|  |  |
| --- | --- |
| Flight price prediction using machine learning  Capstone project 1 | Abstract  The aim of the project is to predict the price of flight ticket of various airlines, passengers travelling from one city to another city through flights. In order to make predictions a machine learning model is built which is trained on a dataset containing flight price and other factors. The model building is done step by step, the dataset is pre-processed and well analysed. Multiple models were built and the best model was concluded by comparing the individual model performance.  Nagalakshmi S  PGA-11 |

# Flight Price Prediction using Machine Learning Algorithms

# Problem Statement :

Flight ticket prices can be something hard to guess now a days. With the increase in number of Airlines in a Country providing the best travel experience at different cost, the price of a ticket varies depending on which airline the passenger is travelling and the duration of Journey and many other factors.

To solve this problem, A machine learning model is built which is trained on the dataset (Source: kaggle) containing the prices of flight ticket for various airlines between the months of March and June of 2019 travelled across various cities of India.

The Dataset contains the following features:

1. Airline: The name of the airline.
2. Date\_of\_Journey: The date of the Journey.
3. Source: The source from which the service begins.
4. Destination: The destination where the service ends.
5. Route: The route taken by the flight to reach the destination.
6. Dep\_Time: The time when the journey starts from the source.
7. Arrival\_Time: The time of arrival to the destination.
8. Duration: Total duration of flight journey.
9. Total\_Stops: Total stops between source and destination.
10. Additional\_Info: Additional Information about the flight.
11. Price: The price of the ticket (target variable).

# Content

In order to build the model, we have performed the following steps:

1. Importing Packages
2. Reading the data
3. Prepare and analyze data
   1. Understand the data
      1. Checking datatype
      2. Statistical Summary
      3. Exploratory Data Analysis
      4. Correlation
      5. Missing value treatment
      6. Checking for Outliers
   2. Prepare data
      1. Check for Normality
      2. Dummy encoding the categorical variables
4. Feature Selection
5. Model Building:

* Linear Regression
* Decision tree
* Random Forest
* Random forest using GridSearchCV
* Bagging meta estimator
* Adaboost
* Xgboost

1. Conclusion

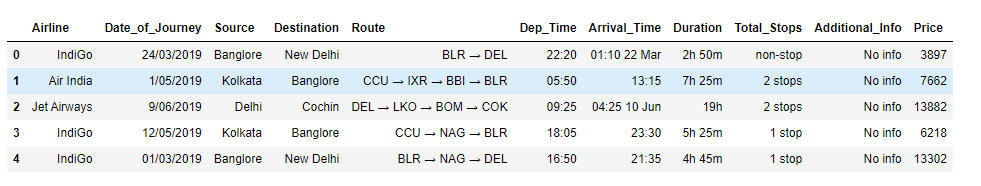
## 1.Importing the Packages

import the necessary packages to build a regression model.

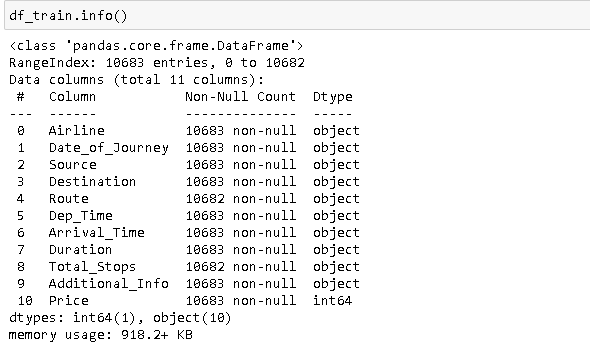
## 2. Read the Data

reading the excel file where our dataset is stored.





Let us check the data type of features using info() method.



We can see that the dataset contains total of 11 features out of which 10 features are object datatype and one is integer which is our target variable (Price). Also, the dataset contains 10,683 observations.

3.1 Understand the Data

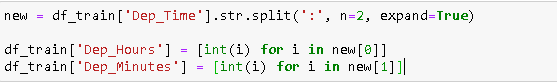
### 3.1.1 Data Type

In order to build the model first we need to check whether the data type of variable is correctly identified, if not we need to convert them.

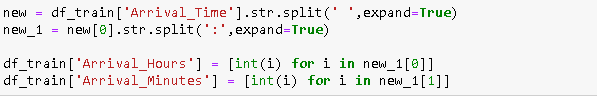
* Date\_of\_Journey: This column contains the date of a journey; from this we extract the day and the month of a journey and store them in new variables.



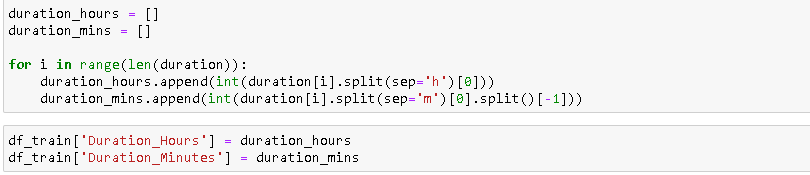
* Dep\_Time: the time when the journey starts from source. The column is split into two and stored them as ‘Dep\_Hours’ and ‘Dep\_Minutes’ in new variable.



* Arrival\_Time: Arrival time includes day and month along with time in some of the rows, so we ignore the day and month and only considered the hours and minutes in it and stored them in new variable.



* Duration: The column contains duration of journey. We extract the hours and minutes and store them in new variables.

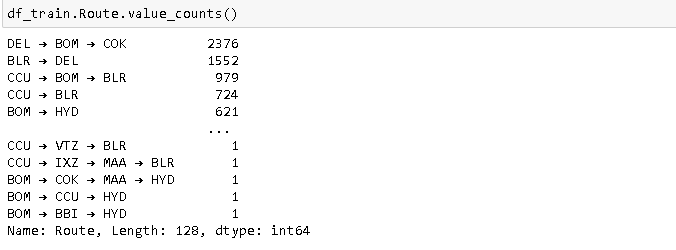


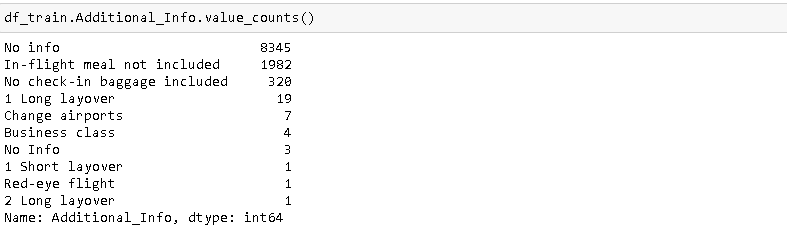
* Total\_Stops: the column contains 5 unique values indicating the number of stops the flight lands at. So, we label encode these values.



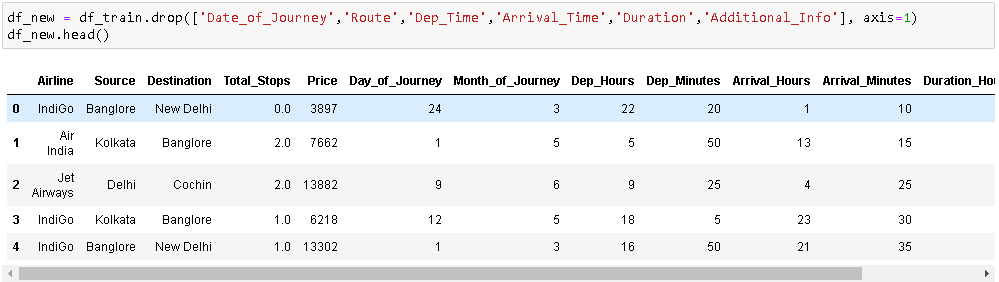


* Route: The columns Route and Total\_Stops are strongly correlated to each other. Hence the column Route is redundant in the presence of Total\_Stops.



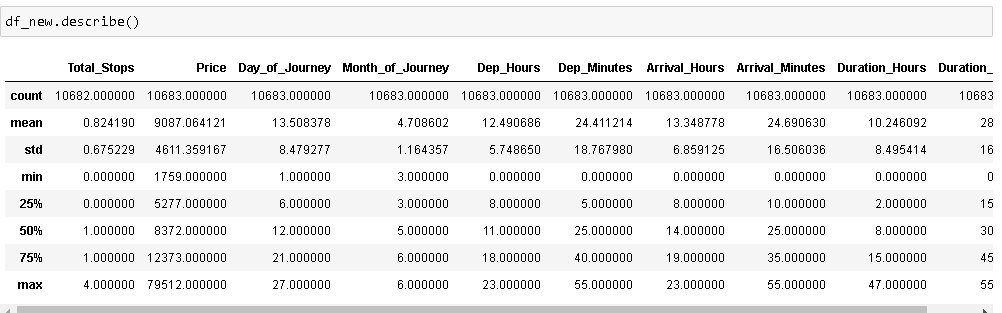
* Additional\_info: In this column 78% of the rows does not contain any Information, so it is better to drop this column.

Let us now drop the redundant columns from the data frame and print the new data frame.



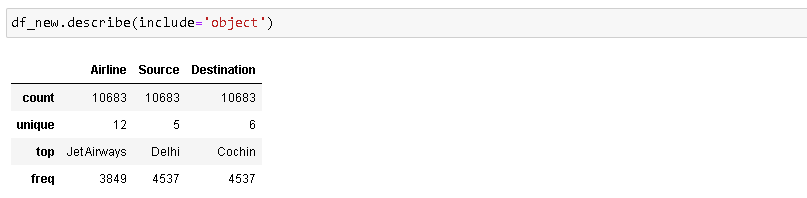
### 3.1.2 Summary Statistics

Checking the Summary Statistics using describe() function.



The summary statistics for numerical variables displays properties like mean, standard deviation, quantiles, minimum, maximum values.

In the above output we can see that the average ticket price is 9,087 rupees and ranges between 1,759 and 79,512. The maximum Stops are 4. Month of Journey between March to June. Maximum duration of journey is 47hours and 55 minutes.

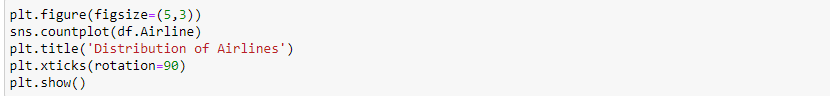


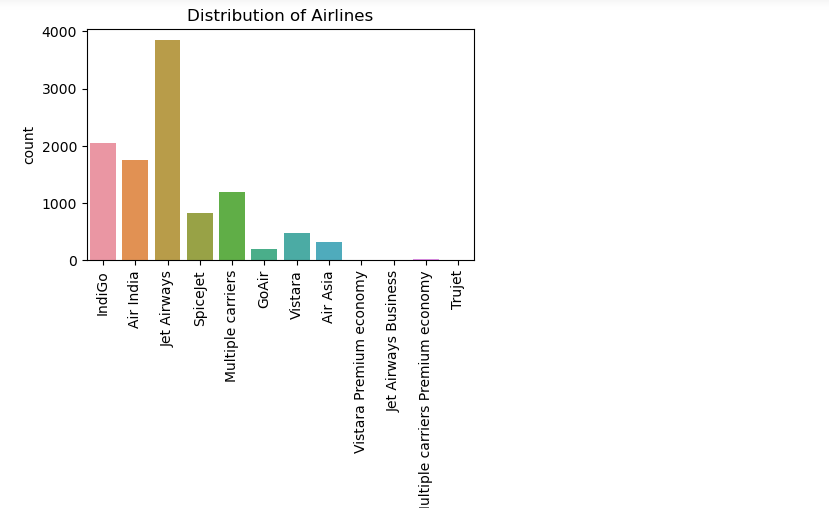
The summary statistics for categorical variables displays properties like count, unique, top, frequency(mode) values.

From the above output we can say that the Airline variable contains 12 unique values that means there are 12 different airlines, most of the passengers travelled in ‘Jet Airways’ and most repeated Source place is ‘Delhi’, most common destination is ‘Cochin’.

### 3.1.3 EDA

Let us visualize few of the variables to get some insights.





#### Insight: from above visualization we can clearly say that 'Jet Airways' is the airline with highest number of people travelled in it, also we can notice that airlines 'Multiple carriers Premium economy', 'Jet Airways Business', 'Vistara Premium economy', 'Trujet' have very least number of people preferring them. This leave us with an assumption that 'Jet Airways' have a best service and reasonable ticket prices which attracts passengers

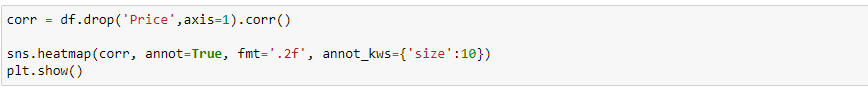
#### 

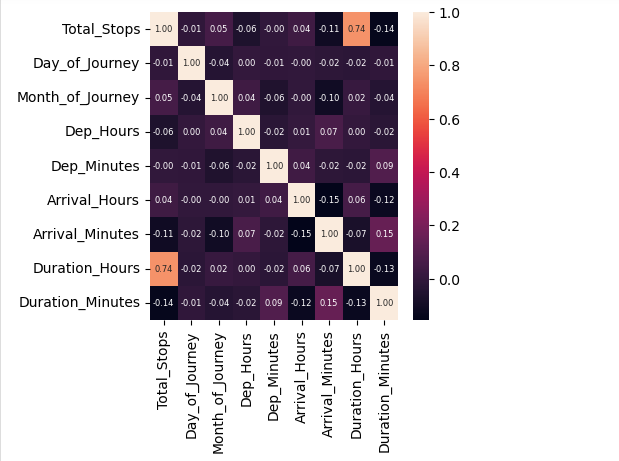
#### Insights: the above output shows the relationship between variables Total\_Stops and Price. from the above result we can say that these two variables are positively correlated, since the ticket prices are increasing as the Total\_stops increasing.

## 3.1.4 Correlation

Correlation explains how one or more variables are related to each other.

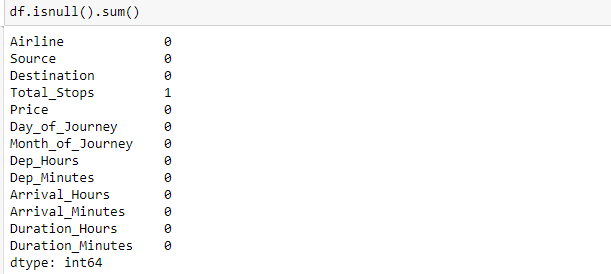
It measures the degree of relationship between two variables, with a correlation >0 indicating positive correlation, <0 indicating negative correlation and 0 indicating no linear relationship.



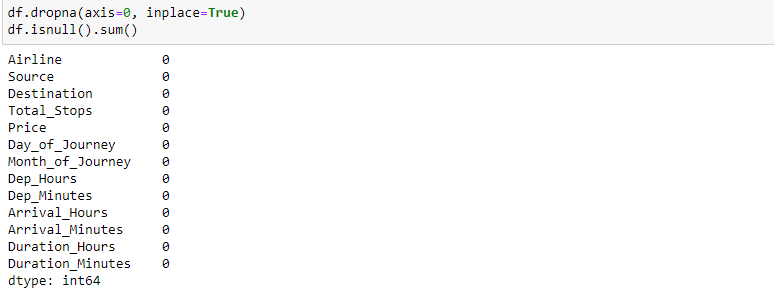


#### Summary: the above heatmap shows that the variable 'Total\_Stops' have a strong positive correlation with 'Duration\_Hours' having a correlation 0.74. and all the other variables do not have any strong correlation between them.

## 3.1.5 Missing Value Treatment

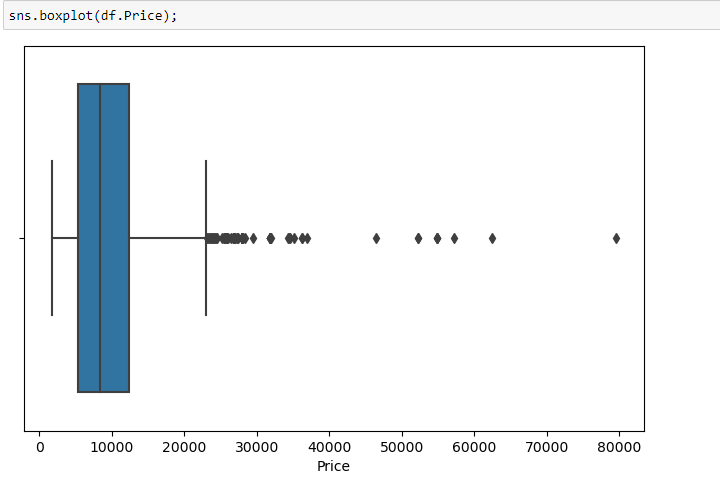


**As we can see that we have only one missing value in our dataset. So, we drop that missing value**.

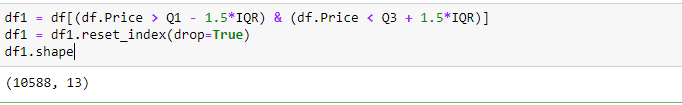


## 3.1.6 Checking For Outliers

Outlier is an extreme high or extreme low data point compare to the rest of the data points present in the dataset.



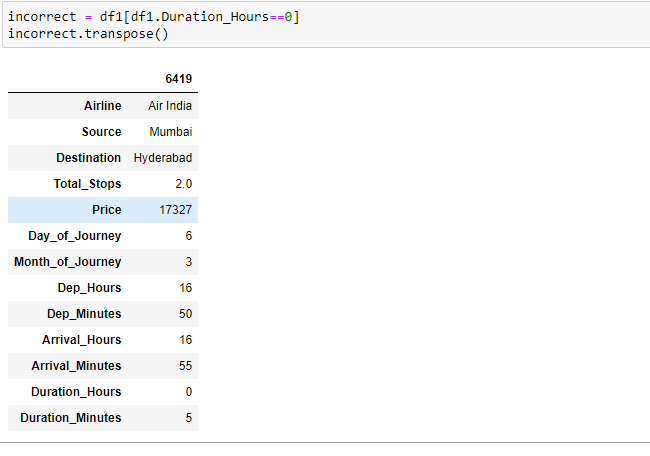
#### Summary: From the above boxplot we can clearly notice that there are many outliers present in the dataset. this will highly deviate the models and results in less accuracy. Hence, we remove them.



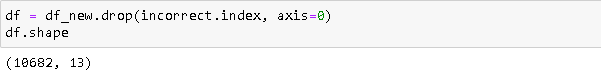
## Identifying Incorrect Records

There might be a chance of having some unusual records in the dataset.

As we know that duration time cannot be zero, let us check if any such records are present.



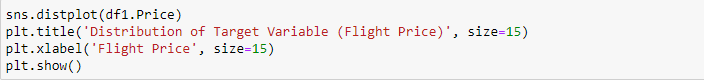
the above output returned one observation with Duration Hours= 0 and Duration Minutes= 5, indicating the Total Duration as 5 minutes from Mumbai to Hyderabad and total stops= 2. This is clearly an incorrect observation because the flight cannot reach to Hyderabad from Mumbai in just 5 minutes. so, we drop this row.

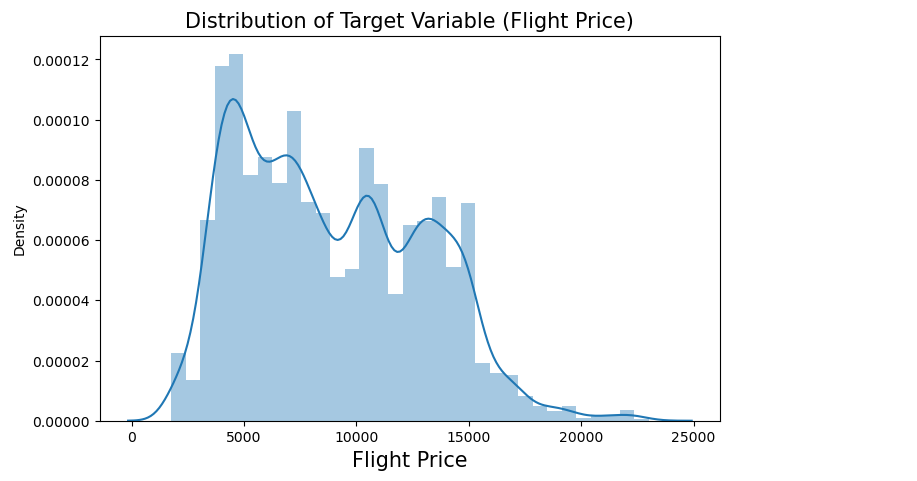


## 3.2 Prepare Data

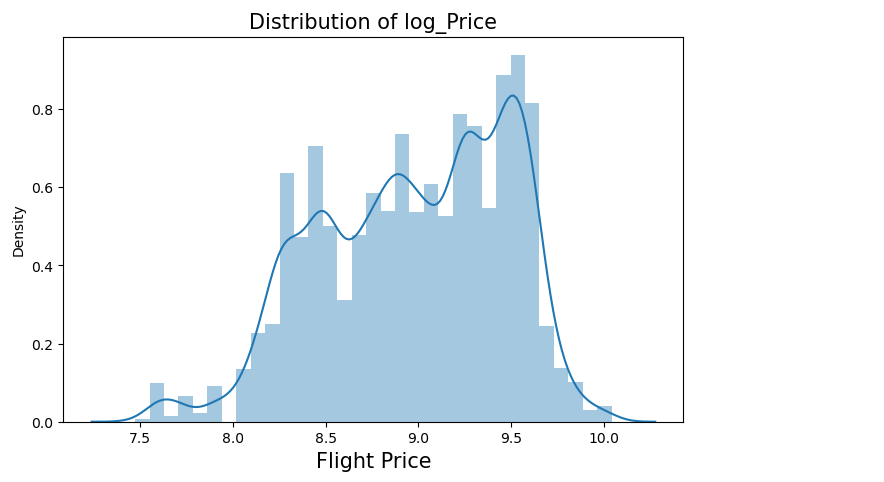
## 3.2.1 Check For Normality

Normality test is used to check if the data is normally distributed.



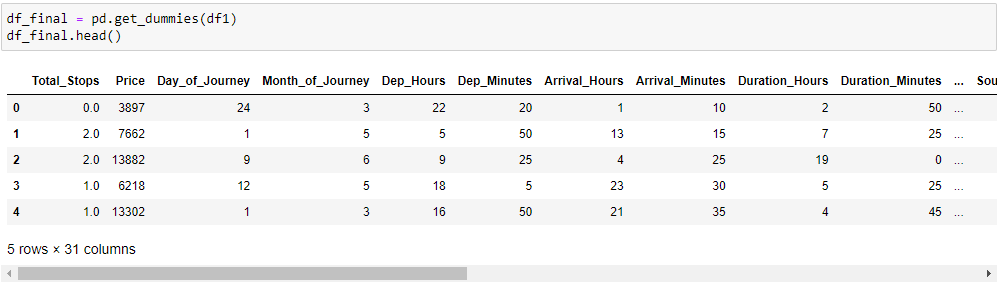


#### Summary: The above result shows that the variable Price is right skewed. So, to normalise the distribution we take log of target variable.



#### Summary: we can see that the log price is left skewed and is not much better than the actual price. Hence, we use the actual price.

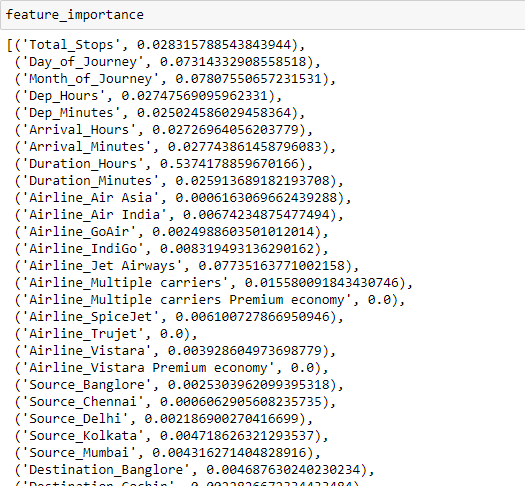
## 3.2.2 Dummy Encoding The Categorical Variables



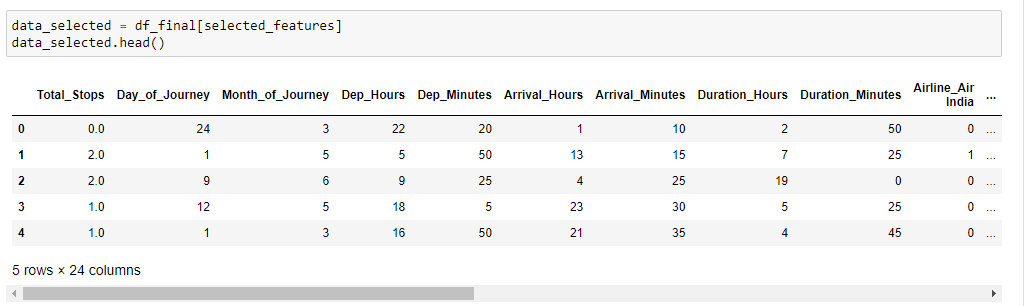
## 4. Feature Selection

Feature selection is used to make the process more accurate. It also increases the prediction power of algorithms by selecting the most critical variables and eliminating the redundant and irrelevant ones.

We can do feature selection using various techniques including Random Forest Importance, Chi-square test, SelectKBest etc.



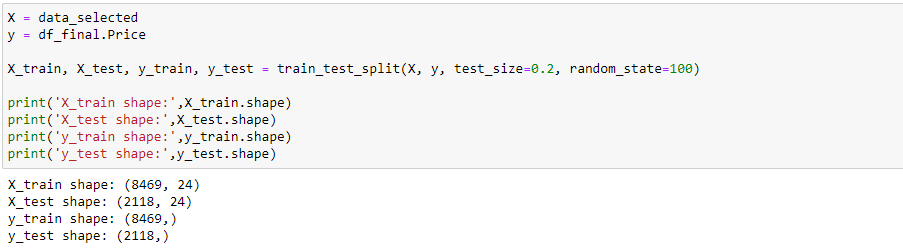
The above result shows the importance of each variable in the dataset obtained by applying the technique Random Forest Importance. As we can see that most of the variable value lies above 0.001. so, we select the important variables by setting the threshold value equal to 0.001.



After feature selection the number of columns reduced to 24 from 30.

## 5. Model Building

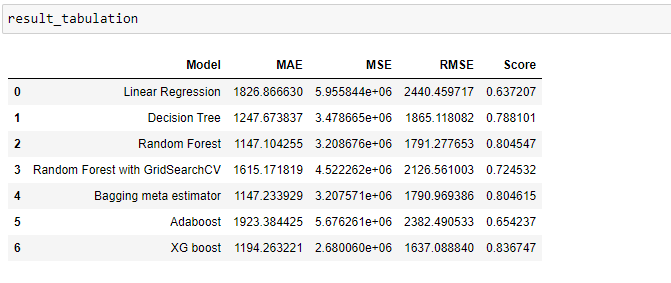
splitting the data into train and test set using 80-20 rule.



Since the target variable is continuous, regression algorithms should be used. multiple regression models are built using machine learning algorithms such as Linear Regression, Decision Tree, Random Forest, Random Forest using GridSearchCV, bagging meta estimator, AdaBoost and Xgboost and tabulated the result of each model, calculating the performance metrics like MAE (mean absolute error), MSE (mean squared error), RMSE (root mean squared error), Accuracy Score.

Performance metrics:

* MAE – measures the average of absolute errors.
* MSE – measures the average of the squares of the errors.
* RMSE – it is the square root of MSE.
* Accuracy Score – calculated by taking average of residuals (difference between actual and predicted values).



## 6. Conclusion :

In the above tabulated result, we can see that the Xgboost model has given a best performance with less RMSE score and the accuracy score of 83% compare to the other machine learning models.

Hence, we can conclude that the model Xgboost is the best performing model.