Flight Price Prediction



Submitted by:

Mini Sharma

**About the Fight Price**

When we generally book a flight to go somewhere generally the price of the flight fluctuates but Do we know the possible reasons behind these differences? There are many factors which affect these variations. These variations may be because of the following reasons:

* The name of the airline.
* The date of the journey
* The source from which the service begins.
* The destination where the service ends.
* The route taken by the flight to reach the destination.
* The time when the journey starts from the source.
* Time of arrival at the destination.
* Total duration of the flight.
* Total stops between the source and destination.
* Additional information about the flight
* The price of the ticket

Here by help of this project we have tried to understand these variations and factors leading these variations. Later we wil predict the flight price depending upon the circumstances.

# INTRODUCTION

## Business Problem Framing

## It is a project related to airlines and its industries. We are basically predicting the price of the flights

## Based on the conditions of the customers.

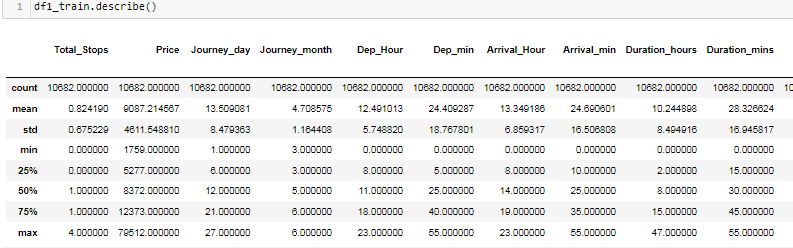
## Review of Literature

In this model we will study different variables and how this independent variables are related with dependent variables and how this will help us to predict whether the price of the flights.

**Analytical Problem Framing**

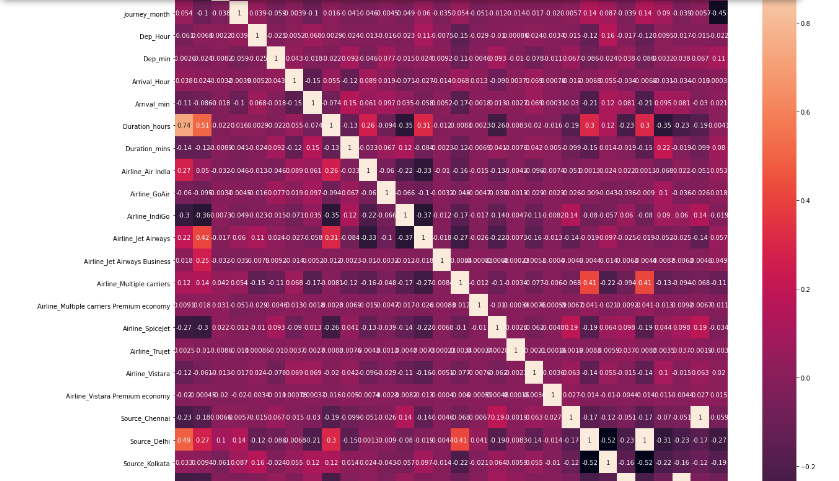
## Mathematical/ Analytical Modeling of the Problem

Lets view some basic statistics about the data like the percentile, mean , maximum, minimum etc.



* + The average value for flight price is 9087.214.
  + The average duration of the flight is 10.244 hours and 28.32 mins.
  + The average number of stops are 0.82.

#### Lets See co-relation between the Columns

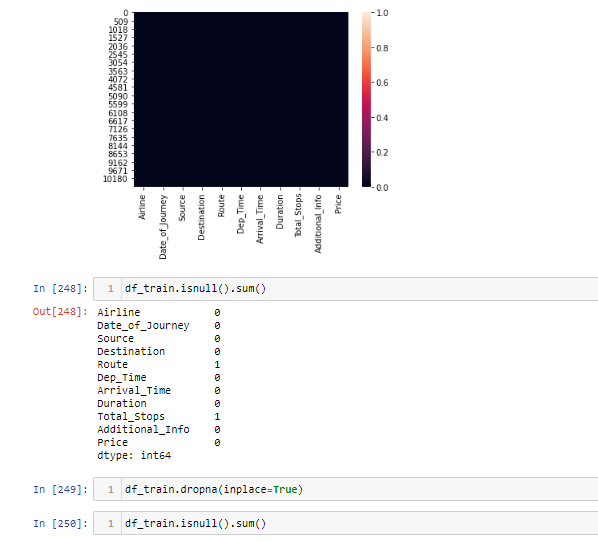


## Data Sources and their formats

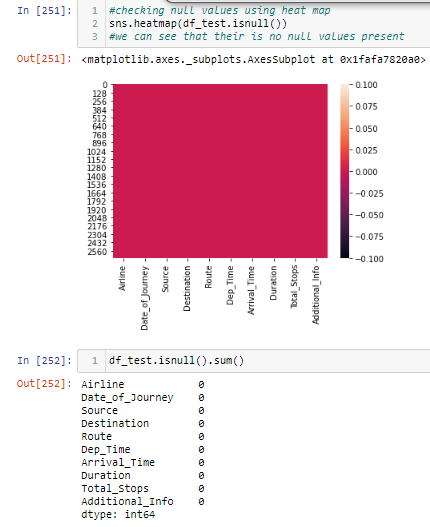
We got this data in excel format and worked on it accordingly.

## Data Preprocessing Done

* + Lets check the shape and see count of the number of empty values in each column.

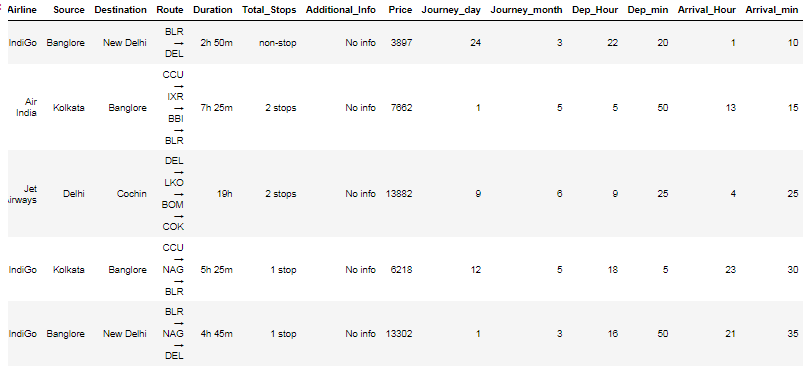


* + As we can see from above train Dataset contains 10683 rows and 11 columns in which label is the dependent target column and rest are independent columns .
  + And we can see dataset contains null values. So we dropped the null values.
  + Our test dataset doesn’t have any null value so far.

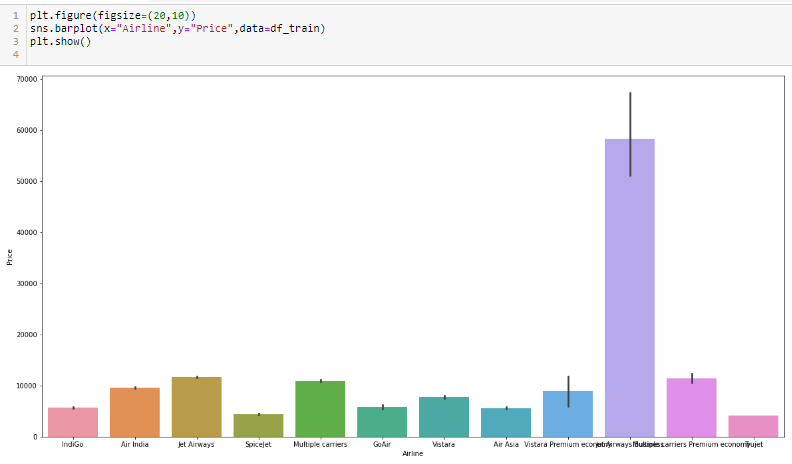


* + Data set contains journey date in format of year ,month and date. We will split the pdate column for further analysis.
  + After checking the unique values of each column we can see that year count is only one soo we will drop journey date , year and unnamed as it is of no use.





## Data Inputs- Logic- Output Relationships

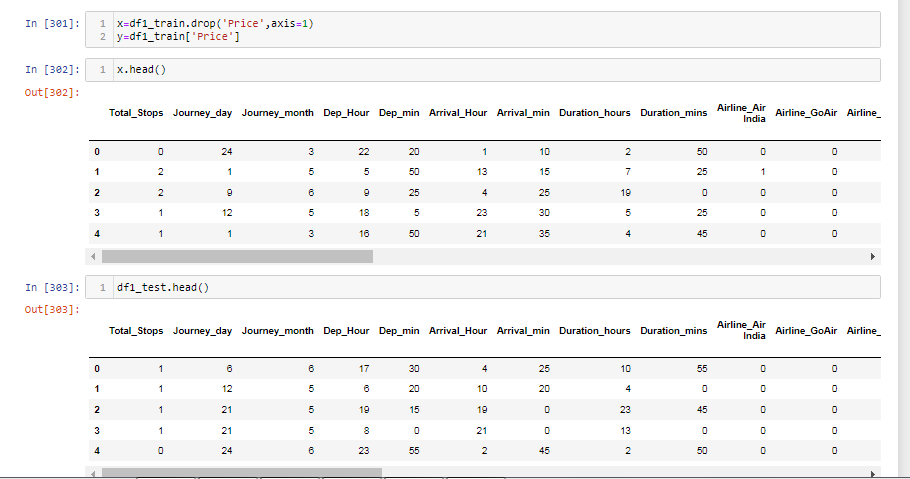


## Hardware and Software Requirements and Tools Used

We will use here Jupyter notebook to make your Prediction Model.

# Model/s Development and Evaluation

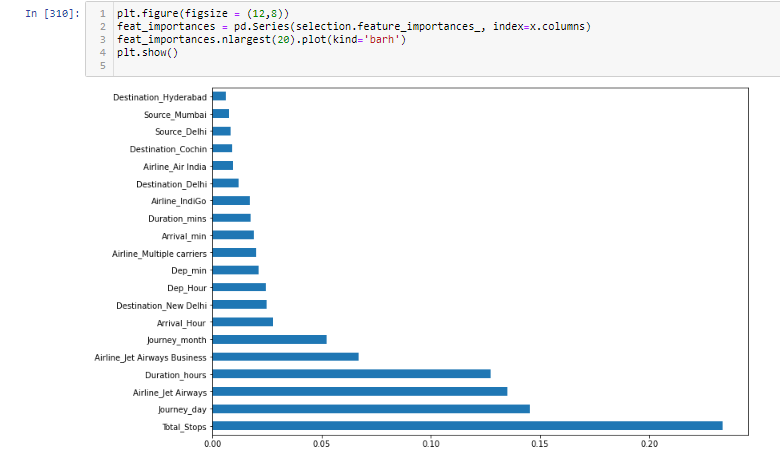
## Identification of possible problem-solving approaches (methods)



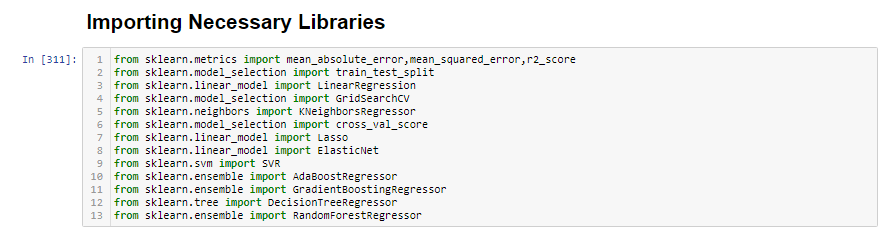
* + Now we will split the data set into input and output variable. As you can see above x is your input variable and y (label) is your target out variable.

**Lets check feature importance of the Data set.**

* + You can get the feature importance of each feature of your dataset by using the feature importance property of the model.
  + Feature importance gives you a score for each feature of your data, the higher the score more important or relevant is the feature towards your output variable.
  + Feature importance is an inbuilt class that comes with Tree Based Classifiers, we will be using Extra Tree Classifier for extracting the top 10 features for the dataset

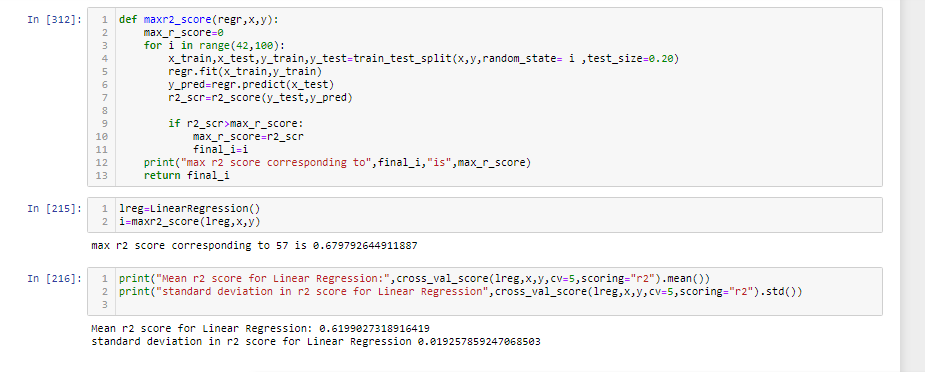


## Testing of Identified Approaches (Algorithms)



1. **Linear Regression**

* In Logistic Regression, we wish to model a dependent variable(y) in terms of one or more independent variables(x). It is a method for regression. This algorithm is used for the dependent variable that is Continous. Y is modeled using a function that gives output in linear relationship.



Further we will use grid search cv to find the best parameter linear regression.

Results

#### Decision Tree Classification

The idea of a decision tree is to divide the data set into smaller data sets based on the descriptive features until you reach a small enough set that contains data points that fall under one label.

Decision trees are easy to interpret. To build a decision tree requires little data preparation from the user- there is no need to normalize data.



Further we will use grid search cv to find the best parameter for Decision Tree Classifier.

#### Random Forest Classification

Random Forest is a supervised learning algorithm, it creates a forest and makes it

somehow random. The "forest“ it builds, is an ensemble of Decision Trees. Step-1Pick at random K data points from the training set.

Step-2 Build the Decision tree associated to these K data points

Step-3Choose the Number of trees(n) you want to build and repeat Step1 and Step2

Step-4For a new data points make each one of your 'n' trees predict the category to which the data point belongs and assign the new data point to the category that wins the majority vote.



Further we will use grid search cv to find the best parameter for Random Forest.

Results:

#### Gradient Boosting-

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

Further we will use grid search cv to find the best parameter for Gradient Boosting Classifier.



#### K Neighbors Classifier-

This is a supervised, non-parametric learning algorithm which classifies a given point based on its neighbours. The choice of the ‘k’ becomes very crucial since the data point is assigned to the class of the nearest ‘k’ neighbors. Once we get to know such ‘k’ nearest data points, the test data is assigned a label by taking the

majority vote from the class labels of the ‘k’ nearest data points



Further we will use grid search cv to find the best parameter for K Neighbors Classifier.

## Key Metrics for success in solving problem under consideration

**Accuracy Score** is the number of correct predictions made as a ratio of all predictions made. It is the most common evaluation metric for classification problems.

**Cross-validation** is to call the cross\_val\_score helper function on the estimator and the dataset.

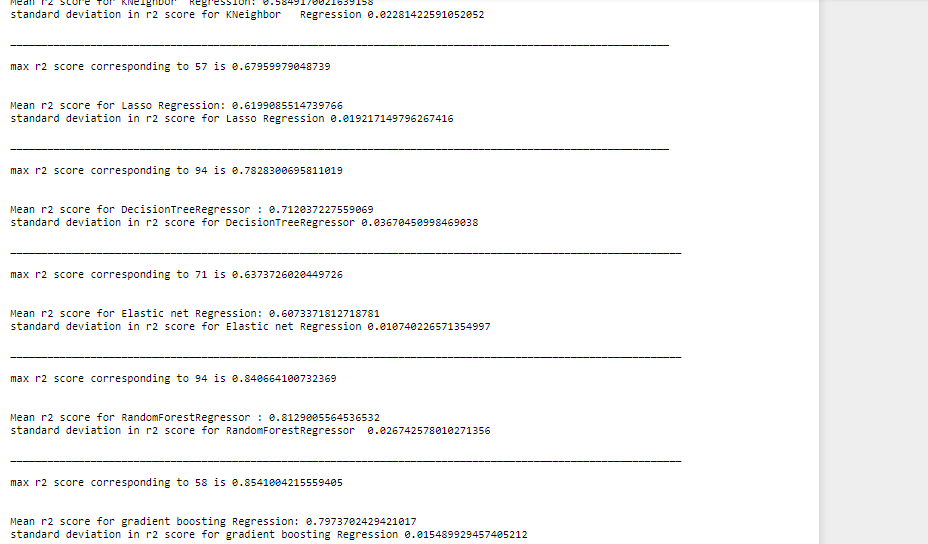
To estimate the accuracy of a linear kernel support vector machine on the dataset by splitting the data, fitting a model and computing the score (n=5 or any number provided by you)consecutive times (with different splits each time):

The **Area Under the Curve (AUC)** is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes

**Receiver Operating Characteristic(ROC)** summarizes the model’s performance by evaluating the trade offs between true positive rate (sensitivity) and false positive rate(1- specificity). For plotting ROC, it is advisable to assume p > 0.5 since we are more concerned about success rate.

ROC summarizes the predictive power for all possible values of p > 0.5. The area under curve (AUC), referred to as index of accuracy(A) or concordance index, is a perfect performance metric for ROC curve. Higher the area under curve, better the prediction power of the model.

* **Results for all algorithms**

****

## Interpretation of the Results

We are getting best result from the gradientboost regressor with 85.4033% accuracy.

