**Project report on**

**CAB FARE PREDICTION Submitted By**

**AKASH SHARMA**

Contact Details-+91-8949122676, [akashsharma8875@gmail.com](mailto:akashsharma8875@gmail.com)

Contents

[Introduction 4](#_Toc78578774)

[Problem Statement 4](#_Toc78578775)

[Data Review 4](#_Toc78578776)

[Exploratory Data Analysis 4](#_Toc78578777)

[Assumptions with the data 5](#_Toc78578778)

[Data Cleaning and Missing Value Analysis 6](#_Toc78578779)

[Visual observations 11](#_Toc78578780)

[Feature Scaling 15](#_Toc78578781)

[Model Selection 18](#_Toc78578782)

[Linear Regression Model 19](#_Toc78578783)

[Gradient Boosting Model 20](#_Toc78578784)

[Decision Tree Model 21](#_Toc78578785)

[Random Forest Model 22](#_Toc78578786)

[New Results Optimizations with Parameters Tuning 23](#_Toc78578787)

[CONCLUSION 27](#_Toc78578788)

[Appendix 29](#_Toc78578789)

[**Figure 1:Training data of Cab fare** 5](#_Toc78499095)

[**Figure 2:Data types of given variables** 5](#_Toc78499096)

[**Figure 3:Data type conversion** 7](#_Toc78499097)

[**Figure 4:Differntiated DateTime** 7](#_Toc78499098)

[**Figure 5:Limiting the longitude and Latitude by dropping some rows** 8](#_Toc78499099)

[**Figure 6:Visualization of Lat long and Pickup time on Tableau** 9](#_Toc78499100)

[**Figure 7:Setting Up the fare amount** 10](#_Toc78499101)

[**Figure 8:Setting up the passenger count** 11](#_Toc78499102)

[**Figure 9:Fare amount vs Passenger count** 12](#_Toc78499103)

[**Figure 10: Cab Period Time observation in a particular hour** 12](#_Toc78499104)

[**Figure 11: Time Fare relation Visualization** 13](#_Toc78499105)

[**Figure 12; Visualization of Trip Fare** 13](#_Toc78499106)

[**Figure 13:Fare and days of week visualization** 14](#_Toc78499107)

[**Figure 14: Impact of the day on the number of cab rides** 14](#_Toc78499108)

[**Figure 15: Relationship between distance and fare** 15](#_Toc78499109)

[**Figure 16:Skewness in fare amount and distance before applying the log transformation** 16](#_Toc78499110)

[**Figure 17:Skewness after applying the log transformation** 17](#_Toc78499111)

[**Figure 18:Skewness of distance in the test set before applying a log transformation** 18](#_Toc78499112)

[**Figure 19:Skewness after applying Log Transformation** 18](#_Toc78499113)

[**Figure 20:Findings of RMSE and R-Squared in LR model** 20](#_Toc78499114)

[**Figure 21:Findings of RMSE and R-square in GB model** 21](#_Toc78499115)

[**Figure 22:Findings of RMSE and R-Squared using Decision Tree Model** 22](#_Toc78499116)

# Introduction

## Problem Statement

You are a cab rental start-up company. You have successfully run the pilot project and now want to launch your cab service across the country. You have collected the historical data from your pilot project and now have a requirement to apply analytics for fare prediction. You need to design a system that predicts the fare amount for a cab ride in the city.

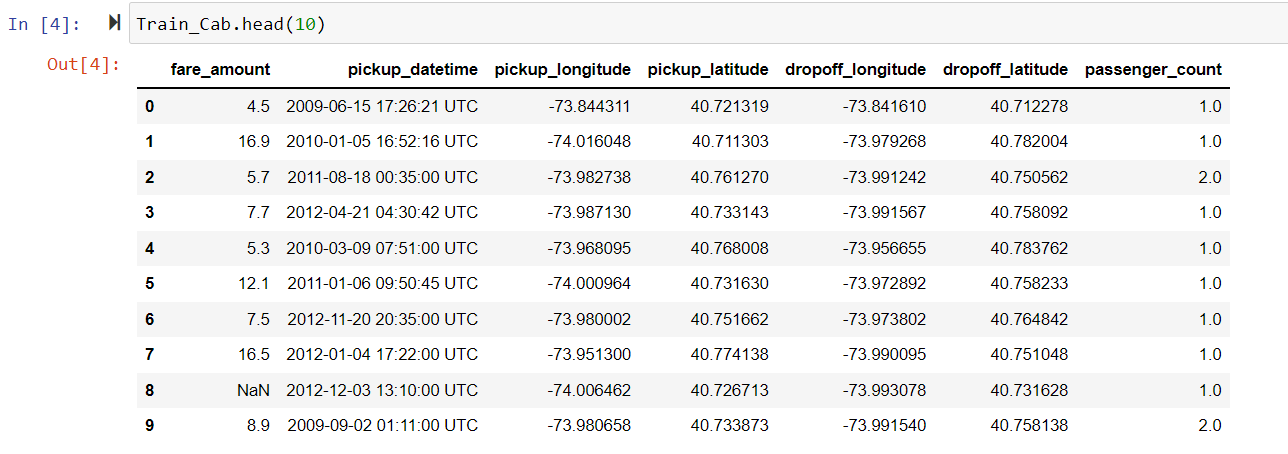
## Data Review

Before starting any new model and project, data reviewing is a very essential step because it helps to find what kind of columns are described and going to be used in further processing. Mainly, there are six attributes information are given with the problem statement which is mentioned below:

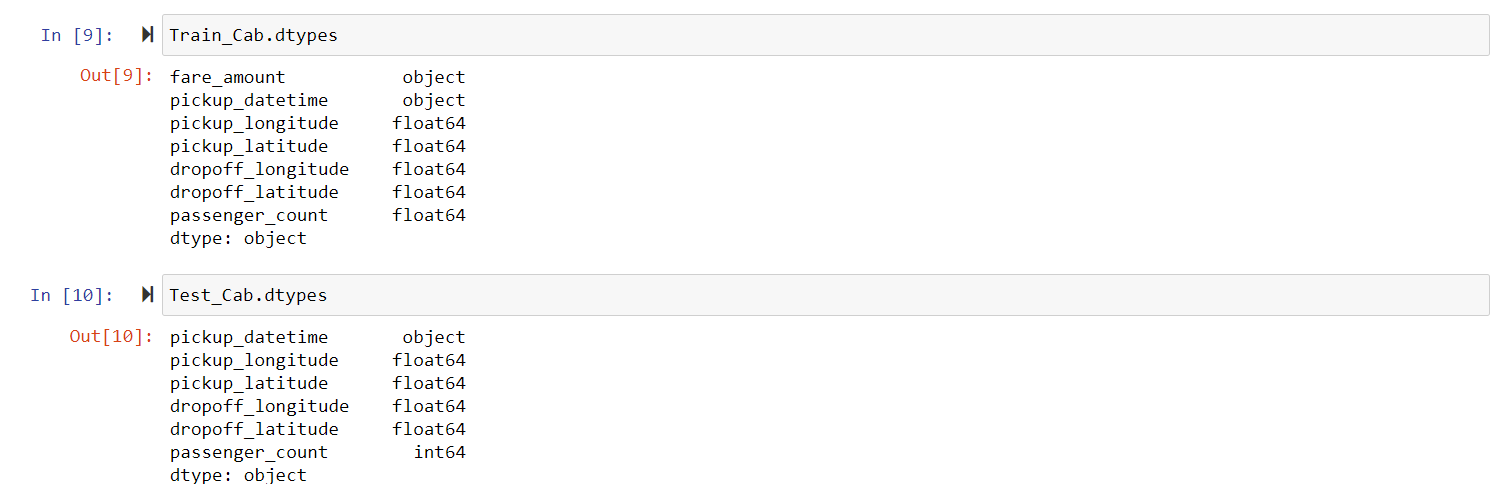
* *pickup\_datetime* - timestamp value indicating when the cab ride started.
* *pickup\_longitude* - float for longitude coordinate of where the cab ride started.
* *pickup\_latitude* - float for latitude coordinate of where the cab ride started.
* *dropoff\_longitude* - float for longitude coordinate of where the cab ride ended.
* *dropoff\_latitude* - float for latitude coordinate of where the cab ride ended.
* *passenger\_coun*t - an integer indicating the number of passengers in the cab ride.

## Exploratory Data Analysis

The train and test cab data are provided for the effective processing of prediction analysis. The data given here is provided from 2009 to 2015 which seems unstructured. The exploration and data understanding is the most important part of any project so as it is done here. Herewith some steps like, data understanding, data visualization, and data cleaning, a better approach of predictions of fare is done. In the below Fig.1, we can see that there are some training data set columns that are going to be used in further analysis.



**Figure 1:Training data of Cab fare**



**Figure 2:Data types of given variables**

In the above Fig.2, the datatypes of given attributes can be seen which needs to be changed. The reason behind changing the data type is the name and the summary of some continuous and categorical variables. Fare amount and distance are the continuous variables and the rest of the columns are categorical variables. Fare amount in $ is our targeted variable which should be afloat as it can be seen that it is not available in the dataset.

## Assumptions with the data

It can be stated in the above statement that target variables may affect the independent variables so there might be multiple assumptions related to the dataset.

1. It can be assumed that the fare amount is dependent on how much time will it take to travel from one place to another place because the traffic may differ according to daytime and night time so apparently, it will affect the fare amount.
2. Let us take a position from New York and it is cleared that with a city like New York, there are multiple cabs available so possibly, the fare will be less
3. The fare amount is very much dependent on the trip distance which can be calculated from the longitude and latitude of pickup and dropout.

Let’s jump on the effective step of this project which is missing value and data cleaning.

# Data Cleaning and Missing Value Analysis

Effective data is a vital part of the analytics of the analytical process which is used to prepare and validate the data. A short word that can be usezd for the data cleaning and missing values analysis process is Garbage in-Garbage out (GIGO). Overall, it is needed to create an effective foundation for the analysis process.

Here are some considerations are given below:

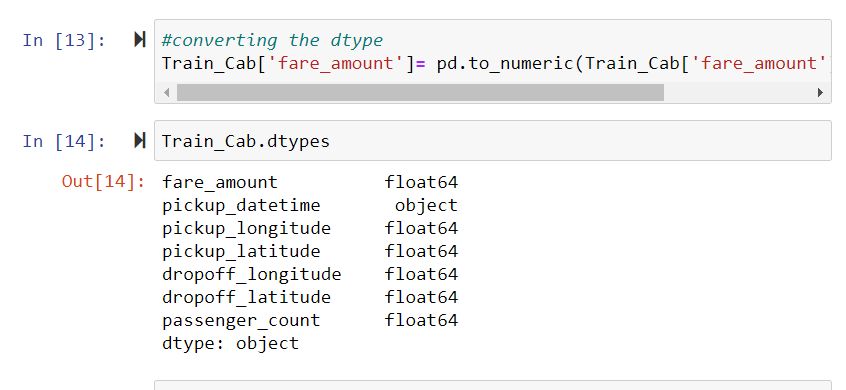
1.Passenger count should not be exceeded to 6(Note: without observations of SUV).

2.The outlier in pickup\_datetime column of value 43.

3.As per the data set, the longitudes range varies from -180 to 180, and latitudes ranges vary from -90 to 90.

Some steps of data cleaning and missing values analysis

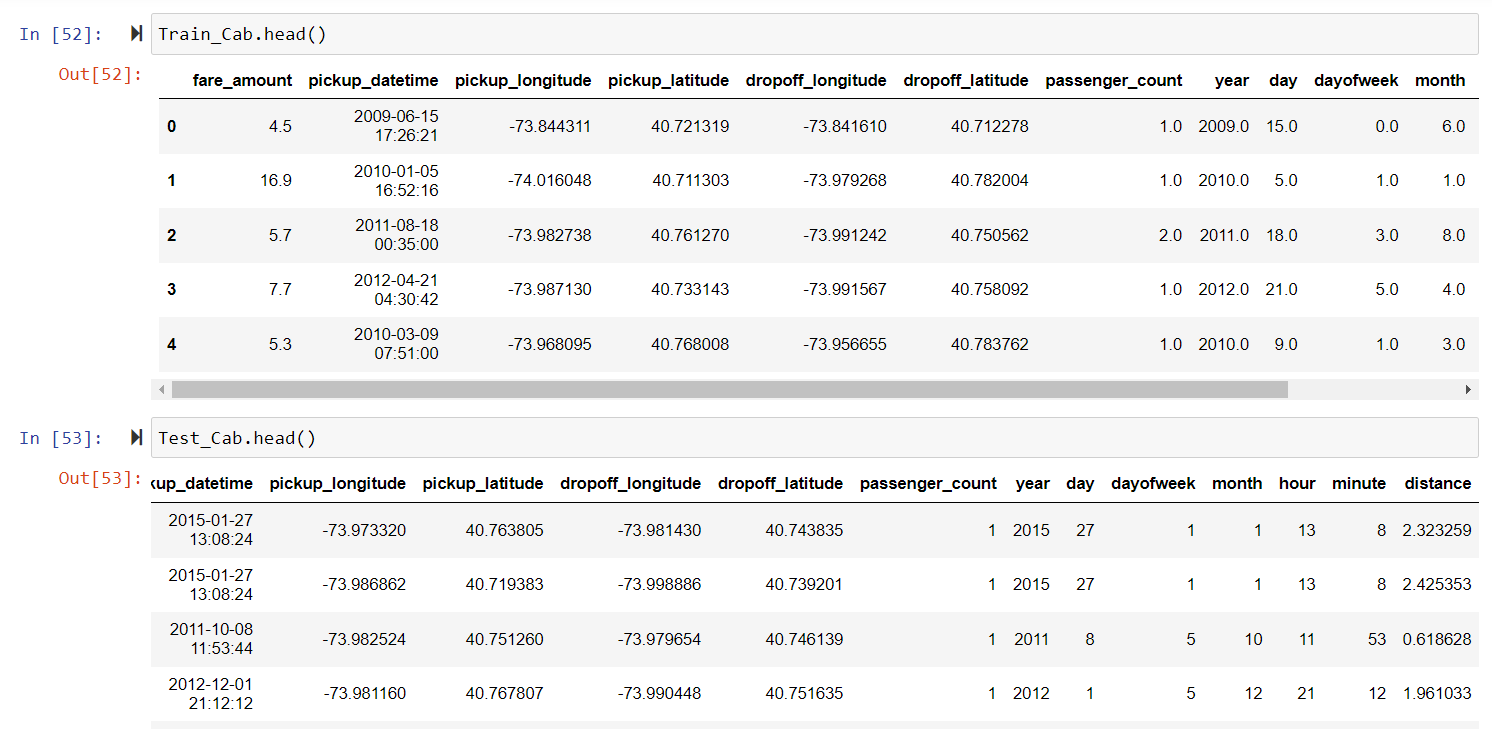
**Data types Conversion**-As it has been mentioned in the above scripts that the data types need to be changed as it doesn’t match with the targeted variables so it has been changed with the help of required steps that have been mentioned in Fig.3



**Figure 3:Data type conversion**

Further, the pickup DateTime datatypes were also converted in Datetime

**Pickup\_Datetime Conversion**-The training dataset has 1 target and 6 independent variables and if it is discussed, it is observed that the journey was started like 2009-06-15,17:26:21 UTC which can be differentiated. The Datetime data can be differentiated in the year, month, day, day of the month, hour, minute, and second.



**Figure 4:Differntiated DateTime**

Fig.4 shows the differentiated DateTime which is more understanding with the purposes.

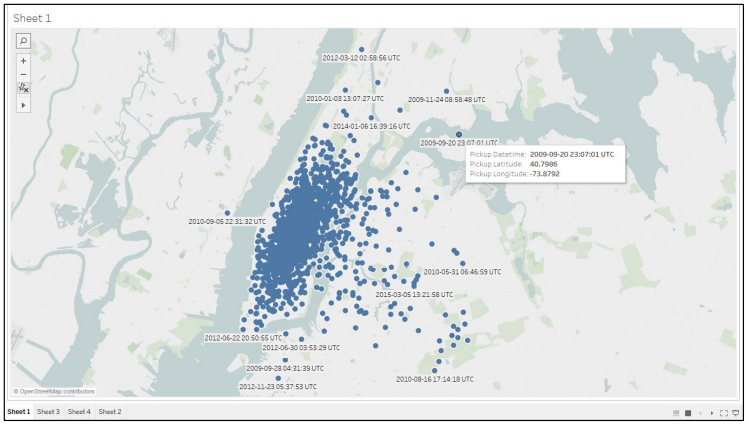
Here in fig.5 shows a pickup time and fare amount analysis with the help of Tableau which shows that the passenger count differs with the fare amount. The high fare is chosen by fewer passengers. The fare has a huge difference if it is compared between 2009 to 2015.

**Pickup and Dropout location**-The distance here is described herein as longitude and latitude which is related to both pickup and drop-off. Firstly, the data was fixed by setting some limits related to it. The latitude and longitude points must be entered within a particular range [-90 to 90] and [-180 to 180] because the increased value of these points shows that the point lies in an ocean or water. Sadly, the cab cannot run on the water so these values are nothing but outliers.

Here are some snaps of these steps with explanations below in Fig.5:



**Figure 5:Limiting the longitude and Latitude by dropping some rows**



**Figure 6:Visualization of Lat long and Pickup time on Tableau**

Above Fig.6 shows that the most frequent pickup location is the airport area which means the airport area comes in the range where the cab frequency is more.

With the setting and fixing the limitation upon data, the trip distance was found out using the Haversine formula which is used to calculate the shortest distance between two points on a sphere. The formula is shown below:

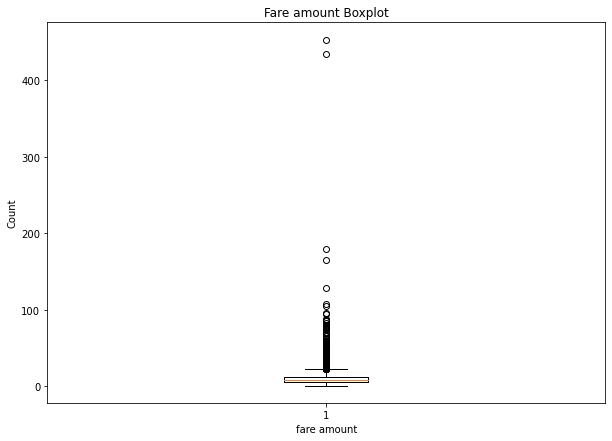
a = sin²(Δφ/2) + cos φ1 ⋅ cos φ2 ⋅ sin²(Δλ/2)

c = 2 ⋅ atan2(√a, √(1−a))

d = R ⋅ c

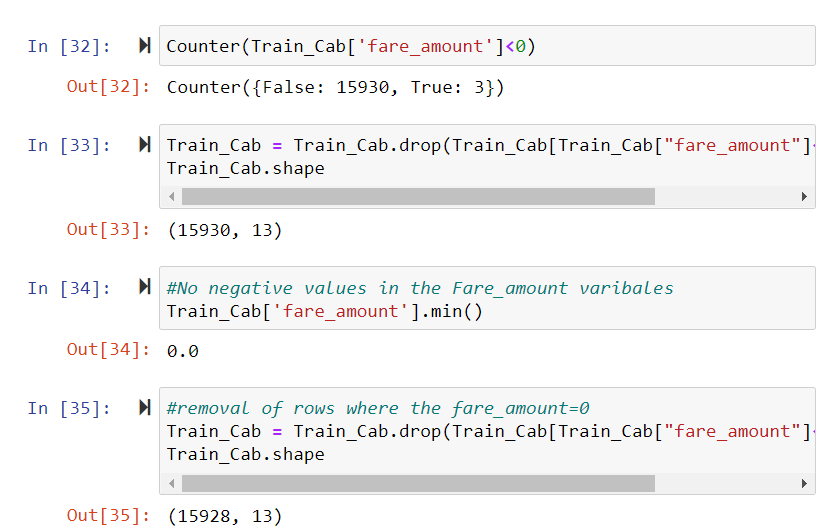
Where φ is latitude in radian, λ is longitude, R is earth’s radius.

The data was converted in km with the help of the Haversine formula for further data analysis, the shape distance was limited to 129kms because it was considered as the outlier for the further process.[



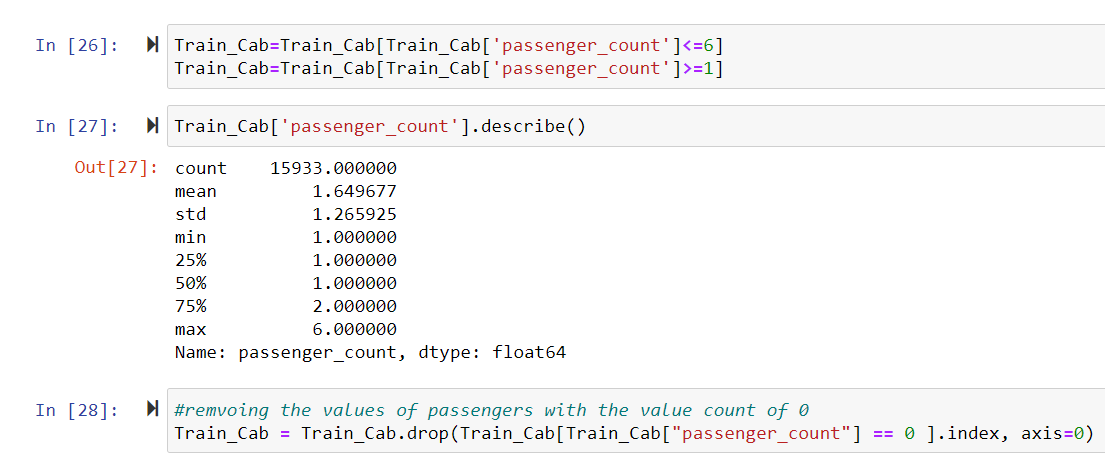
**Fare amount**-During the analysis of fare-amount, it was found out that some values are negative with a min value of $-3 and the max amount is $54343.The cost of the journey cannot be negative so it is found out that it is impure values which are needed to be filtered out from the existing dataset.

Here it is shown in fig.7 given below:



**Figure 7:Setting Up the fare amount**

**Passenger Count**-The passenger count is a major part of assumption and consideration. It was observed that the value for some passenger trips was around 0 at min and 5345 at a maximum which is not possible because the cab has a desire sitting capacity. Here the cab dataset with a capacity of more than 6 is considered as the outlier.



**Figure 8:Setting up the passenger count**

So far, the data has been cleaned by exploring the practical aspects and using multiple approaches, here are some high-level approaches that are used are mentioned below:

1.Get rid of unwanted observations: eliminated the rows in which the fare amount is missing.

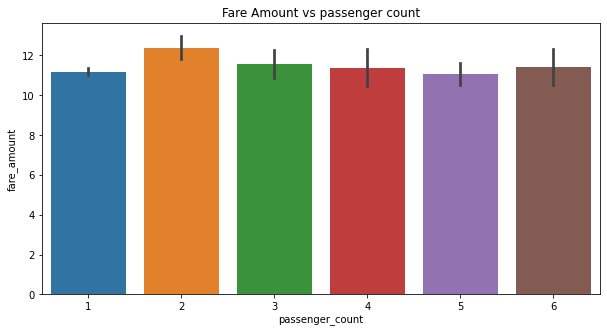
2.Fix structured error:(a). changed the Datatypes, (b). dropping the fare>0

3.standardized the data:(a). changed the values count, (b). No negative values, (c). clearing the longitude and latitude for easy understanding.

The process was done in both the test and train datasets.

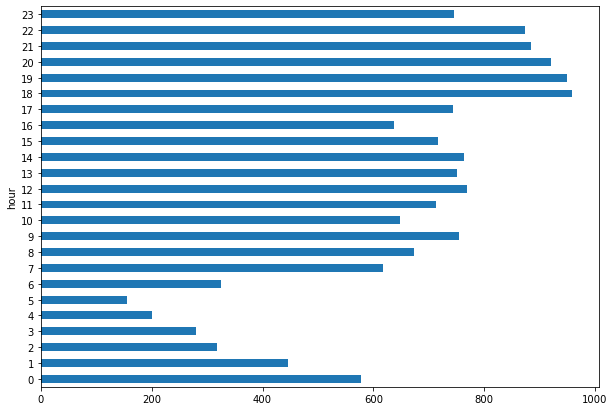
# Visual observations

Visual observations are a very important part of any project because it makes the presentation easy to understand. It helps in understanding the actual behavior and accurate understanding (more in-depth) of the process. Here the visual observations are used in this project to measure any specific behavior that occurs.



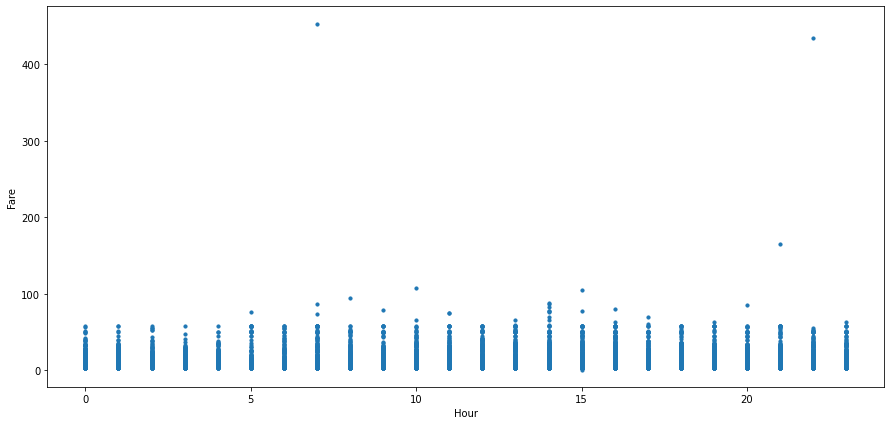
**Figure 9:Fare amount vs Passenger count**

Here Fig.9 shows an observation which stated that 2 is the most common passenger count with the high level of fare amount. Here the fare amount is measured according to the distance traveled. A Cab with the passenger count=2 has traveled the most. Other passenger counts are showing almost the same kind of observations



**Figure 10: Cab Period Time observation in a particular hour**

Above Fig.10 shows the Lowest cabs in the early morning (5:00 am) and the highest cab between office rush hours (18:00 and 19:00)



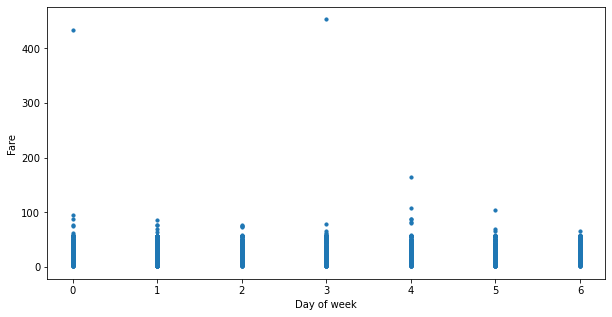
**Figure 11: Time Fare relation Visualization**

With the observation in above fig.11, it can be seen that the cab fare in the early morning at 7:00 am and late night at 23:00 are the costliest



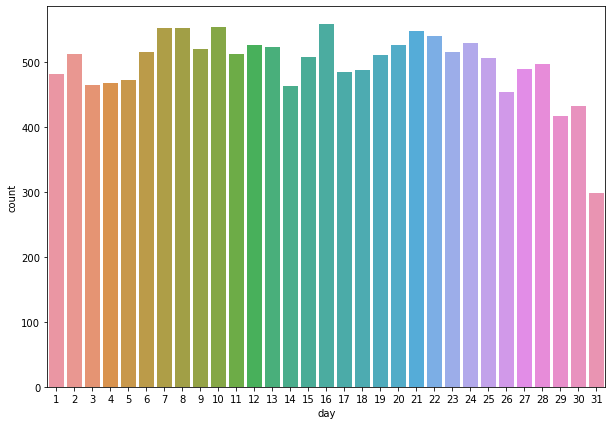
**Figure 12; Visualization of Trip Fare**

Fig.12 is the visualization of trip fare and it is defining a peak value of trip distance and fare amount. This is the test distance which lies between 0-400. It can be observed that most of the fare lies within the range of $100. Some fewer passengers booked the cab for a long distance.



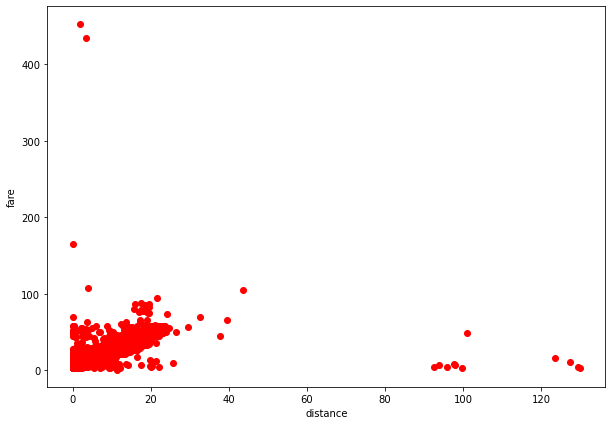
**Figure 13:Fare and days of week visualization**

The observation in Fig.13 can be stated that the Highest fare is on Sunday, Monday, and Thursday. the lowest fare on Wednesday and Saturday. the cab fare is low and the high demand for cabs on Sunday and Monday shows the high fare prices. maybe low demands on the weekend.



**Figure 14: Impact of the day on the number of cab rides**

Fig.14, it can be observed that cab rides are affected at the start and end of the month.



**Figure 15: Relationship between distance and fare**

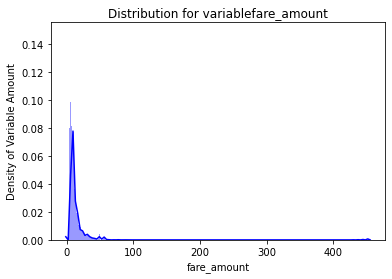
Fig.15 is showing that the cab fare is increasing with the distance which is an obvious statement.

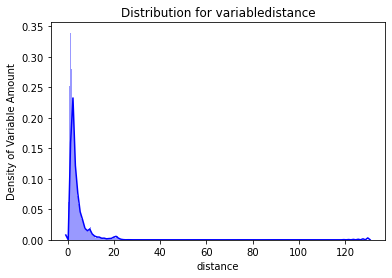
# Feature Scaling

Feature scaling is considered a technique used to standardize the data with independent features present within a fixed range. Feature scaling is performed du ring the data pre-processing which is used to maintain some highly varied magnitudes or units or values.

Skewness is an asymmetry in a statistical distribution where the curve looks like a distorted one or partly skewed on the left or right side. Skewness can be stated as part of the distribution that is different from the normal distribution. Here in this ongoing project, the normalization was checked within fare amount and distance and it was found out that the skewness of the target variable (distance, fare\_amount) is high which is needed to be reduced by performing log transform. Here it is shown in the figure of distribution plot before applying the log transformation.

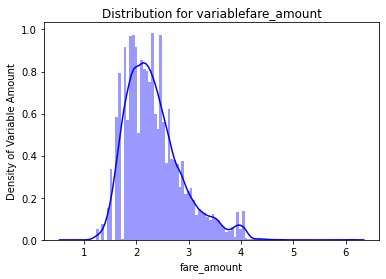
.

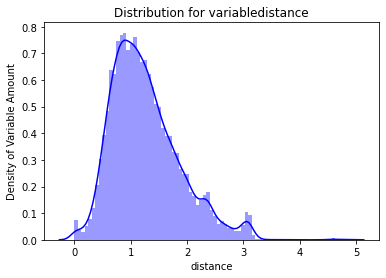




**Figure 16:Skewness in fare amount and distance before applying the log transformation**

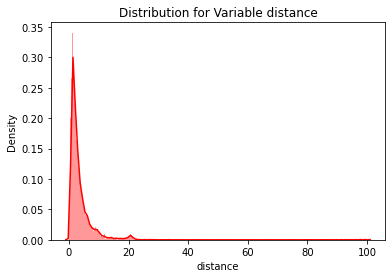
An accurate even number was computed in both fare amount and distance and this step was done because it was showing a lack of precision. Log transformed helped to assist the skewness in distribution. The bell-shaped curved shown in the above visualize area means that the continuous variable is now normally distributed so there is no need of using the feature scaling technique i.e., normalization or standardization



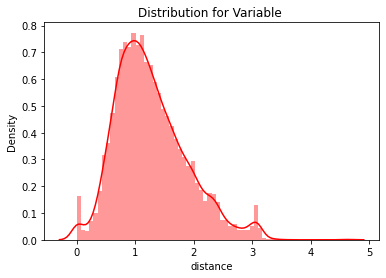


**Figure 17:Skewness after applying the log transformation**

The above process was done in the training dataset and the same process was followed up in the test dataset. Here are some figures showed below with the pre-application and post-application of log transformation to assist the log transformation.



**Figure 18:Skewness of distance in the test set before applying a log transformation**



**Figure 19:Skewness after applying Log Transformation**

# Model Selection

With the completion of data cleaning and data pre-processing, the model selection phase is finally here. In this section, some machine learning algorithms are applied to the test cab. As it has been described that our target variable is fare\_amount which is numeric (the problem is based on prediction and forecasting) so all the models on structure data to predict data cases that are used based on the regression.

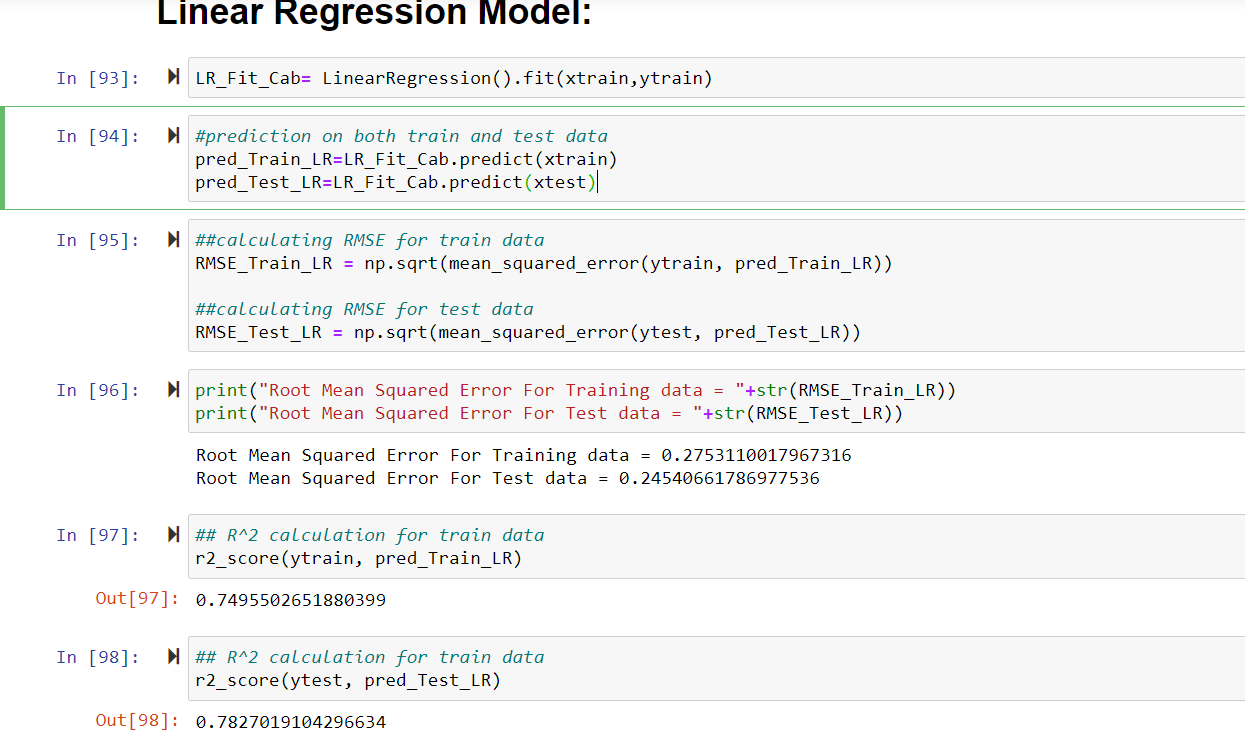
The predicted fare amount is based upon the numeric variable so it comes to know that supervised machine learning algorithms are used to predict the test set case. The reason behind choosing the regression matrix model is that the target variable is continuous.

RMSE (Root Mean Square Error) is used to measure the percentage of error in between two datasets. It is used to compare the predicted value with a pre-observed value. It is an effective error measurement technique because it is directly interpretable and measures proper goodness of fit. In this case, the model that is built should have a lower value of RMSE and a Higher Value of R squared.

## Linear Regression Model

Linear Regression Model is one of the most used statistical methods used for prediction. It is used to find a linear relationship between the target variable and one or more predictors. As it is a part of regression so it is applied when the target variable remains continuous. The main idea behind the LR is to find a line that best fits the data. The algorithm is not so flexible and also highly biased.

Below is the finding of the RMSE and R square using the linear regression model shown below in Fig.20



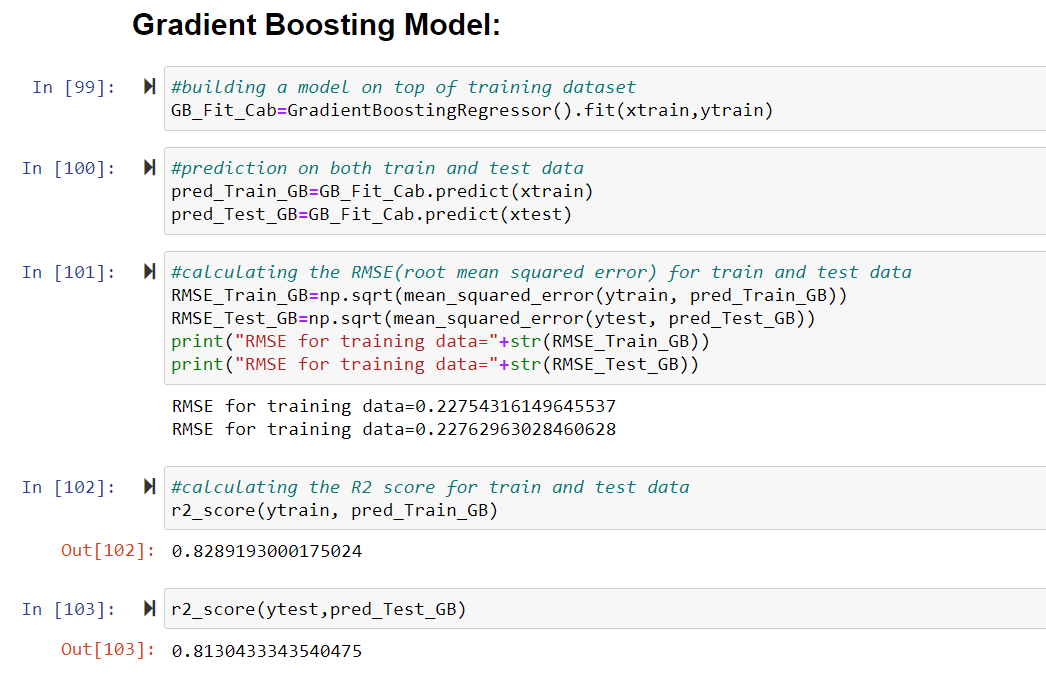
**Figure 20:Findings of RMSE and R-Squared in LR model**

The RMSE of Linear Regression is somewhere acceptable but the R-Squared value is very low so let’s check and compare another model.

## Gradient Boosting Model

It is a sequential ensemble ML technique where the model performance can be improved over iterations. This is widely used to create the model in a stage-wise fashion. An absolute differentiation loss function is enabled which optimizes and infers the model. Most of the time, it is used to get a precise estimation of the response variable.

Below is the finding of the RMSE and R square using the Gradient Boosting model shown below in Fig.21



**Figure 21:Findings of RMSE and R-square in GB model**

The results are better than Linear regression. The RMSE and R-Squared are quite acceptable but the R-Squared score can be found better in some other models. Here the main reason behind the average R-squared score to due to lack of weak learners and revised residuals.

## Decision Tree Model

A Decision Tree is a flowchart-like structure in which each test feature where each leaf node represents a class label that decides to compute all given features. The paths are discovered by a roof to leaf but it is used most in classification. The model works with the target variables and takes a continuous set of values in regression trees.

Below is the finding of the RMSE and R square using the Decision Tree model shown below in Fig.22



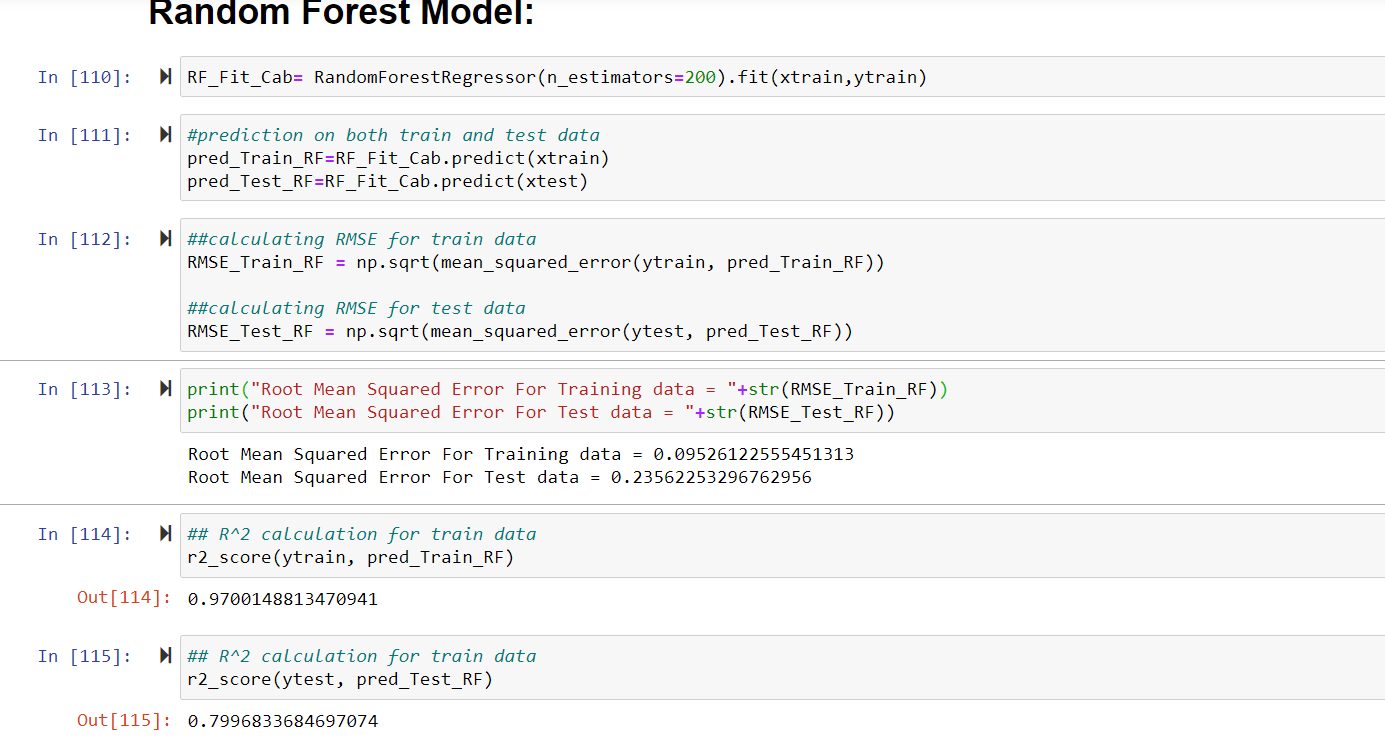
**Figure 22:Findings of RMSE and R-Squared using Decision Tree Model**

The RMSE and R-Squared scores are not acceptable here while other models are showing better results than the decision tree model.

## Random Forest Model

Random Forest contains a collection of Multiple Decision Trees. In Random Forest, the output is quite average with the prediction by each of the other trees. The method works only when the baseline models have a lower bias. The random forest contains n number of trees which is prone to effective accuracy in the dataset. it uses the bagging method for prediction which improves the accuracy of machine learning and it also reduces variance and avoids overfitting.

Below is the finding of the RMSE and R square using the Decision Tree model shown below in Fig.23



The RMSE and R-Squared look much balanced and acceptable. The test model is tuned here with the help of the bagging method without deleting the variable estimation.

The model evaluation shows good results in LR, GB, DT, and RF but these models can be tuned for optimizing the result to balance the weight of variable datasets.

# New Results Optimizations with Parameters Tuning

It is known as the set used for training and controlling implementation aspects of the model. The weight learned during training of the LR model is known as a parameter where the number of trees in a random forest is a model because these are set by the data scientist. These hyperparameters can be thought of as a model set which is needed to be tuned for each problem because the best hyperparameters for one dataset might not be the best across all the datasets.

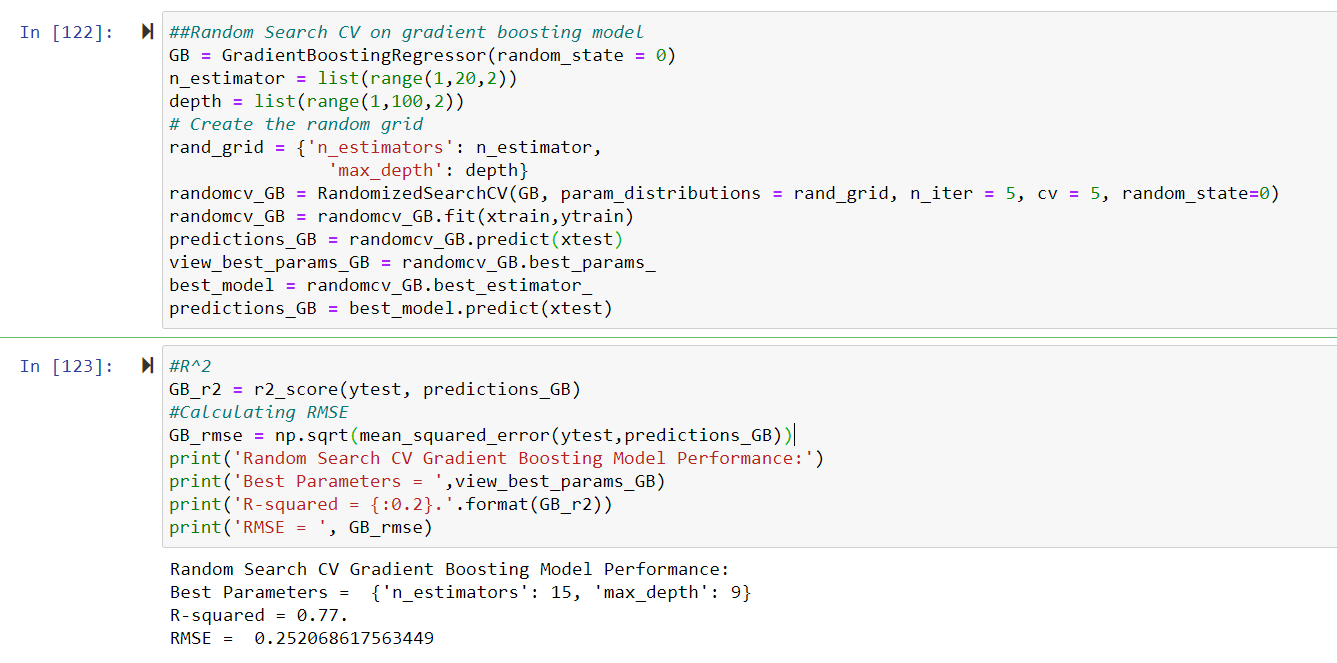
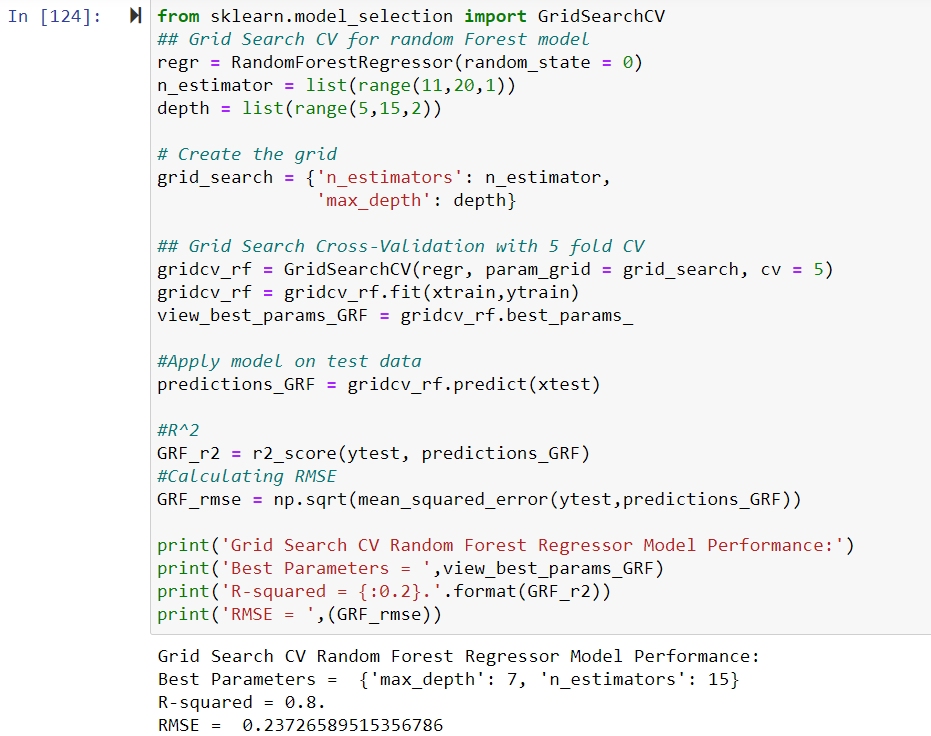
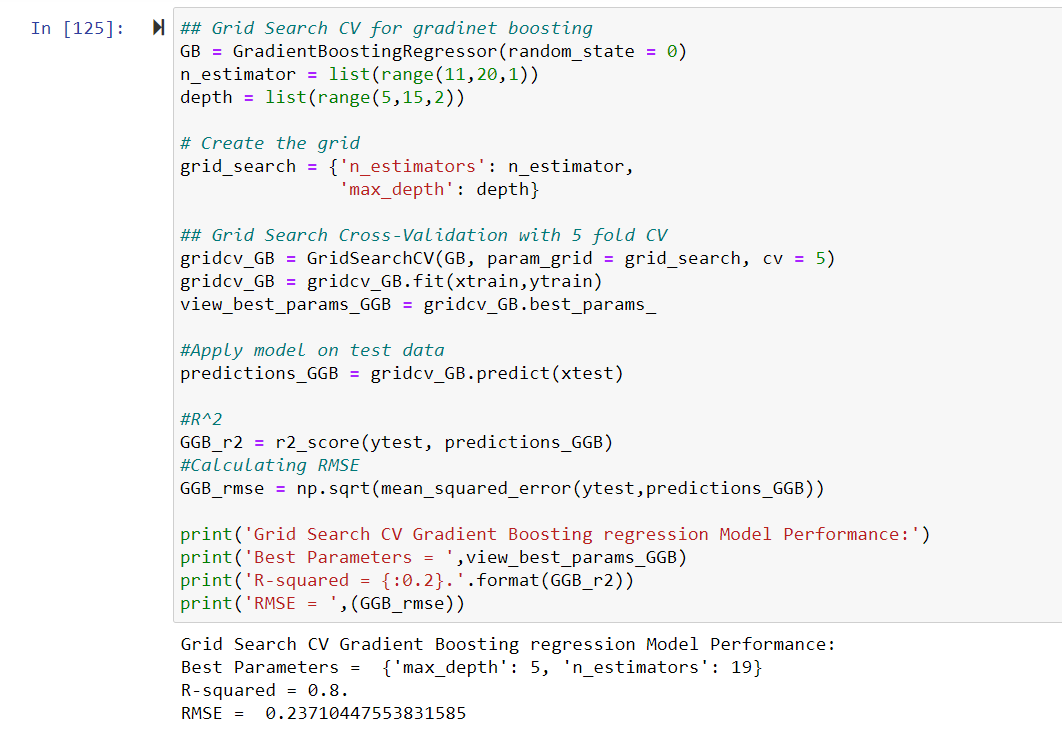
This process is used to find the best combination of hyperparameter values which performs the best with the machine learning model as restrained on a validation dataset.

Two hyperparameters tuning techniques are used in this project:

* Random Search CV
* Grid Search CV

1. **Random Search CV**: It sets hyperparameter grid values and select combinations to train the score as well as model. In this tuning, the search iterations are set based on resources/time.
2. **Grid Search CV**: It works the same as Random Search but in the Grid Search CV, every single combination of hyperparameters values is tried for tuning the model efficiency.

Here are some process images of hyperparameter tuning on each model given below:



# CONCLUSION

In the above sections, the structural format is maintained by applying different steps of the machine learning algorithm and data pre-processing. This section is going to summarize and finalize the performance of the model.

* **Model Evaluation**: All the model that used in above section helped us to calculate the RMSE and R-Squared. RMSE is the standard deviation of predicted errors. Residuals are a measurement of regression line data points distance. On the other side, R-Squared is relatively used as a measure of fit which explains the degree to which the input variable explains the variation of output. In simple words, R-Squared. It is measured between the range of 0-1 where 0 means the independent variable unable to explain the target variables and 1 means the independent variable completely explains the target variable. The high value of R-Squared and lower value of RMSE is a better fit for the model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Name** | **RMSE** | | **R-Squared** | |
| **Train** | **Test** | **Train** | **Test** |
| Linear Regression | 0.27 | 0.24 | 0.74 | 0.78 |
| Gradient Boosting | 0.22 | 0.22 | 0.82 | 0.81 |
| Decision Tree | 0.29 | 0.28 | 0.70 | 0.70 |
| Random Forest Model | 0.09 | 0.23 | 0.97 | 0.79 |

* **Model Selection**: Here is a table of results of model performance after hyperparameter tunning:

|  |  |  |  |
| --- | --- | --- | --- |
| Model Name | Parameter | RMSE(Test) | R-Squared (Test) |
| Random Search CV | Random Forest | 0.23 | 0.8 |
| Gradient Boosting | 0.23 | 0.8 |
| Grid Search CV | Random Forest | 0.25 | 0.77 |
| Gradient Boosting | 0.23 | 0.8 |

Based on RMSE and R-squared, It can be concluded that the Gradient Boosting and Random Forest Model comparatively performed well in comparison to another model. Grid Search was also applied to these chosen models and it was found that Random Forest Model is much effective so finally, it can be stated that the RF model is the best chosen for predicting ongoing projects with a highly explained variance of target variables and with some lower error of chances with GRID SEARCH CV.

Finally, RANDOM FOREST MODEL was applied to predict the Fare amount of Test Data set. Results of predicted fare are attached with the submissions.

THANK YOU

# Appendix

The Python Code file is separately attached with the submission folder and predicted cab fare file has been imported in .csv format which is also attached with overall submissions of the project.