

**FLIGHT PRICE PREDICTION**



Submitted by:

CHAITRA H N

**INTRODUCTION**

**Business Problem Framing**

Flight ticket prices can be something hard to guessandwe know how unexpectedly they vary. The cheapest available ticket on a given flight gets more and less expensive over time. This usually happens as an attempt to maximize revenue based on Time of purchase patterns. The last-minute purchases are expensive. Raising prices on a flight will reduce sales.

Airplane tickets can be ludicrously expensive. Especially when traveling shorter distances, the difference in cost for air travel and other modes of transport, would seem comically large. And the time saving aspect decreases for short distances when you consider the time it takes to cut through all the hassle and red tape at the airport.

Flight booking systems are dynamic in nature. They depend on a lot of features like Airline company,Source, Destination, duration, arrival time, departure time, number of stops and date of the flight. Inthis project, I plan to use machine learning algorithms on a dataset based on the above parameters topredict flight prices. There are basically two approaches to solve this problem. These involve considering it as a regressionor classification problem. Algorithms can be applied to predict whether the price of the ticket will dropin the future, thus considering it as a classification problem. In this project, I will consider it as aregression problem, thus predicting the ticket price.

**Conceptual Background of the Domain Problem**

Airline Industry is one of the most sophisticated industries in its use of dynamic pricing strategies tomaximize its revenue. The model used by airline companies for the prediction of a ticket price is notpublic and it is really very complicated because it is based on proprietary algorithms and hiddenvariables. There is a need to develop a model for the consumers from which they can predict theseprices and moreover analyze which feature is the most influential in determining these prices. In thisproject I have applied Machine learning algorithms on a dataset consisting of various factors which caninfluence the flight fares to prepare such a model.

**Review of Literature**

In the past of few decades ago, airline industry is being control very tightly with lots of regulations. For example, United State (US) air transportation industry is being control tightly by the Civil Aeronautics Board (CAB) on price, route and schedule of flight. After that in 1978, domestic air transportation market of US is having a free competition among airlines which allowed by the Airline Deregulation Acts of 1978. Through this act, every airline is allowed to set their own price, how frequent they are flying and the destination they want to fly to (Thomas, O.G., 2004).

After the deregulation of American airline market, European air transport also experienced deregulation in the middle of 1980. The result of regulation is an increase competition of airlines and to open new entry to new airlines. The airline structure has changed due to deregulations. The airline industry becomes more competitive with numbers of competition. Changes of pricing strategy, marketing strategy and airlines networks such as hubs and spokes had made.

**Motivation for the Problem Undertaken**

Air travel is the fastest method of transport around, and can cut hours or days off of a trip.But we know how unexpectedly the prices vary. So, I was interested in Flight Fares Prediction listings to help individuals and find the right fares based on their needs. And also, to get hands on experience and to know that how the data scientist approaches and work in an industry end to end.

**Analytical Problem Framing**

**Mathematical/ Analytical Modelling of the Problem**

In this project, I have collected the details of Flights from different websites like Yatra.com, Easy my trip.com, Via.com so for collecting the data. I have used web-scraping, (selenium) for scraping the details. Then I have collected all the data from different locations in India and collected the details like Air-linesname, Source, Destination, duration, arrival time, departure time, number of stops, Price and date of the flight. So, by using all the information of the data collected we need to predict the flight prices. After collecting the data, I have put all together in a data frame and saved the data as excel file.

Then by using my Jupiter notebook I have imported required libraries like pandas, numpy, matplotlib, seaborn where we use them for our problem using then I imported my collected excel file data and checked for top 5 rows using head method and then I have checked for the shape of the data. From shape method I got to know that there are 3435entries and 9 columns. Then I checked for is null () method to find the NaN’s then there are no null values present in the columns and then checked for d-types of the data in the columns using info method. Then I checked for the describe method and then plotted some graphs and visualized and tried removing outliers from the data and then scaled the data. As the label is a continuous variable, I have used regression models for predicting our label. I have used various algorithms for drawing the patterns and concluded a final model on the basis of performance and evaluation Metrics.

**Data Sources and their formats**

By using pandas, I have first imported the Excel file and it consists of different columns which includes data in it. Our dataset consists of Features and label. After importing I have checked for shape of the dataset and which consists of rows and columns. Then I checked for null values and need to be treated and then I checked for info () method for knowing the type of the data then I checked for stats using describe method.

Our label is price prediction which is a continuous variable based on the values of independent variables our dependent variable depends.



The dataset used in the project are scrapped from different flight websites and collected the data.

It consists of 9 attributes with over 3435 entries. The features of the dataset are as follows:

1.airline\_name: The name of the airlines

2.date\_of\_Journey: The date of the journey.

3.Source: Source city of the flight.

4.Destination: Destination city of the flight.

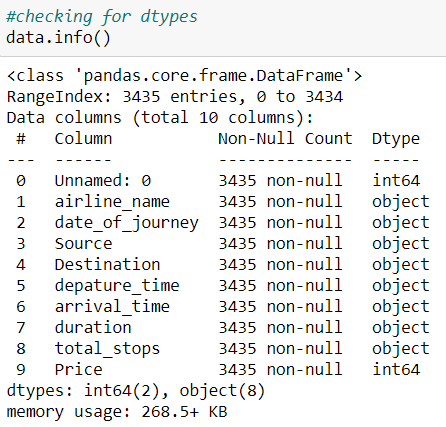
5.Depature\_time: Time of the departure of flight.

6.Arrival\_Time: Time of arrival of the flight.

7.Total\_stops: No of stops

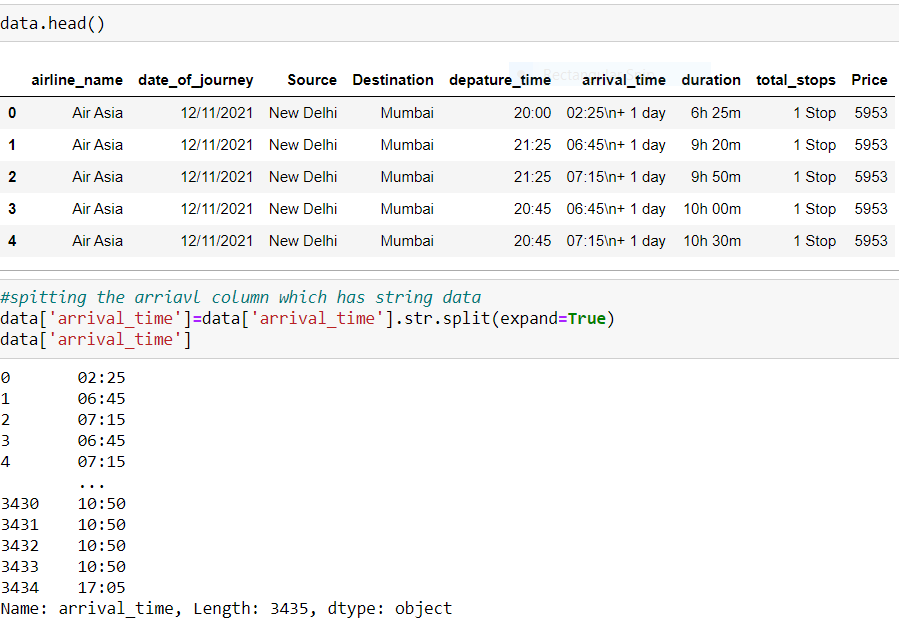
8.Price: Fares of different flights

9.Duration: The total time a person sits in the flight.



**Data Pre-processing Done**

There are no null values in the dataset. The arrival time column has some extra characters added in the time those are to be cleaned so I have used str.split() method to split the time and the unwanted data and cleaned the arrival time column and converted all the time and duration columns which are having special characters and letters using date time library in pandas as shown below.



**Feature Engineering**

I have replaced the column classes which are having same name with change in the format as shown below.

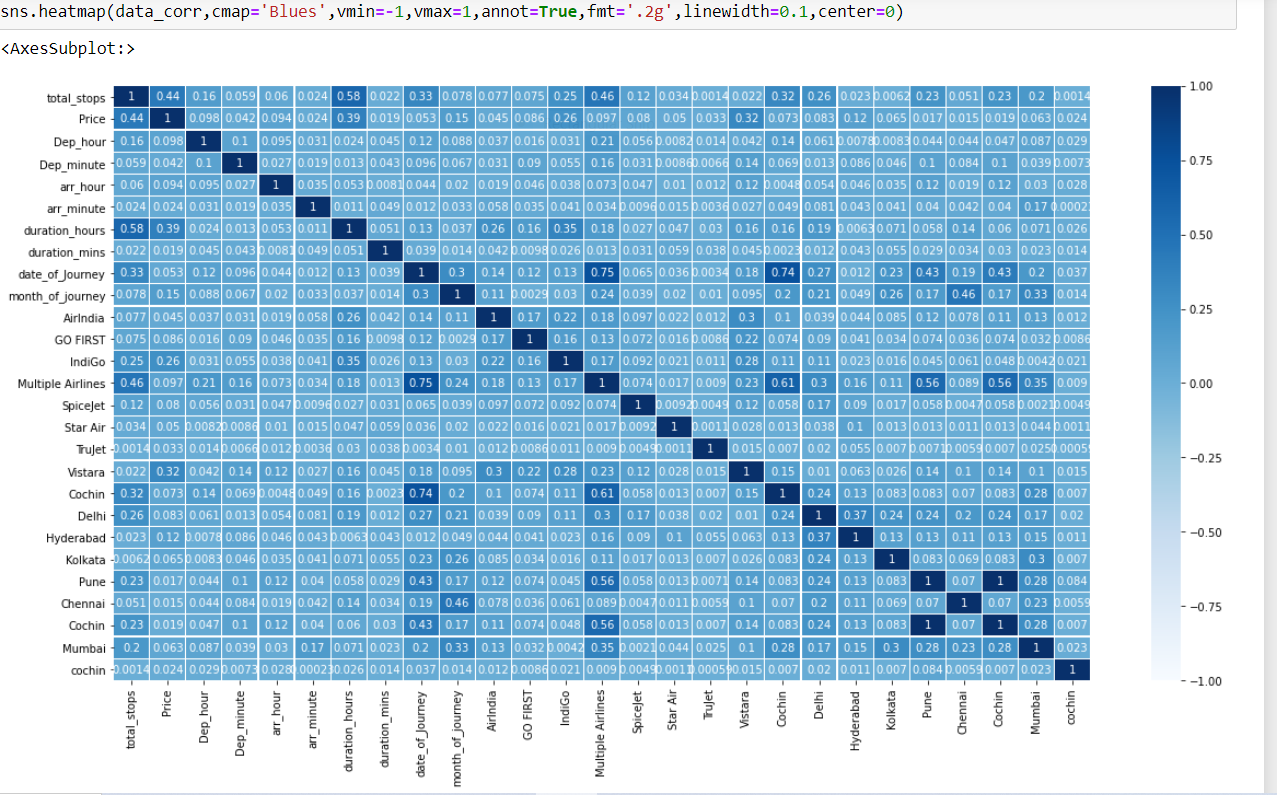




**Data Inputs- Logic- Output Relationships**

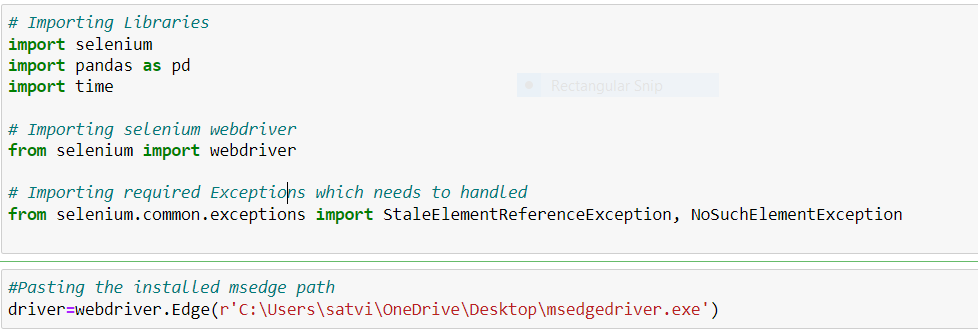
For checking the relation between the columns, I have used correlation matrix to find the relation and plotted heat map to visualise the percentage of the correlation. The below are the observations from the heatmap.

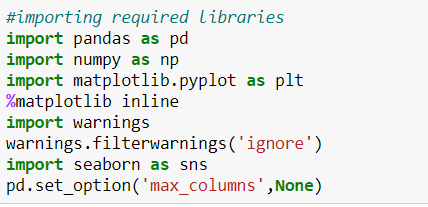
1.The Dark blue indicates high correlation and light blue indicates less correlation.

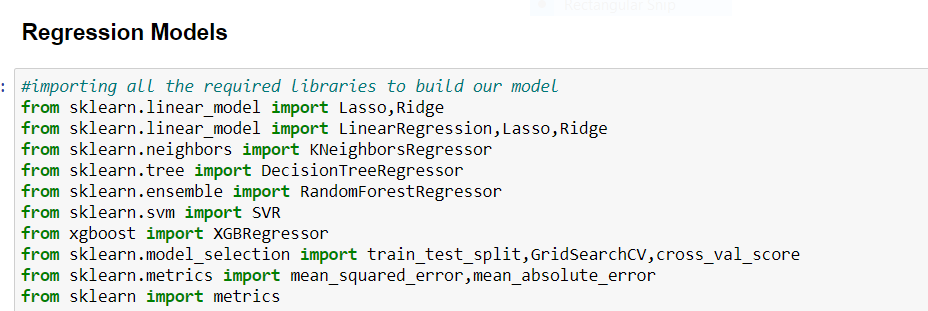
2. Even there are some columns which are highly correlated with each other which means there exists multi collinearity problem with Pune and cochin columns.

**Hardware and Software Requirements and Tools Used**

I have used my laptop, wed server, micro-soft edge, Jupiter Notebook which is having GUI interface. Imported necessary libraries from python such as pandas, NumPy, seaborn, matplotlib, then imported the required model libraries from Scikit learn to import our algorithms.



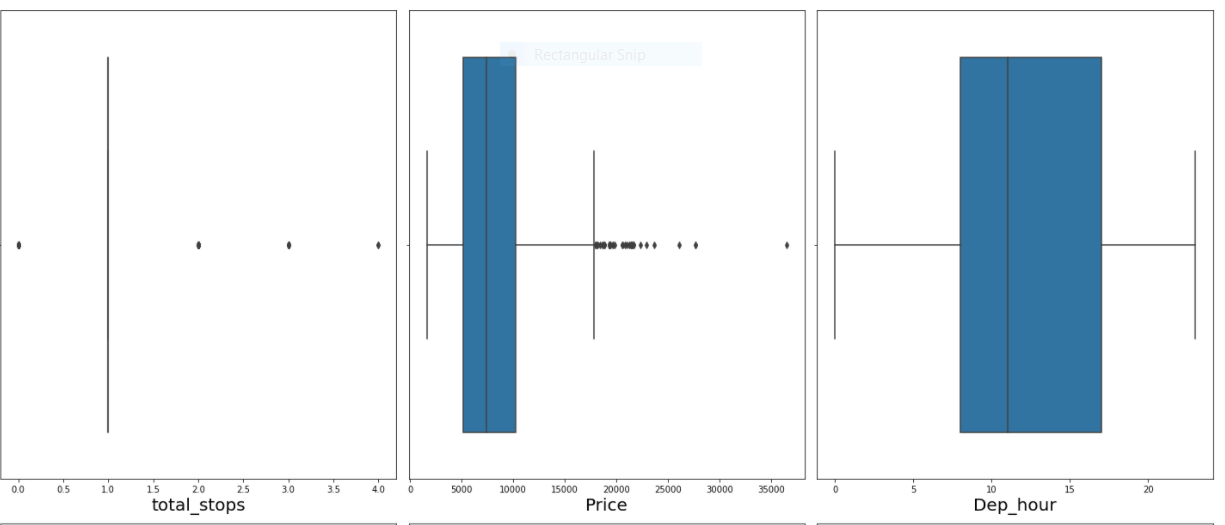




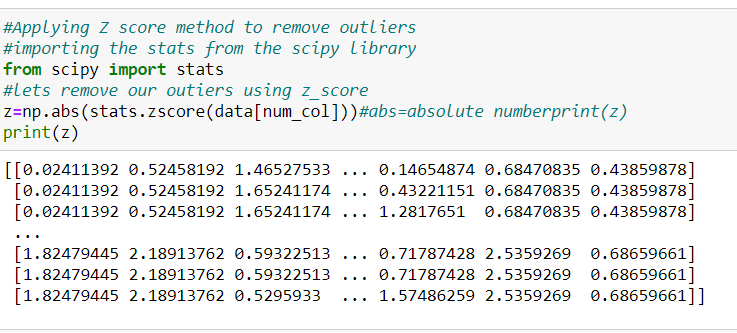
**Model/s Development and Evaluation**

I have plotted box plots to check for outliers and distribution plots to check the skewness so I found the presence of outliers and skewness in the continuous data. So, I have used Z-Score method for removing the outliers and log transformation method the remove the skewness.

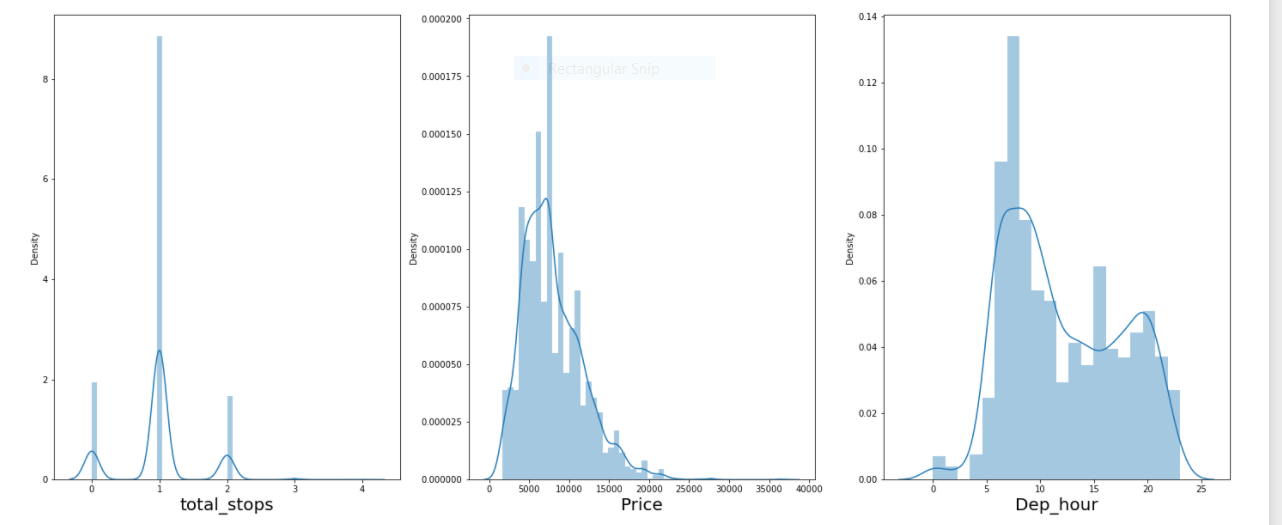
**Box Plots**:



**Z-Score**



**Distribution Plots** are used to check the flow of the data in the columns.

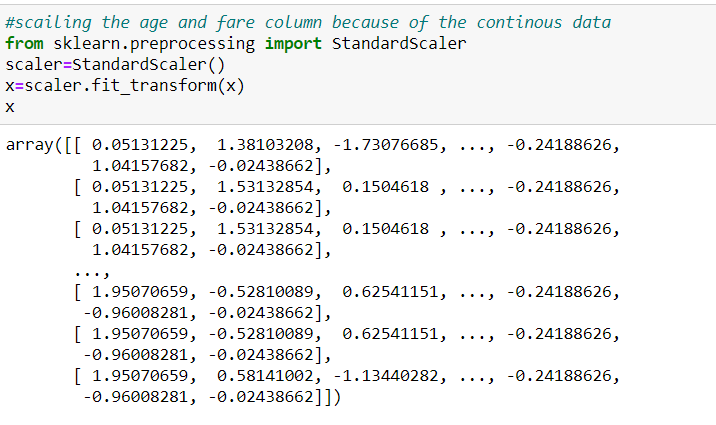


**Testing of Identified Approaches (Algorithms)**

The price of the Flights prediction is a numerical variable so it comes under regression problem, So I have used 6 different algorithms to check the model patterns. In order to final a model we have checked on different evaluation metrics like finding the score of the training data and testing data and finding the errors like mean absolute error (MAE), mean squared error (MSE) and Root mean squared error (RMSE). In order to tell a model is good their RMSE value should be as less as possible then we can say the model is efficiently working on the given data.

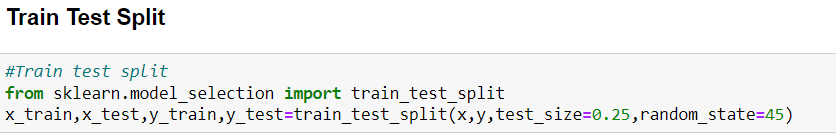
**Scaling the data**

I have used standard Scaler to scale the data**.**



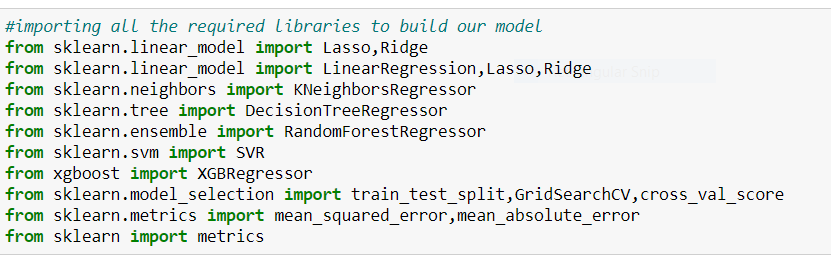
**Train Test Split**

I have imported the train\_test\_split from the module sklearn from model\_selection. And used 75% of the data for training and 25% of the data for testing and splitted the data into x-train, x-test, y-train, y-test.



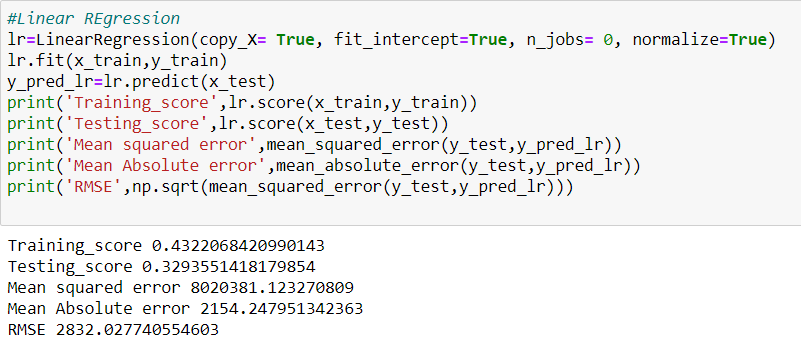
**Regression Algorithms used for our prediction**

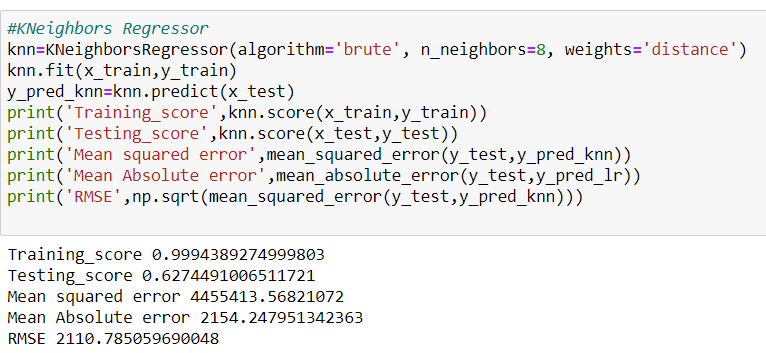
Since our price prediction is a continuous variable then this comes under Regression problem so I have used different Regression algorithms for predicting our label.

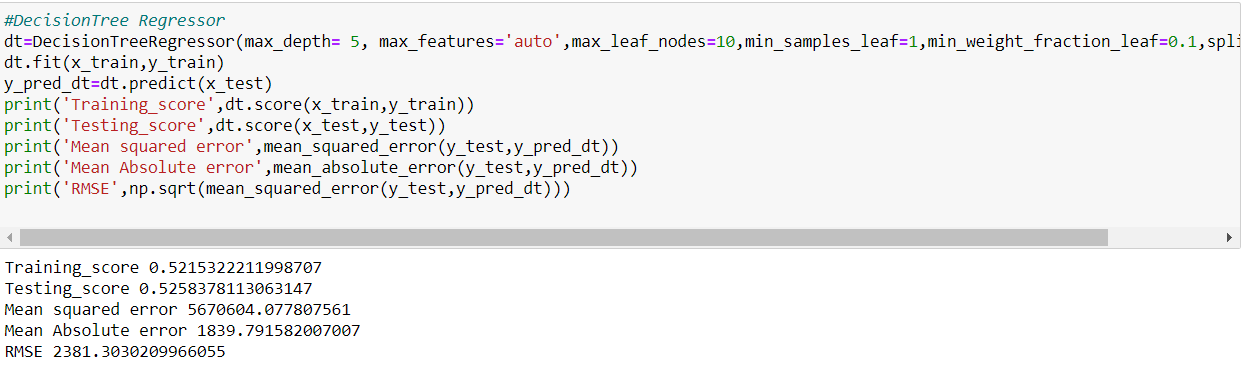


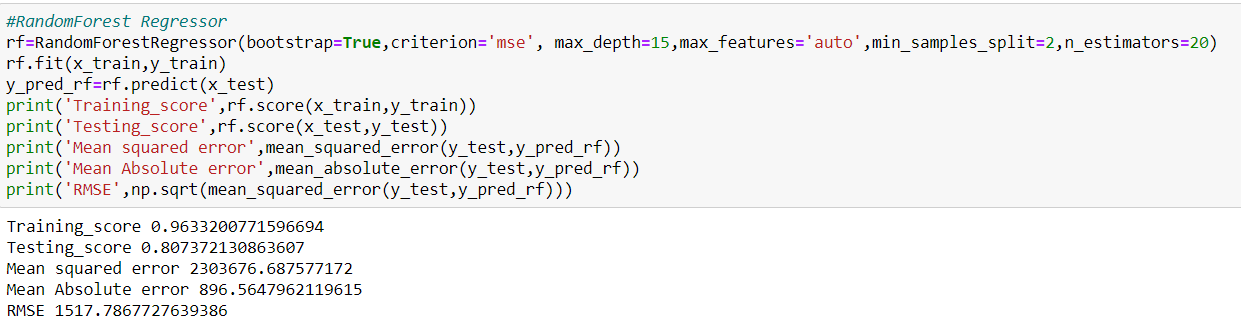
**Run and Evaluate selected models**

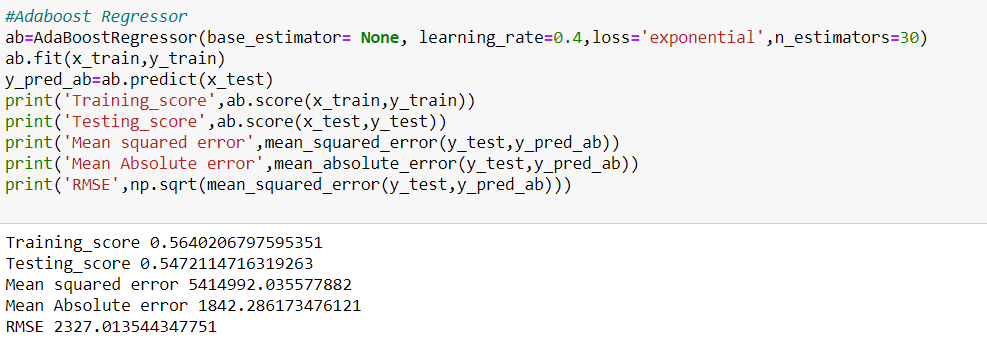
I have used various algorithms for predicting our label like Linear Regression, KNeighbors Regression, Decision Tree Regression, Random Forest Regression,Support Vector Regression, Ada Boost Regression, XGBoost Regressor. For evaluating the model, I have used Mean Squared Error(MSE), Mean Absolue Error (MAE), training score, testing score and root mean squared root(RMSE).

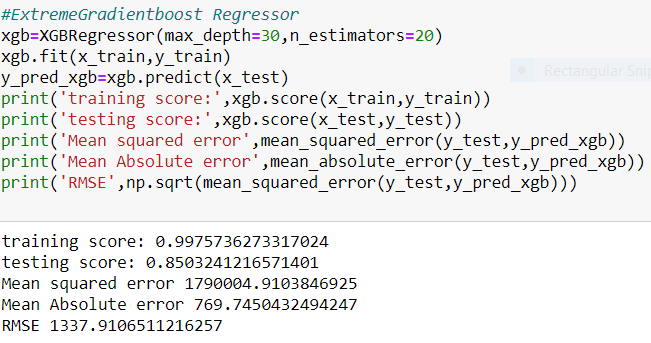






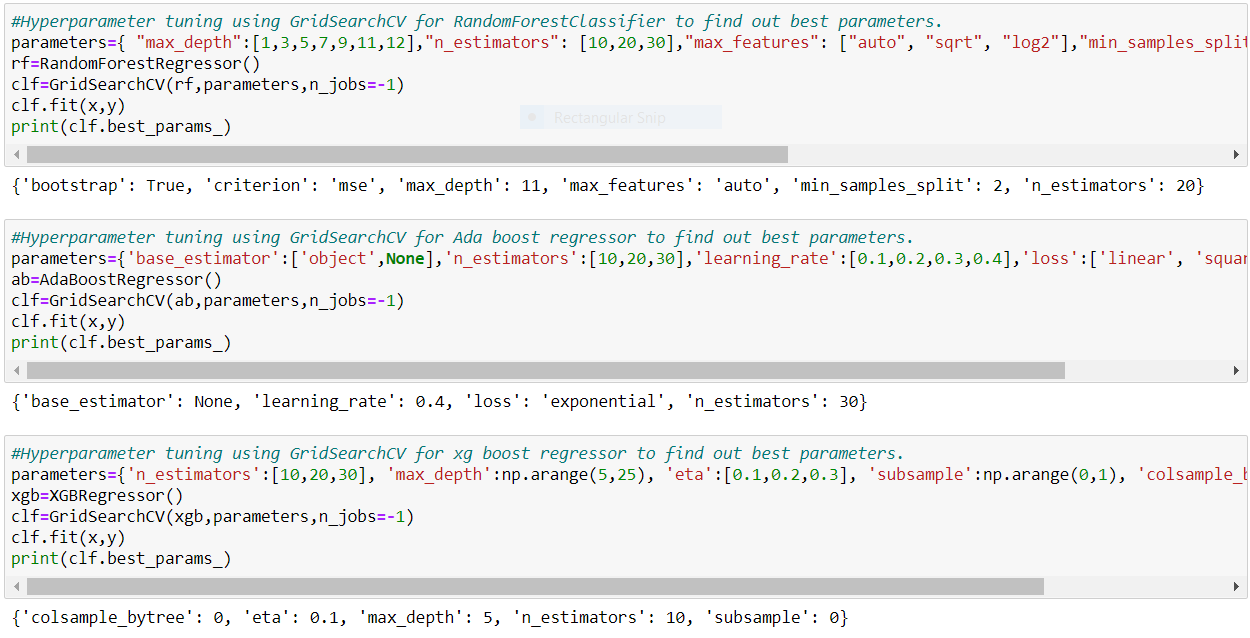






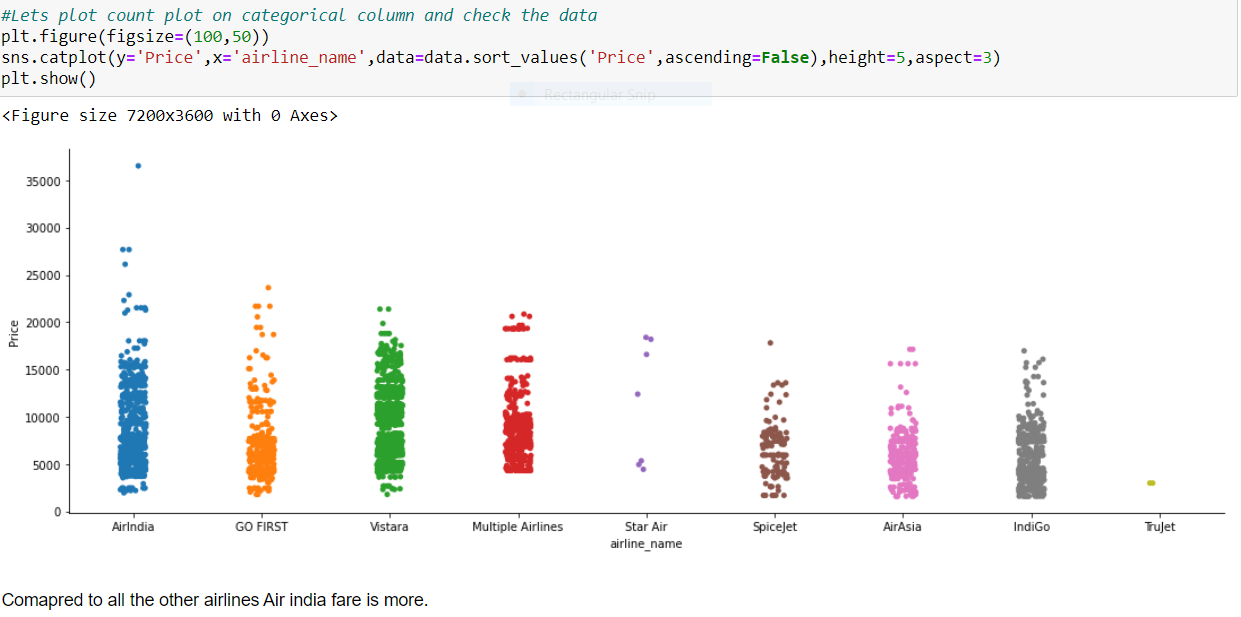
**Key Metrics for success in solving problem under consideration**

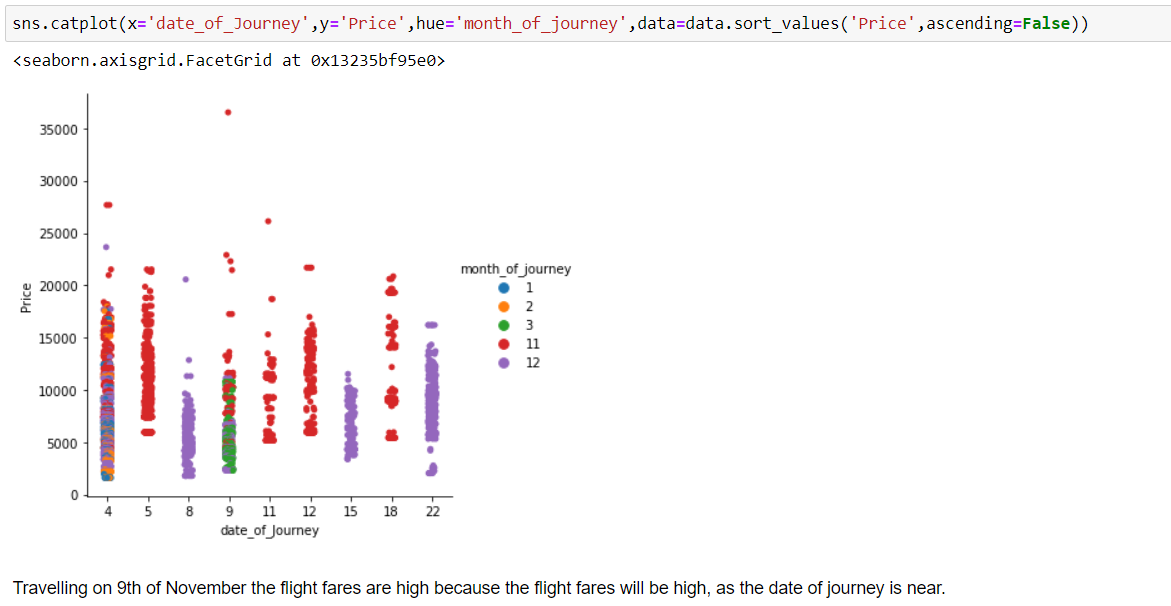
After Checking the model evaluation, I have performed hyperparameter tuning to improve the score of the model. By using Grid search cv we are going to pass different parameters for the algorithms which improves the score of the model and reduction in errors. After applying grid search cv there is change in the score and error.

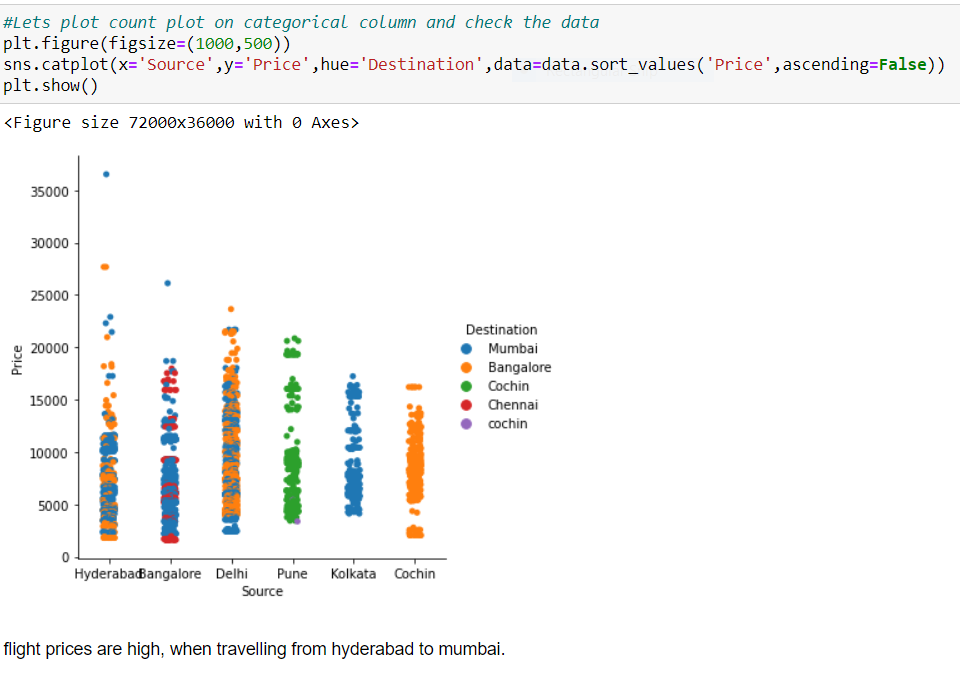


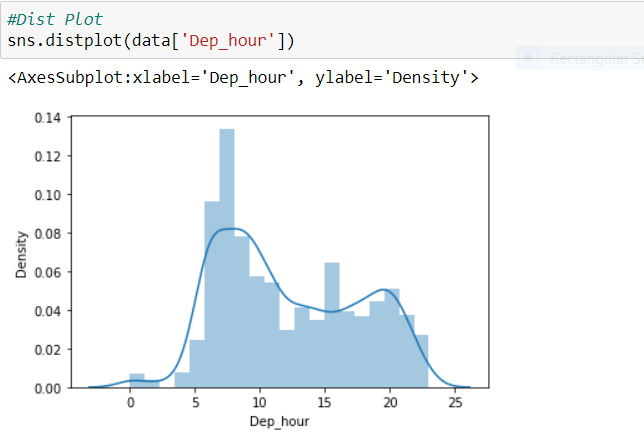
**Visualizations**

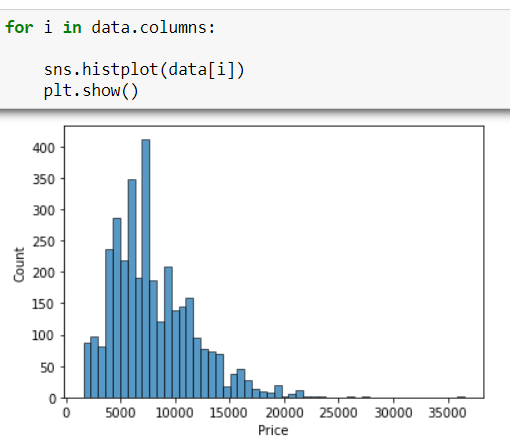
I have compared by plottingcat plots on differentfeatures with the labeland compared the variation of fares with different classesin the features and tried to know how label is varying with the features and even plotted hist plots for all thecolumnsand checked the variation of the val



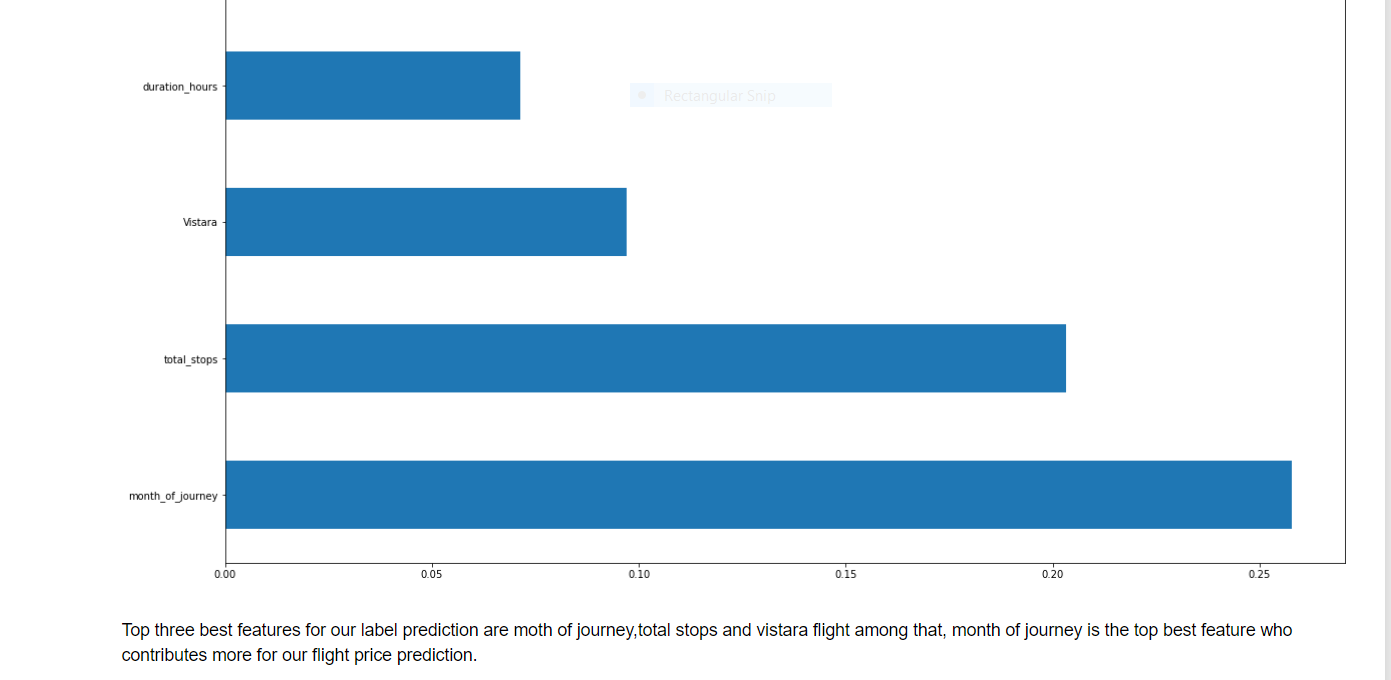






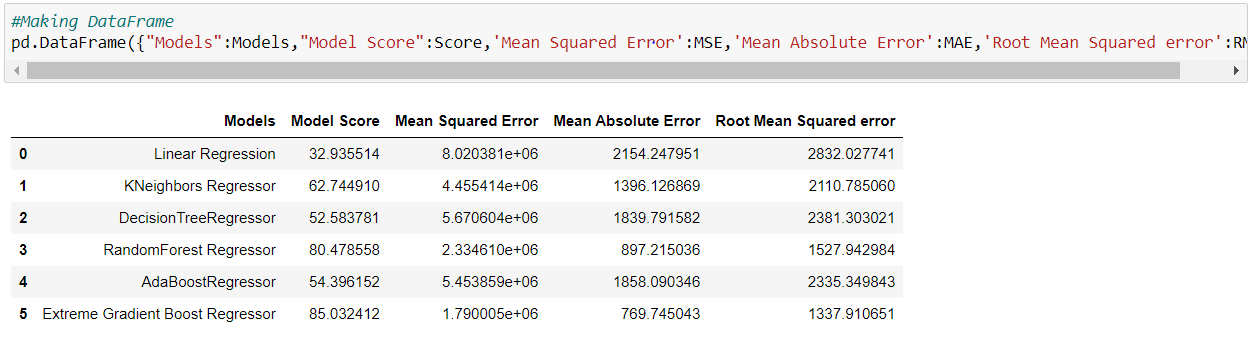


Extra Tree Regressor:



**Interpretation of the Results**

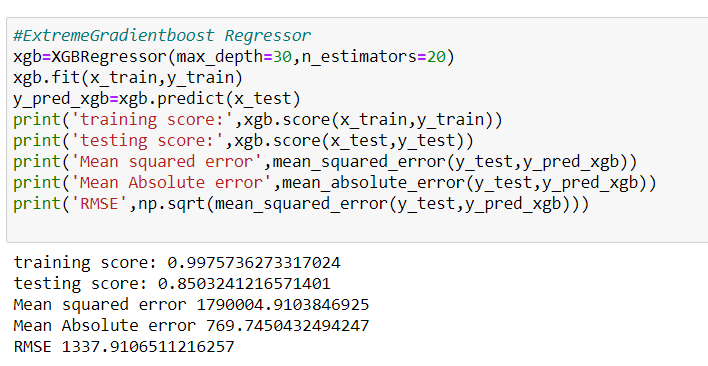
I have created a data frame using all the models used for prediction, their scores and errors given by the models as shown in the below screenshot.

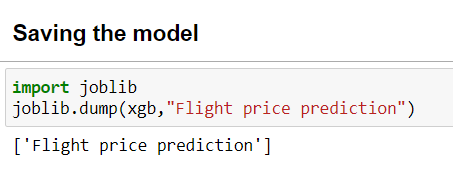


**CONCLUSIONS**

**Key Findings and Conclusions of the Study**

I have used various models for predicting the price of flights and used various evaluation metrics for evaluating the model like finding the training score, testing score, Mean Squared error (MSE), Mean Absolute error (MAE), Root Mean squared error (RMSE). So, after evaluating on different models, Extra Gradient boost Forest giving high score and low RMSE Value. So I finalised the model and saved the model using job-lib library.



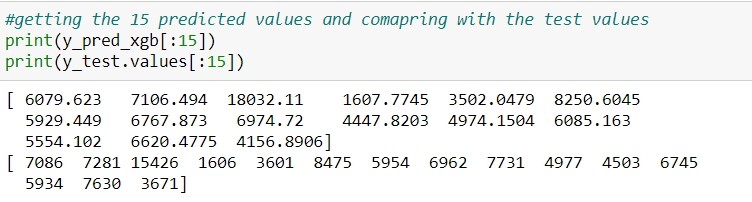


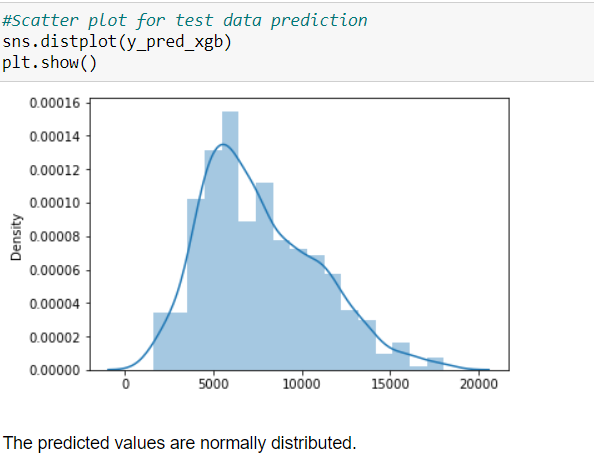
**Conclusions on our model building**

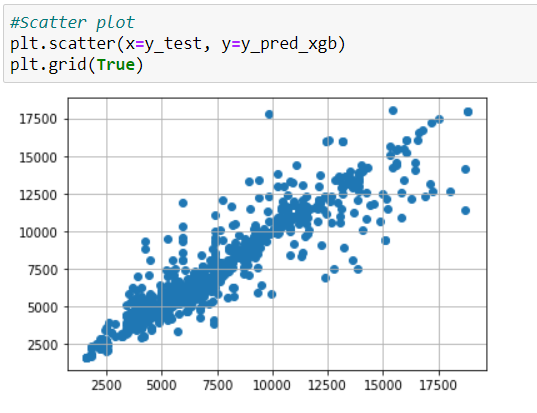
We got our best model as XGBoost Regressor with the r2 score of 85% and the RMSE value is also less compared to all other models. So, we can go further build our model using XGBoost Regressor.

**Learning Outcomes of the Study in respect of Data Science**

After finalising the model XGBoost Regressor, I have taken the values of prices which are predicted by the model and compared with the actual Price values and checked the relation between them by plotting the scatter plot and plotted distribution plot for predicted value and checked the data distribution and visualised the predicted price values are normally distributed.







**Limitations of this work and Scope for Future Work**

This study shows that it is feasible to predict the airline ticket price based on historical data. One possible way to increase the accuracy can be combining different models after carefully studying their own performance on each individual bin. Additionally, as the learning curve indicates, adding more features will increase the accuracy of our models. However, limited by the current data source that we have, we are unable to extract more information of a particular flight. In the future, more features, such as the available seat and whether the departure day is a holiday or not, can be added to the model to improve the performance of the predicting model.

The Root Mean squared error (RMSE) errors calculated for all the algorithms are very high. Statistical methods work better, on large set data. But the length of the dataset is very less so using different methods that match the time-series data will be used in the future research to obtain smaller error prediction values (RMSE) and using more data to get the better result. In future this machine learning model may bind with various website which can provide real time data for price prediction. In Future, we need to add extra historical data of flight price which can help to reduce the RMSE error.