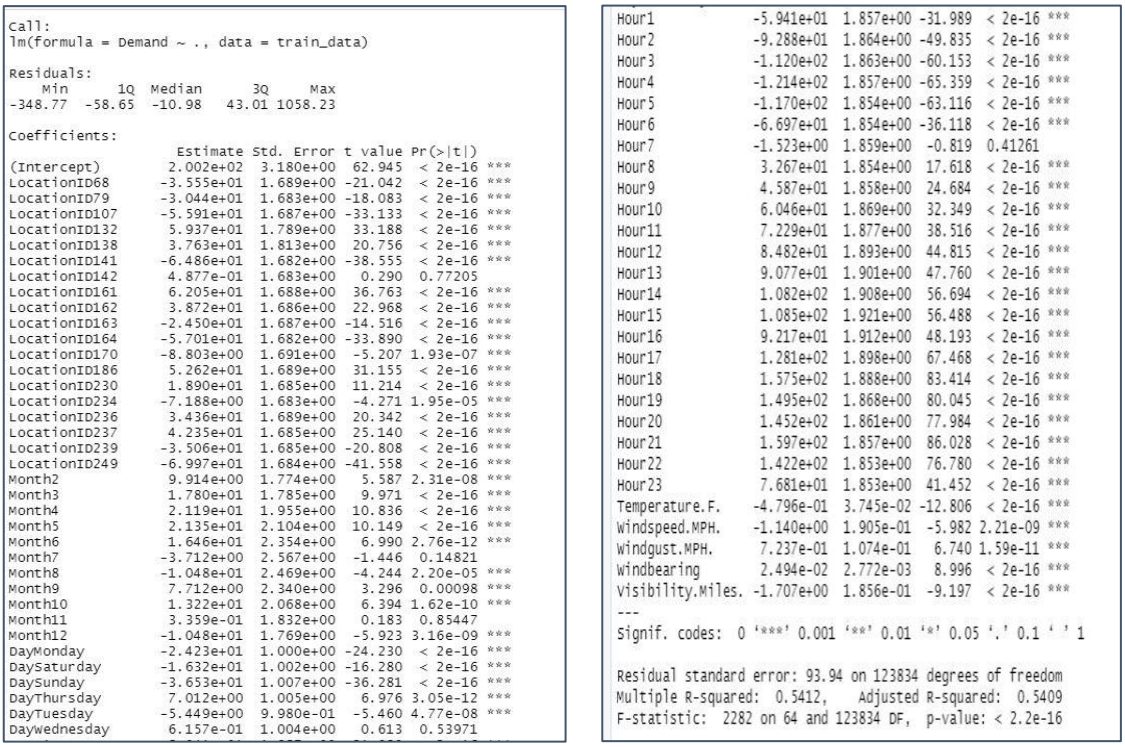
Finally, the model with best accuracy and performance is chosen and the results were interpreted and concluded accordingly.

CHAPTER 6

STUDY RESULTS

6.1 LINER REGRESSION MODEL

From the model summary (Fig. 6.1.1), we can obtain each factors contribution and significance in explaining the variation of the response variable. By training the LR Model, we have achieved RMSE of 93.93, R_2 of 0.5412 and Adjusted R_2 of 0.5409 for the train data. After testing the model with the test data, we have achieved a test RMSE and R_2 of 95.6695 and 0.5332249 respectively. The poor R_2 and RMSE value explains that the Linear model isn't suitable for predicting the response in this data set. Even though the model accuracy is not as expected, we can gain other statistical insights from the model summary. It is evident that atmospheric conditions like Temperature, Wind Speed, Wing Gust, Wind Bearing and Visibility are statistically significant and therefore these factors does have an impact in predicting the demand. Also, there are some statistically insignificant factors like some categories from Location, Month, Day and Hour.



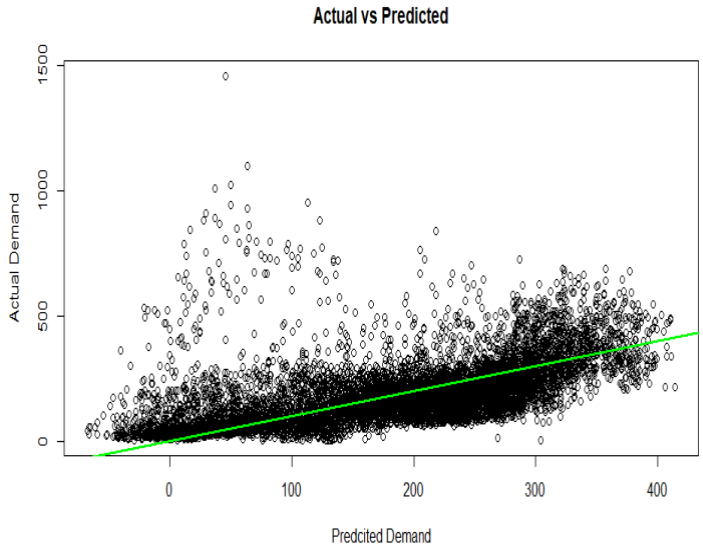
```
Call:
lm(formula = Demand ~ ., data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-348.77  -58.65  -10.98   43.01 1058.23

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.002e+02  3.180e+00  62.945  < 2e-16 ***
LocationID68     -3.555e+01  1.689e+00 -21.042  < 2e-16 ***
LocationID79     -3.044e+01  1.683e+00 -18.083  < 2e-16 ***
LocationID107    -5.591e+01  1.687e+00 -33.133  < 2e-16 ***
LocationID132     5.937e+01  1.789e+00  33.188  < 2e-16 ***
LocationID138     3.763e+01  1.813e+00  20.756  < 2e-16 ***
LocationID141    -6.486e+01  1.682e+00 -38.555  < 2e-16 ***
LocationID142     4.877e-01  1.683e+00   0.290  0.77205
LocationID161     6.205e+01  1.688e+00  36.763  < 2e-16 ***
LocationID162     3.872e+01  1.686e+00  22.968  < 2e-16 ***
LocationID163    -2.450e+01  1.687e+00 -14.516  < 2e-16 ***
LocationID164    -5.701e+01  1.682e+00 -33.890  < 2e-16 ***
LocationID170    -8.803e+00  1.691e+00  -5.207 1.93e-07 ***
LocationID186     5.262e+01  1.689e+00  31.155  < 2e-16 ***
LocationID230     1.890e+01  1.685e+00  11.214  < 2e-16 ***
LocationID234    -7.188e+00  1.683e+00  -4.271 1.95e-05 ***
LocationID236     3.436e+01  1.689e+00  20.342  < 2e-16 ***
LocationID237     4.235e+01  1.685e+00  25.140  < 2e-16 ***
LocationID239    -3.506e+01  1.685e+00 -20.808  < 2e-16 ***
LocationID249    -6.997e+01  1.684e+00 -41.558  < 2e-16 ***
Month2            9.914e+00  1.774e+00   5.587 2.31e-08 ***
Month3            1.780e+01  1.785e+00   9.971  < 2e-16 ***
Month4            2.119e+01  1.955e+00  10.836  < 2e-16 ***
Month5            2.135e+01  2.104e+00  10.149  < 2e-16 ***
Month6            1.646e+01  2.354e+00   6.990 2.76e-12 ***
Month7           -3.712e+00  2.567e+00  -1.446  0.14821
Month8           -1.048e+01  2.469e+00  -4.244 2.20e-05 ***
Month9            7.712e+00  2.340e+00   3.296  0.00098 ***
Month10           1.322e+01  2.068e+00   6.394 1.62e-10 ***
Month11           3.359e-01  1.832e+00   0.183  0.85447
Month12          -1.048e+01  1.769e+00  -5.923 3.16e-09 ***
DayMonday        -2.423e+01  1.000e+00 -24.230  < 2e-16 ***
DaySaturday      -1.632e+01  1.002e+00 -16.280  < 2e-16 ***
DaySunday        -3.653e+01  1.007e+00 -36.281  < 2e-16 ***
DayThursday       7.012e+00  1.005e+00   6.976 3.05e-12 ***
DayTuesday       -5.449e+00  9.980e-01  -5.460 4.77e-08 ***
DayWednesday      6.157e-01  1.004e+00   0.613  0.53971
Hour1            -5.941e+01  1.857e+00 -31.989  < 2e-16 ***
Hour2            -9.288e+01  1.864e+00 -49.835  < 2e-16 ***
Hour3            -1.120e+02  1.863e+00 -60.153  < 2e-16 ***
Hour4            -1.214e+02  1.857e+00 -65.359  < 2e-16 ***
Hour5            -1.170e+02  1.854e+00 -63.116  < 2e-16 ***
Hour6            -6.697e+01  1.854e+00 -36.118  < 2e-16 ***
Hour7            -1.523e+00  1.859e+00  -0.819  0.41261
Hour8             3.267e+01  1.854e+00  17.618  < 2e-16 ***
Hour9             4.587e+01  1.858e+00  24.684  < 2e-16 ***
Hour10            6.046e+01  1.869e+00  32.349  < 2e-16 ***
Hour11            7.229e+01  1.877e+00  38.516  < 2e-16 ***
Hour12            8.482e+01  1.893e+00  44.815  < 2e-16 ***
Hour13            9.077e+01  1.901e+00  47.760  < 2e-16 ***
Hour14            1.082e+02  1.908e+00  56.694  < 2e-16 ***
Hour15            1.085e+02  1.921e+00  56.488  < 2e-16 ***
Hour16            9.217e+01  1.912e+00  48.193  < 2e-16 ***
Hour17            1.281e+02  1.898e+00  67.468  < 2e-16 ***
Hour18            1.575e+02  1.888e+00  83.414  < 2e-16 ***
Hour19            1.495e+02  1.868e+00  80.045  < 2e-16 ***
Hour20            1.452e+02  1.861e+00  77.984  < 2e-16 ***
Hour21            1.597e+02  1.857e+00  86.028  < 2e-16 ***
Hour22            1.422e+02  1.853e+00  76.780  < 2e-16 ***
Hour23            7.681e+01  1.853e+00  41.452  < 2e-16 ***
Temperature.F.   -4.796e-01  3.745e-02 -12.806  < 2e-16 ***
Windspeed.MPH.   -1.140e+00  1.905e-01  -5.982 2.21e-09 ***
Windgust.MPH.     7.237e-01  1.074e-01   6.740 1.59e-11 ***
Windbearing       2.494e-02  2.772e-03   8.996  < 2e-16 ***
Visibility.Miles. -1.707e+00  1.856e-01  -9.197  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 93.94 on 123834 degrees of freedom
Multiple R-squared:  0.5412,    Adjusted R-squared:  0.5409
F-statistic:  2282 on 64 and 123834 DF,  p-value: < 2.2e-16
```

Fig. 6.1.1 LR Model Summary



TEST DATA

	RMSE	Rsquare
1	95.6659	0.5332249

Fig. 6.1.2 Actual vs Predicted

From the above plot between Actual response and Predicted response, most of the data points were scatter around the best fit line and significantly a greater number of data points lies too far away from the best fit line and this explains the poor RMSE and R^2 value. Hence Linear Regression isn't a suitable model and to further improve our prediction, we have decided to perform other models.

6.2 THE CART MODEL

The Classification and Regression Tree (CART) Model obtained us the following results (Fig. 6.2.1). In summary, from Variable importance sections we can understand that the factors Hour, Location ID, Windspeed, Day and Wind gust were the most significant factors in predicting the response. Also we can infer the total number of nodes generated at each stage and the number of split and relative error at each nodes.

After testing the model, we have achieved a slight improvement in the RMSE and R^2 when compared to LR Model. Here we have obtained a RMSE of 92.38536 and an R^2 of 0.5646891. Even though the there is an improvement, the accuracy of the model isn't enough. Hence there is a need for us to try other advanced models.

```
> summary(cart)
Call:
rpart(formula = Demand ~ ., data = train_data, method = "anova")
  n= 123899

          CP nsplit rel error    xerror        xstd
1 0.33683841      0 1.0000000 1.0000239 0.004654295
2 0.10709163      1 0.6631616 0.6631880 0.004271745
3 0.02868218      2 0.5560700 0.5567960 0.003878659
4 0.02682664      3 0.5273878 0.5249038 0.003734003
5 0.02078889      4 0.5005611 0.5012854 0.003595746
6 0.01474666      5 0.4797723 0.4827574 0.003525613
7 0.01290251      7 0.4502789 0.4512876 0.003363346
8 0.01213361      8 0.4373764 0.4393461 0.003357821
9 0.01000000      9 0.4252428 0.4293483 0.003319633

Variable importance
          Hour     LocationID Windspeed.MPH.            Day  Windgust.MPH.
            68             25              3              2              1

Node number 1: 123899 observations,    complexity param=0.3368384
  mean=201.1816, MSE=19223.04
  left son=2 (41047 obs) right son=3 (82852 obs)

Node number 2: 41047 observations,    complexity param=0.02868218
  mean=86.85882, MSE=9277.395
  left son=4 (30787 obs) right son=5 (10260 obs)
```

TEST DATA

```
      RMSE   Rsquare
1 92.38536 0.5646891
```

Fig. 6.2.1 CART Model Summary

6.3 RANDOM FOREST MODEL

Fig. 6.3.1 represents the summary and results of Random Forest Model. The RF model seems to perform efficiently when compared to the previous models. The optimal parameter values were identified by trial and error method and we have set the Tree Size as 50 and the Node Size as 10 for our RF Model. Therefore, from the results we can say that the model was able to explain 88.47% variation in our response variable. From the importance summary and the variable importance plot we can identify the variables that were most significant in the model. It is shows that the factor Hour has the most significant IncNodePurity value.

IncNodePurity quantifies the importance of these variables in a numerical form. Finally after testing the model we have achieved a considerably significant improvement of our model accuracy. The model obtained a R_2 of 0.8897974 and RMSE of 46.48354. Therefore the RF

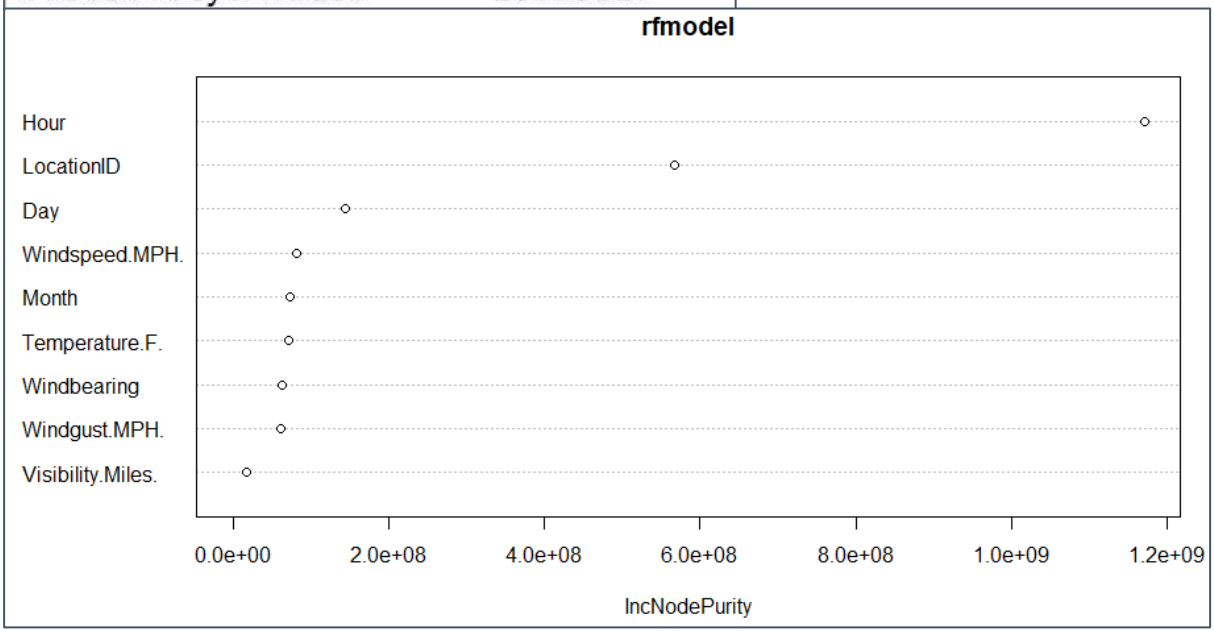
Model has achieved the best model accuracy so far, but in order to further improve our prediction we have decided to use more advanced tree based model.

```
> rfmodel

Call:
 randomForest(x = rfx, y = train_data$Demand, ntree = 50, nodesize = 10)
               Type of random forest: regression
                     Number of trees: 50
No. of variables tried at each split: 3

          Mean of squared residuals: 2217.047
                    % Var explained: 88.47
```

```
> importance(rfmodel)
                   IncNodePurity
LocationID             566371520
Month                   73413008
Day                    144256705
Hour                  1170559655
Temperature.F.          70236002
Windspeed.MPH.          81395942
Windgust.MPH.           61317710
Windbearing             61748760
Visibility.Miles.       16099017
```



TEST DATA

```
      RMSE   Rsquare
1 46.48354 0.8897974
```

Fig. 6.3.1 RF Model Summary

6.4 EXTREME GRADIENT BOOSTING (XG BOOST)

In XG Boost model the data needs to be in the form of numerical and also in DMatrix form. Initially we have converted the actual data into a numerical only form by creating dummy variables for the categorical variables. Then this data is further converted into DMatrix form with help of a function called *xgb.DMatrix()*. This function converts the input data into a form which is suitable for training the XGB Model. Simultaneously the Label i.e. the response variable is created by parsing the response variable from the actual data.

6.4.1 TRAIL RUN

Initially we conducted a trail run of the model with 5 iterations in order to visualise the descent in the RMSE value and to make informed decision in performing Cross – Validation procedures. After 5 iterations we have obtained a RMSE of 109.6013, which is worse than the LR Model. Therefore, this proves that it is important to perform Cross – Validation procedures and Hyper Parameter Tuning in finding the optimal values for the input parameters of the model to achieve better accuracy. The Fig. 6.4.1.1 represents the results obtained from the trail run.

```
> xg_trail
##### xgb.Booster
raw: 19.9 Kb
call:
  xgb.train(params = params, data = dtrain, nrounds = nrounds,
    watchlist = watchlist, verbose = verbose, print_every_n = print_every_n,
    early_stopping_rounds = early_stopping_rounds, maximize = maximize,
    save_period = save_period, save_name = save_name, xgb_model = xgb_model,
    callbacks = callbacks)
params (as set within xgb.train):
  silent = "1"
xgb.attributes:
  niter
callbacks:
  cb.print.evaluation(period = print_every_n)
  cb.evaluation.log()
# of features: 68
niter: 5
nfeatures : 68
evaluation_log:
 iter train_rmse
    1   189.0590
    2   154.6388
    3   131.6893
    4   118.0710
    5   109.6013
```

Fig. 6.4.1.1 Trail Run Summary

6.4.2 PARAMETER TUNING

XG Boost package contains an inbuilt function called *xgb.cv()*. We have used this function to perform the cross validation with number of folds as 5 and number of iterations (nrounds) as 5000, to find the region at which there is no improvement in the Test RMSE. Fig. 6.4.2.1 represents the results of the 5000 iterations. While performing these iterations only the train data is fed into the model and the function correspondingly splits the train data again as train and test to run the respective number of iterations. Hence the initial test data has never fed to the model while parameter tuning and model training. Fig. 6.4.2.2 represents the optimal region for selecting the value for the parameter nrounds. While training the model we have set the nrounds as1200.

```
> xgbcv
##### xgb.cv 5-folds
    iter train_rmse_mean train_rmse_std test_rmse_mean test_rmse_std
       1     189.062982     0.17032595     189.07136     0.6303392
       2     154.640750     0.20400215     154.65633     0.7235479
       3     131.637973     0.21216843     131.72567     0.8340708
       4     118.012210     0.17790404     118.16867     0.8847549
       5     109.638730     0.18634174     109.76388     1.0221359
---
    4996       8.062113     0.04486976      40.17822     0.3576077
    4997       8.059264     0.04479327      40.17857     0.3571611
    4998       8.058018     0.04478532      40.17860     0.3573501
    4999       8.056914     0.04507360      40.17846     0.3573435
    5000       8.054824     0.04451647      40.17871     0.3572806
```

Fig. 6.4.2.1 XGB.CV Results

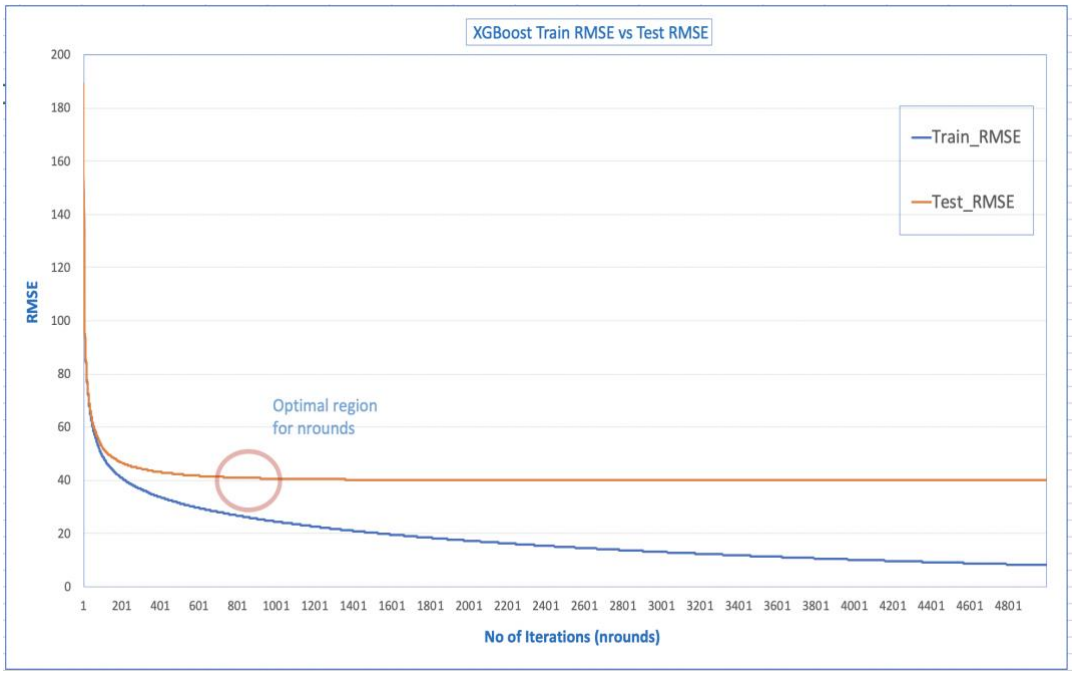


Fig. 6.4.2.2 Nrounds Optimal Region

6.4.3 MODEL

```
> xg
##### xgb.Booster
raw: 4.5 Mb
call:
  xgb.train(params = params, data = dtrain, nrounds = nrounds,
    watchlist = watchlist, verbose = verbose, print_every_n = print_every_n,
    early_stopping_rounds = early_stopping_rounds, maximize = maximize,
    save_period = save_period, save_name = save_name, xgb_model = xgb_model,
    callbacks = callbacks)
params (as set within xgb.train):
  silent = "1"
xgb.attributes:
  niter
callbacks:
  cb.print.evaluation(period = print_every_n)
  cb.evaluation.log()
# of features: 68
niter: 1200
nfeatures : 68
evaluation_log:
    iter train_rmse
       1  189.05904
       2  154.63882
---
    1199   23.96046
    1200   23.94482
```

TEST DATA

	RMSE	Rsquare
1	39.08002	0.9197254

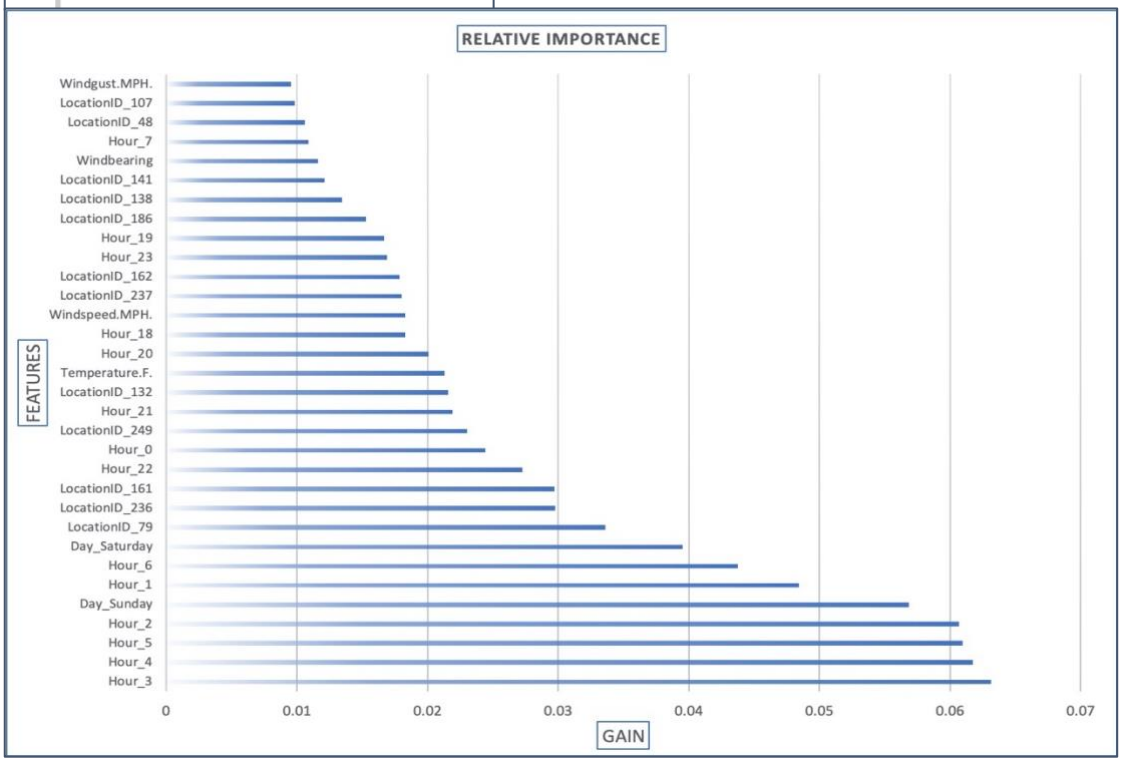


Fig. 6.4.3.1 XG Boost Model Summary

The XGB Model is trained with the optimal parameter values obtained from the cross validation procedure. Summary and results from the model are represented in Fig. 6.4.3.1. At 1200_{th} iteration we have obtained a Train RMSE of 23.94482. Then the model performance is tested with the test data and we have achieved a R_2 of 0.919 and a RMSE of 39.08. The model have performed well and have achieved good accuracy when compared to other models.

Also from the relative importance plot we can identify the factor which had significant important in the model in predicting the Demand. Therefore, we have decided to infer the results and make conclusion from the XGB Model because of its high accuracy comparatively.

CHAPTER 7

RESULT INTERPRETATION

S.No.	MODEL	RMSE	R Sq
1	Linear Regression	95.6659	0.5332249
2	CART	92.38536	0.5646891
3	Random Forest	46.48354	0.8897974
4	Extreme Gradient Boosting	39.08002	0.9197254

Table 7.1 Model Results

7.1 CONCLUDING REMARKS

- This project focuses upon studying the behaviour of people who utilize yellow taxi in NYC and predict the demand by considering various factor like atmospheric conditions, etc.
- There were a total of 255 different pickup and drop-off location zones in NYC. Due to computational limitations, we have decided to predict the demand for the top 20 zones were most number of trips were recorded.
- For these 20 zones, exploratory data analysis was conducted. From the EDA we have identified unique patterns and anomalies in taxi demand with respect to people's behaviour and atmospheric conditions.
- Informed decision were made in choosing, tuning, training and testing the models and the results were interpreted based upon the best model. The model with best accuracy seems to be the Extreme Gradient Boosting (XG Boost).
- The XGB Model states that the factors Temperature, Wind Speed, Hour, Day and Pickup Location are the most significant factors in predicting the demand. Therefore, it is critical for taxi companies to consider these factors in allocating cabs for various locations in NYC.

7.2 LIMITATIONS

- Only pickup locations were considered in this model. This restricts us in predicting the demand uniquely between two different zones i.e. node to node demand.
- Due to limited availability of sources for temperature data, we have taken daily weather data rather than hourly data. This restricts us in predicting the demand for a specific hourly atmospheric condition.
- Since we have considered only the top 20 zones, its limits us in predicting the demand for other locations.

7.3 FUTURE SCOPE

- We can consider special occasions throughout the year in predicting the demand.
- Further we can consider the drop-off locations in the model and predict the demand between various zones within NYC.
- We can implement advanced Machine Learning Algorithms like Spatial Temporal Algorithm, Recurrent Neural Network, Graph Models, etc., in predicting and identifying pattern between nodes (Pickup & Drop-off Zones).
- Using data from the different vehicles including private and public transports and efficiently predicting the traffic flow during extreme weather condition and congestions. It may offer transportation authorities an upper hand in solving traffic problem and take precaution during tough times.

Team Contributions

1. **Bhagutharivalan Natarajan Muthukkannu – Data Collection and Literature Review**
2. **Harish Kannan Venkataramanan – Exploratory Data Analysis**
3. **Rajarajeswaran Chandramohan - Data Cleaning and Future Scope**
4. **Praveen Mohan – Model Building**

CHAPTER 8

REFERENCE

1. Yazici, M. A., C. Kamga, and K. C. Mouskos. Analysis of Travel Time Reliability in New York City Based on Day-of-Week Time-of-Day Periods. In Transportation Research Record: Journal of the Transportation Research Board, No. 2308, Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 83–95.
2. ZhuoYanga,c,MarkL.Franzb,d,ShanjiangZhua,c,JinaMahmoudib,e,ArefehNasrib,e,LeiZhang. Analysis of Washington, DC taxi demand using GPS and land-use data.
3. Sabiheh Sadat Faghih, Understanding and Modeling Taxi Demand Using Time Series Models. A dissertation submitted to the Graduate Faculty in Civil Engineering in partial fulfillment of the requirements for the degree of Doctor of Philosophy.
4. Saumya Shah, Harshil Shah, Taxi Demand Prediction System
5. Kai Zhao1, Denis Khryashchev3, Juliana Freire1,2, Cl´audio Silva1,2, and Huy Vo1,3 1Center for Urban Science and Progress, New York University, Predicting Taxi Demand at High Spatial Resolution: Approaching the Limit of Predictability.
6. Sreejita Biswas, Oklahoma State University, Stillwater, Oklahoma, Taxi Ride Prediction: Does The Yellow Cab Supply Meet Customer Demands?, MWSUG 2019 – Paper BL – 068.
7. Neema Davis, Gaurav Raina, Krishna Jagannathan, Taxi demand forecasting: A HEDGE based tessellation strategy for improved accuracy.
8. Schaller, B., " Regression Model of the Number of Taxicabs in the United States" Journal of Public Transportation, Vol.8(5), 2005, pp. 63-78.
9. Abel Brodeur., "An empirical analysis of taxi, Lyft and Uber rides: Evidence from weather shocks in NYC". Presented at Journal of economic behaviour and organization.
10. Cravo, V. S., Cohen. J. E., "The impact of weather on transit revenue in New York City". Presented at the 88th Annual Meeting of the Transportation Research Board, Washington D.C., 2009.