

**Flight Price Prediction**

**Submitted by:**

Akash Kanjwani

**ACKNOWLEDGMENT**

During the process of completing this project, I have referred following materials for which I owe them great gratitude.

1.Data collection- Using Web scraping tool i.e. Selenium

2. Car24.com for collecting the data *easemytrip.com/*

3. Data trained video tutorials.

4. Scikit-learn <https://scikit-learn.org/stable/>

5. Machine Learning for Dummies by John Mueller and Luca Massaron - Easy to understand for a beginner book.

6. Geeksforgeeks. <https://www.geeksforgeeks.org/>

Besides that all the observation, creations of the models and graphs done by self help.

**Problem Statement-Flight** ticket prices can be something hard to guess, today we might see a price, check out the price of the same flight tomorrow, it will be a different story. We might have often heard travellers saying that **flight** ticket prices are so unpredictable.

**Motivation for the Problem Undertaken**

There are two main use cases of flight price prediction in the travel industry. OTAs and other travel platforms integrate this feature to**attract more visitors looking for the best rates.** Airlines employ the technology to forecast rates of competitors and adjust their pricing strategies accordingly.

**Analytical Problem Framing**

Mathematical/ Analytical Modelling of the Problem

1) The size of table is 8946\*6 i.e. no. of rows are 8946 and no. of columns are 6(including target).

2) All the columns are object type in nature because we got the data using web scraping, In the upcoming steps we will change it where necessary.

3) Null values are not present in the data set as we can see in this seaborn heatmap, so there is no need to adopt imputation technique, but we can get some blank spaces, so we will drop that blank spaces.

4) In case of object data type, we will apply the encoding technique to convert the values in the numeric format.

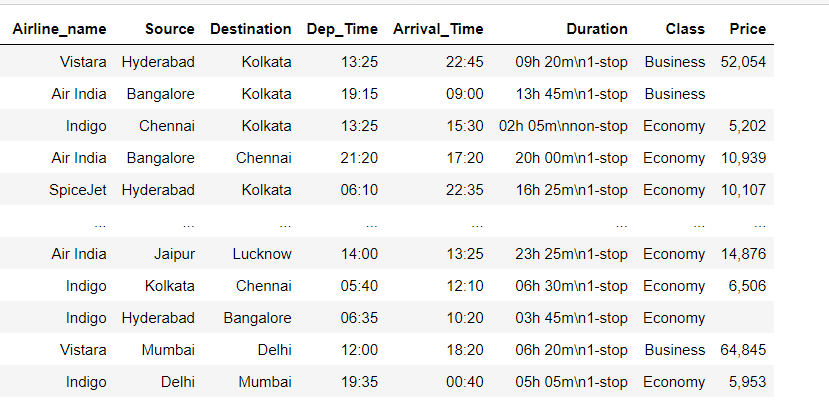
In order to scrape the data we got some garbage data as well like:

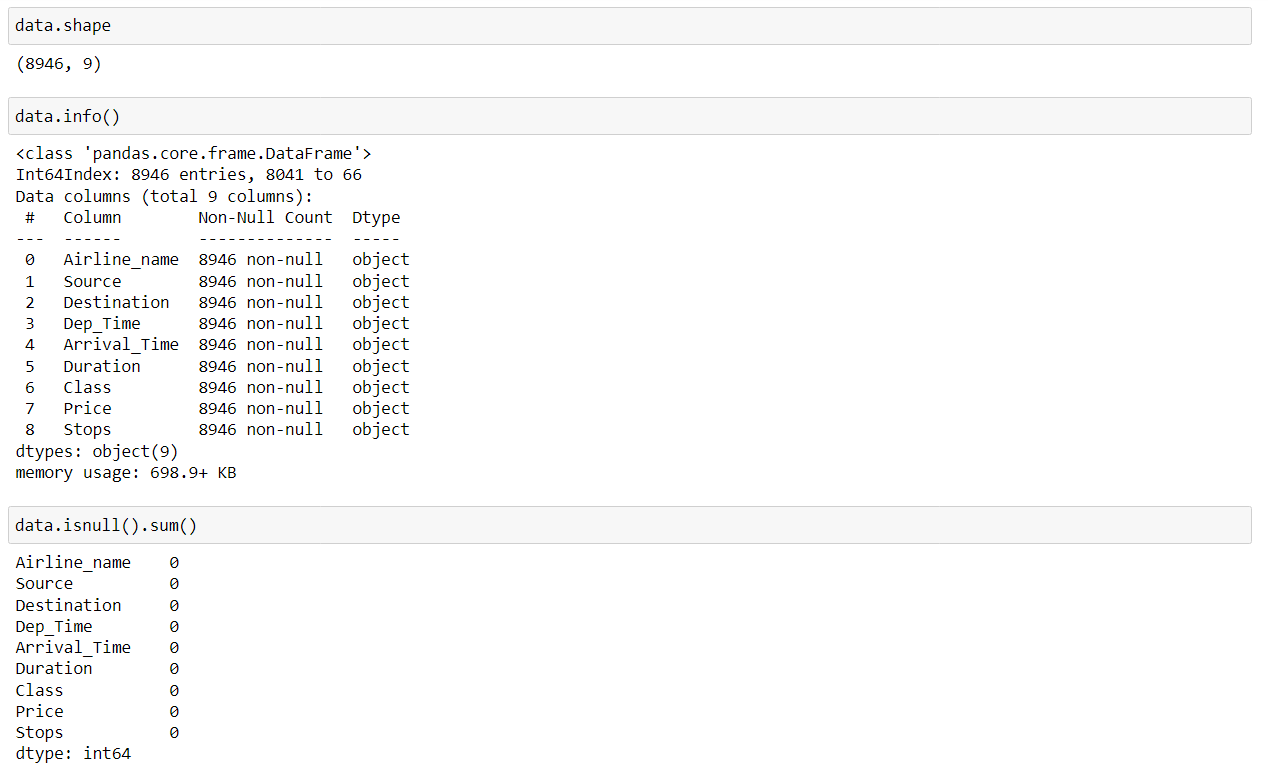
* Price column is present in the the form of object datatype.
* In every row of price column comma(,) was present.
* we had some empty space in between the columns.
* We done some changes in the stop column because 1-stop had been written in the some other formats, so we changed that and winded up in the 1-stop name.

All the above garbage data we removed from the dataset with the help of pandas functions.

* In our dataset only 1 columns are showing some outliers that is price. Because the mean is 23062 and standard daviation is 21016 and maximum max values is 100947 which is very higher and reason of outliers.
* High Skewness is present in the kilometers driven column(1.19), shows that data data are not equally distributed.

Data Sources and their formats

Initially we did not data, For getting the data we used web scraping selenium tool Data source- *easemytrip.com*

**

**Data Pre-processing Done**

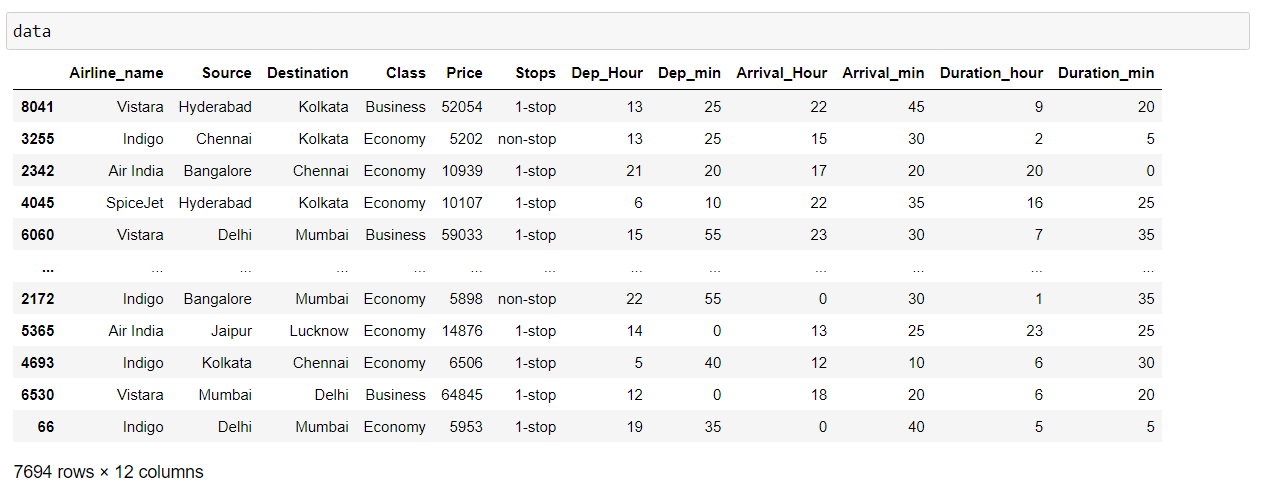
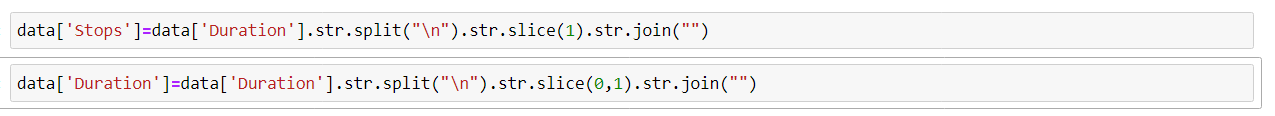
**Outliers, Skewness, Data scaling, Data Cleaning**

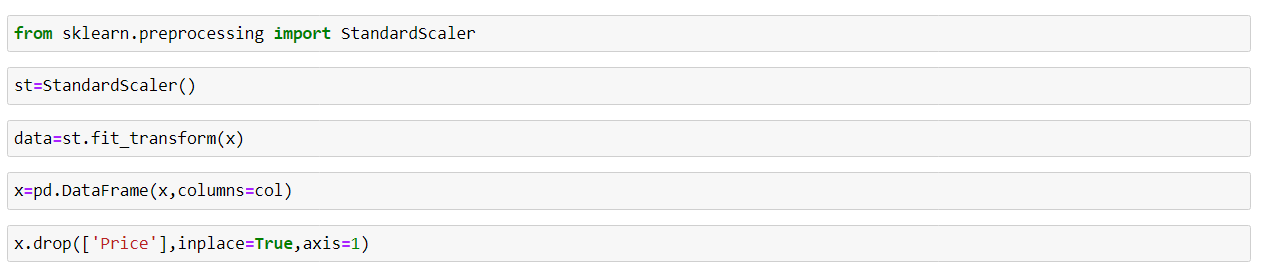
Because we had categorical data and absence of numerical data that is why there was no column which can contain outliers and skewness only we had one column i.e. price column, which is numerical in nature but here we can not pre-process the target column.

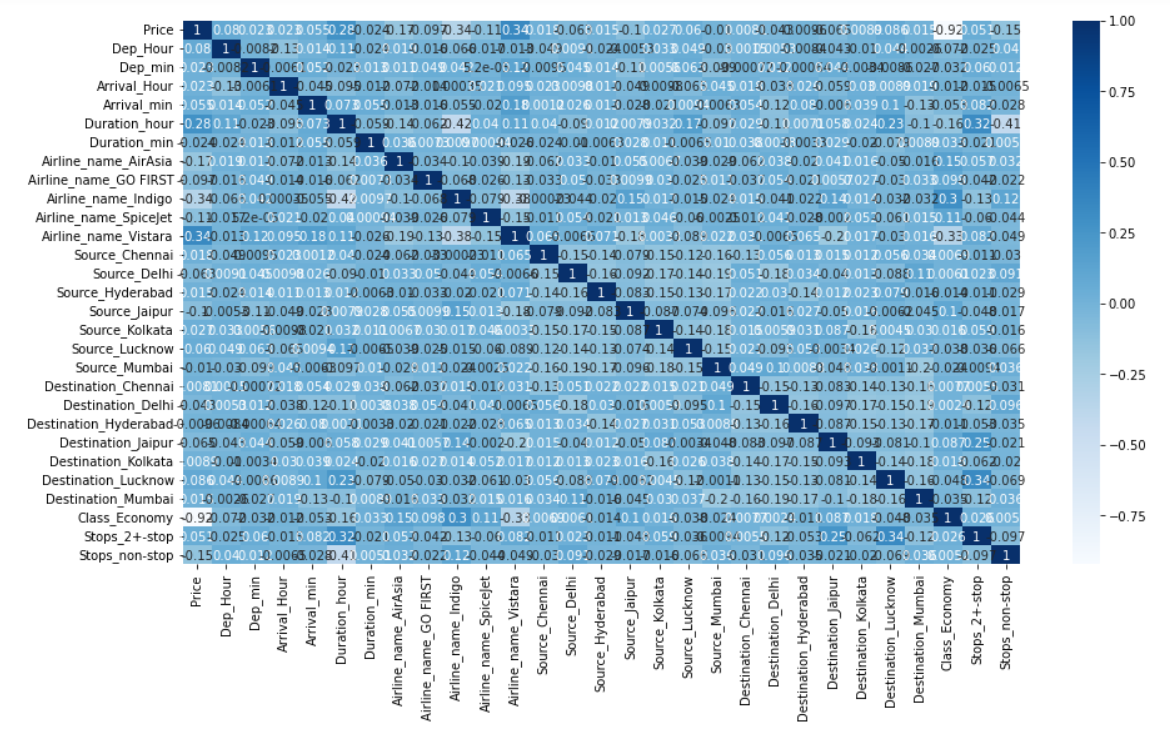
\* For Data scaling we used Min max scaler.

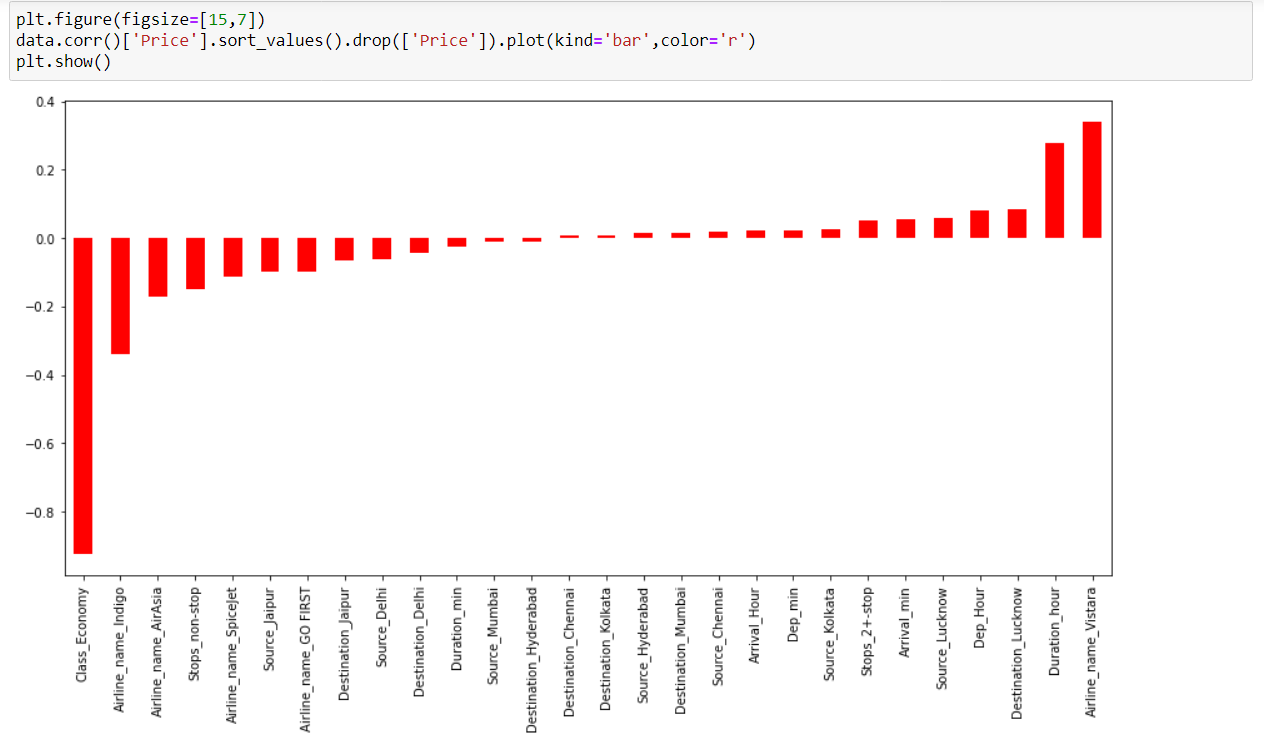
\* Encoding the data with the help of label encoder.

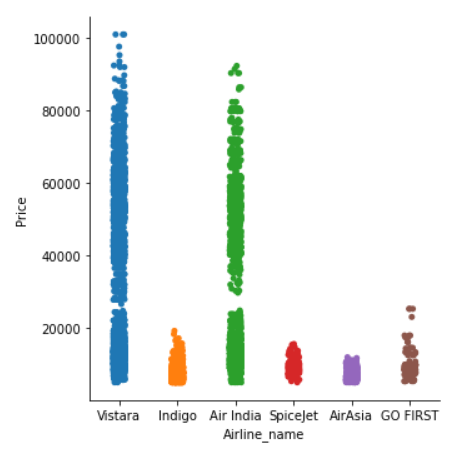
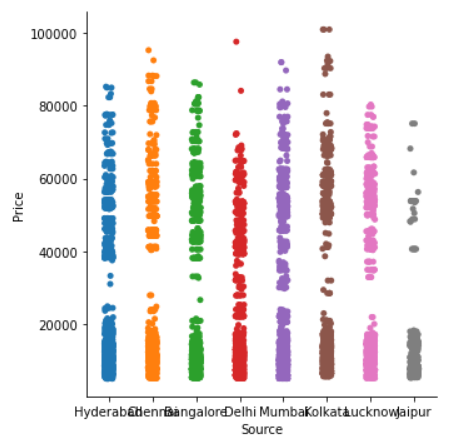
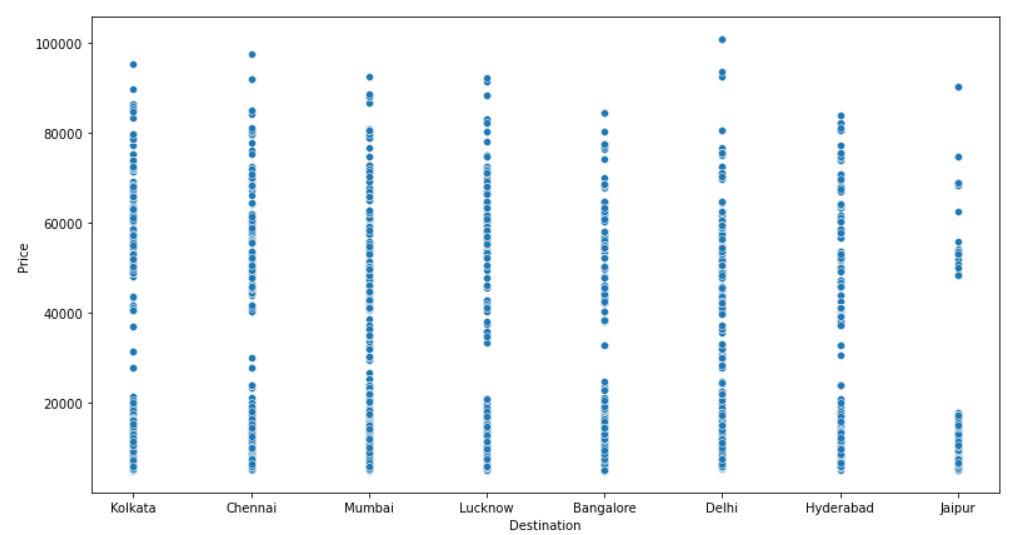
\* Data cleaning using pandas function

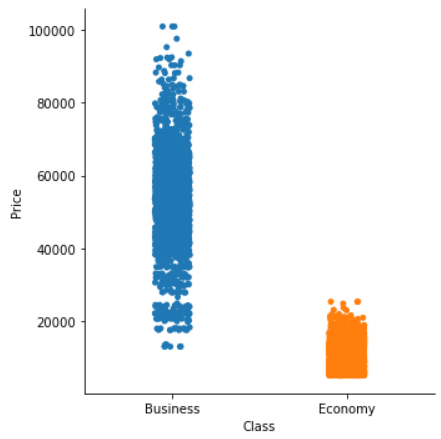
 

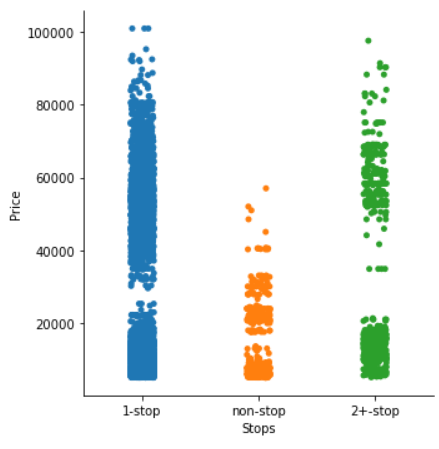
**Scaling the data**

**Data Inputs- Logic- Output Relationships**



****

****

****

1) Price of Vistara and Air india are higher as compare to other airlines.

2) The Ticket price of all the sources and destinations are around same with veay less diffrence.

3) The Ticket price is higher when the class type is business.

4) Most of the airlines takes 1 or 2 stops to reach the destination point, and very less flights takes 0 stops to reach the destination point.

**Hardware and Software Requirements and Tools Used**

**Anaconda Navigator**

**Jupyter Notebook**

**Language-Python**

Selenium

**Many lib.-------**

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

import numpy as np

import warnings

import selenium

warnings.filterwarnings('ignore')

from sklearn.preprocessing import power\_transform

from scipy.stats import zscore

from sklearn.preprocessing import MinMaxScaler

import statsmodels.api as si

from scipy import stats from statsmodels.stats.outliers\_influence

import sklearn

from sklearn.linear\_model import LinearRegression,Lasso,Ridge,ElasticNet

from sklearn.model\_selection import train\_test\_split,GridSearchCV,cross\_val\_score

from sklearn.tree import DecisionTreeRegressor

from sklearn.svm import SVR

from sklearn.neighbors import KNeighborsRegressor

from sklearn.ensemble import RandomForestRegressor,AdaBoostRegressor,GradientBoostingRegressor

import xgboost as xg from sklearn.metrics

import mean\_squared\_error,mean\_absolute\_error,r2\_score

**Pandas**- For making data frame

**Matplotlib and seaborn-** For data visualization

**Numpy-** For numerical python

**MInMAxScaler-** For data scaling

**Power transform**-For removing skewness

**From metrice** - mean\_squared\_error,mean\_absolute\_error,r2\_score -For checking the model accuracy.

**Regression-** For regression modeling

**Ensamble-** For boosting and bagging

**Grid search cv-** For hyperparameter tuning

**Cross\_Val\_Score**- For cross validation

**Model/s Development and Evaluation**

Approaches Firstly it is import to know about which type of modelling we are going to construct, For this problem we used regression models because our target variable is numerical type and we had to predict the flight price. When we go for regression we have to use some metrics like mean\_squared\_error, mean\_absolute\_error, r2\_score In order to do this work we have to find out the best random state by which we can achieve good accuracy. Then we split our data set into the train part and test part using train test split. When we done with the modelling we have to use cross validation for real accuracy(without underfitting and overfitting). Hyperpameter is must for increasing the model accuracy in order to build good model for this we use Grid search cv. We followed all the above approaches to build our Machine learning model.

**Algorithms**

• Linear Regression

• Decision Tree Regressor

• KNeighbors Regressor

• Support Vector Regressor

* Ridge regressor
* Lasso regressor

**For Bagging and boosting:**

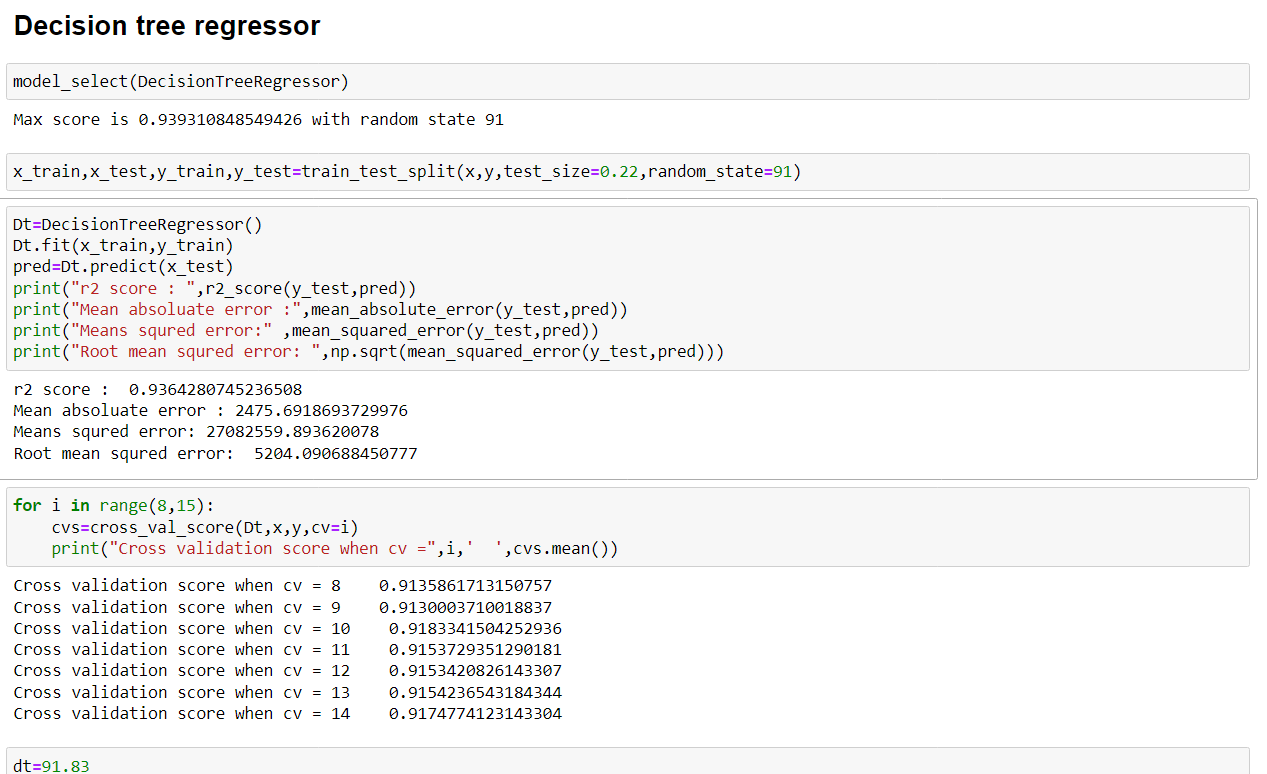
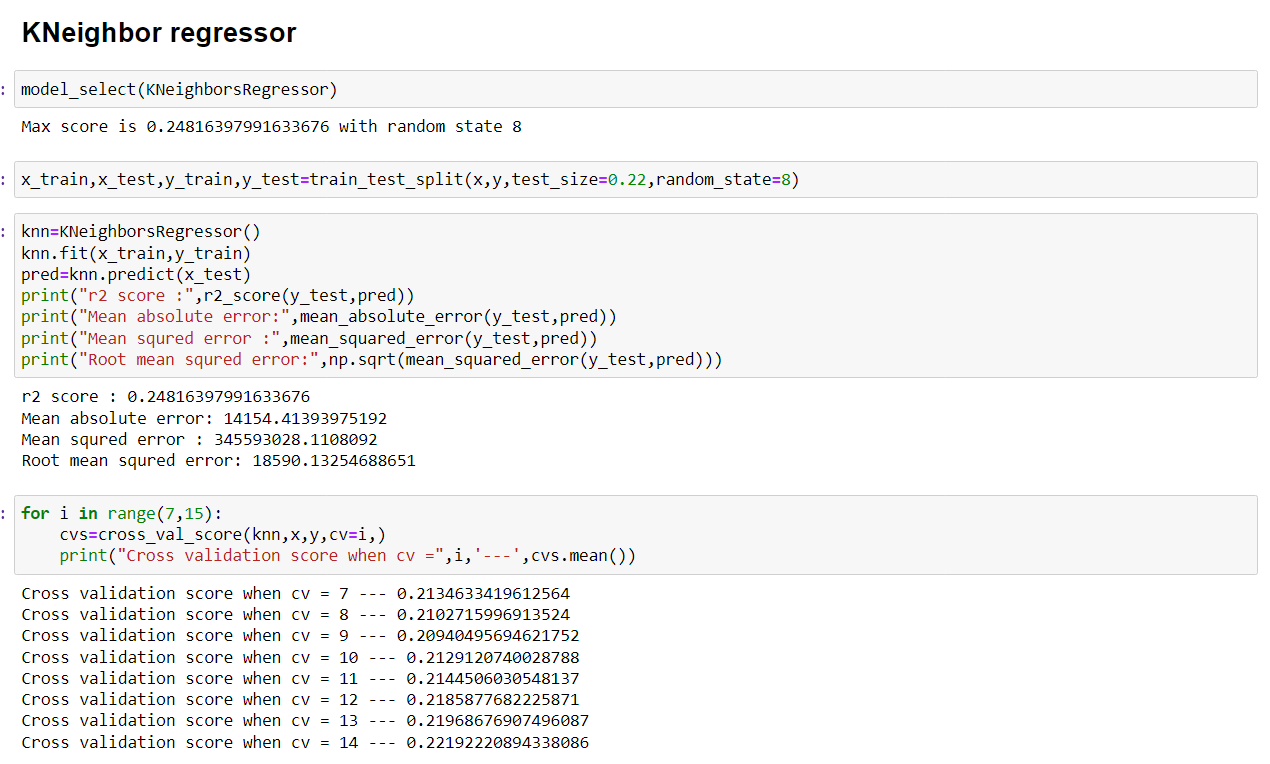
• Random Forest Regresssor

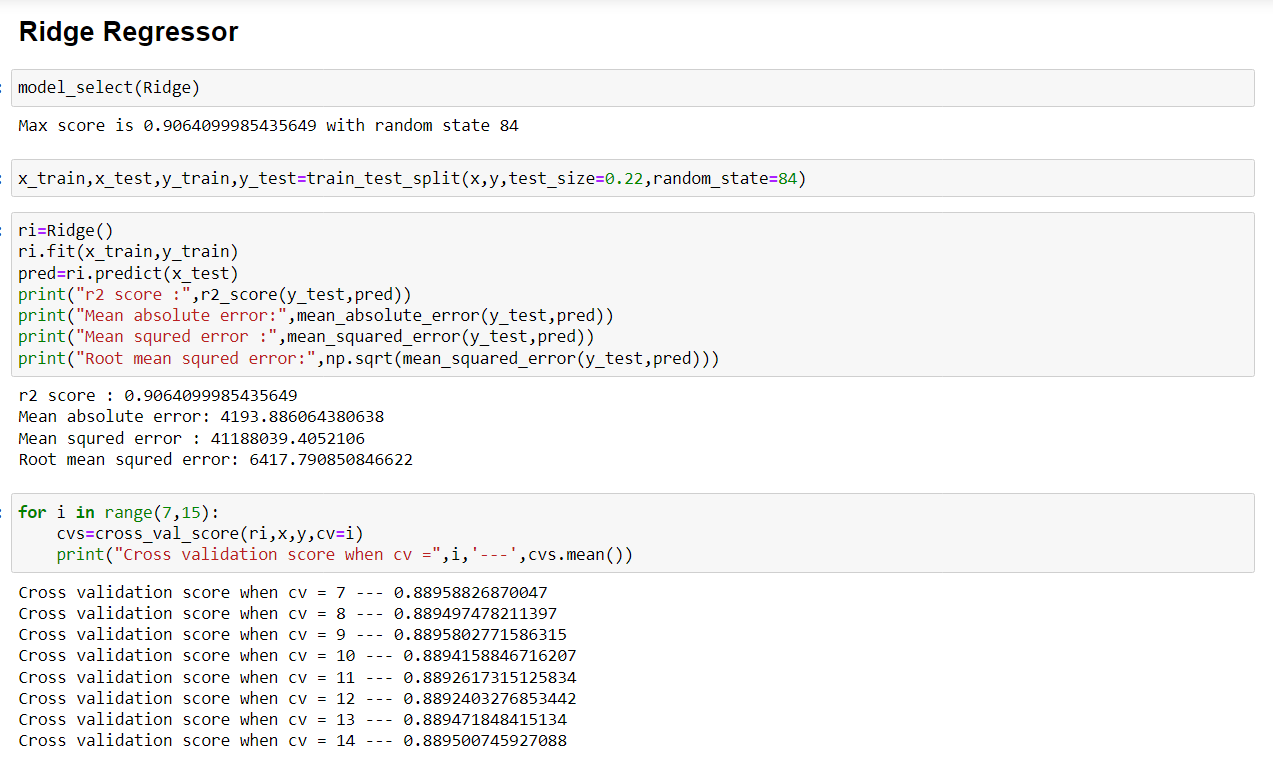
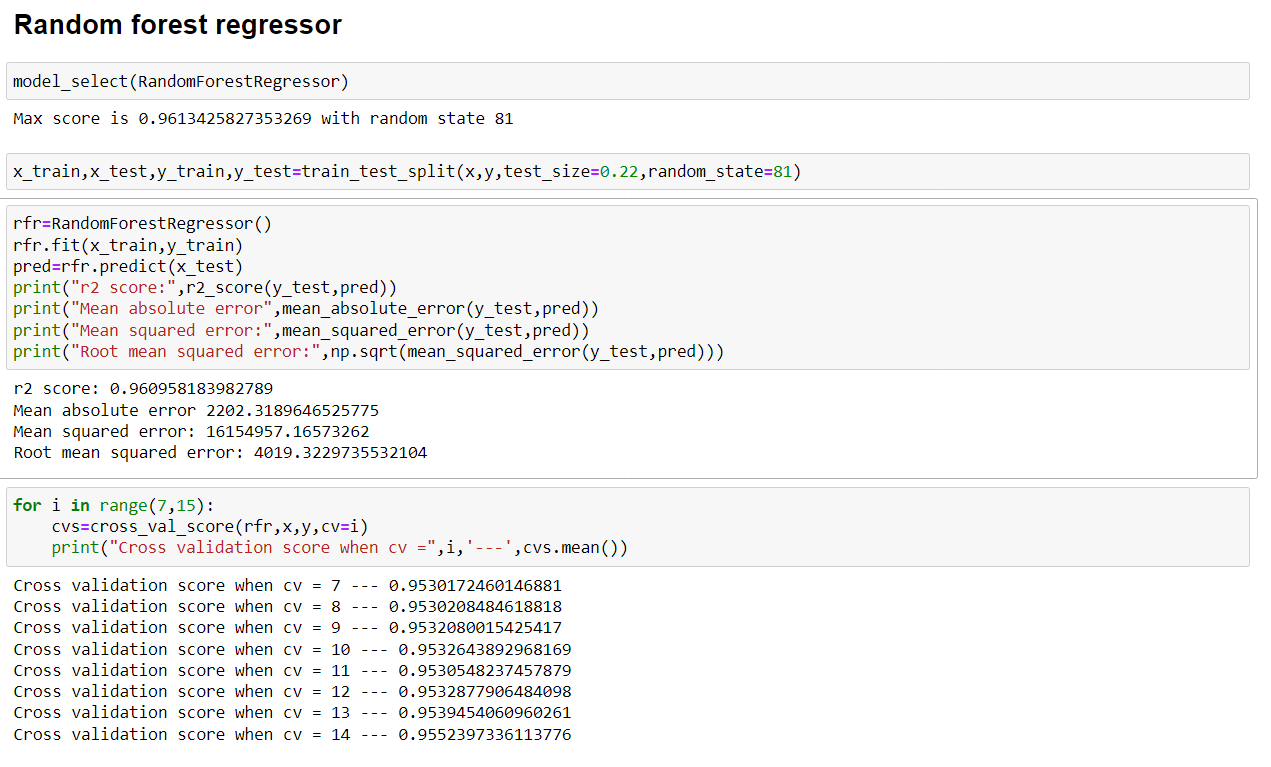
• Gradient Bossting Regressor

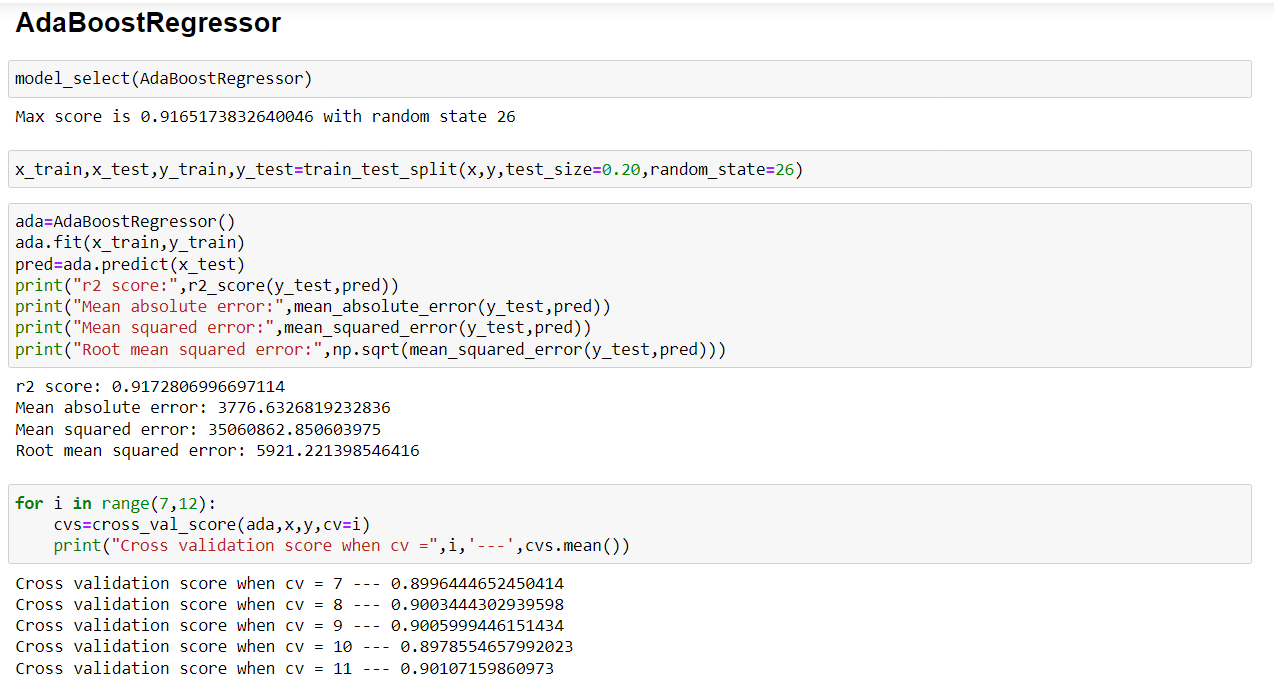
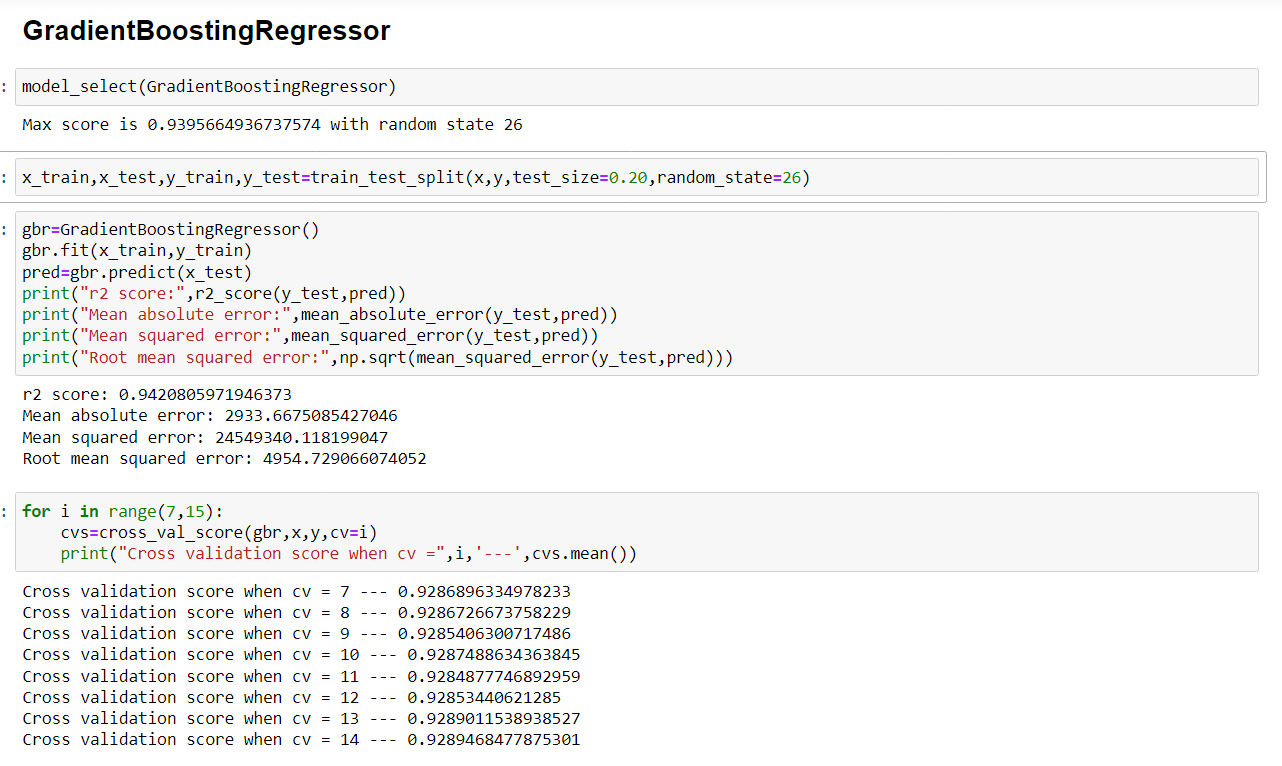
• AdaBoost Regressor

• XgBoost Regressor

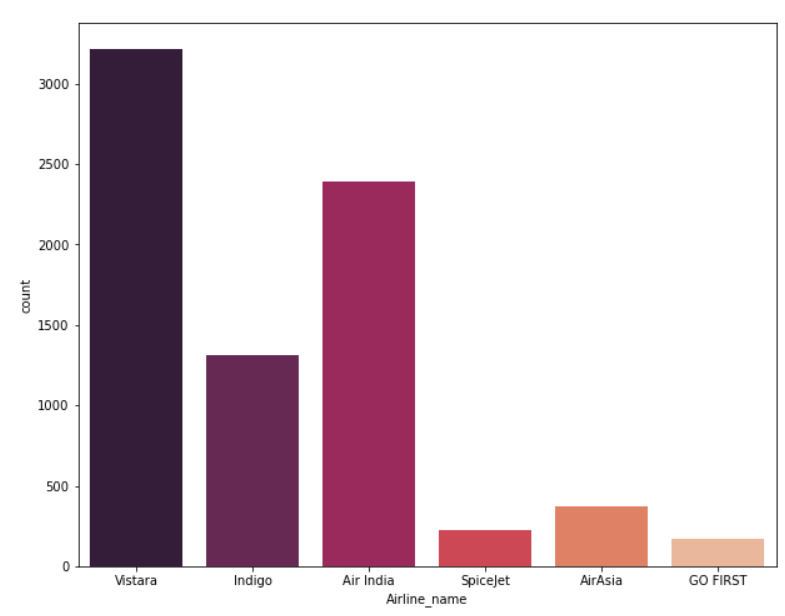


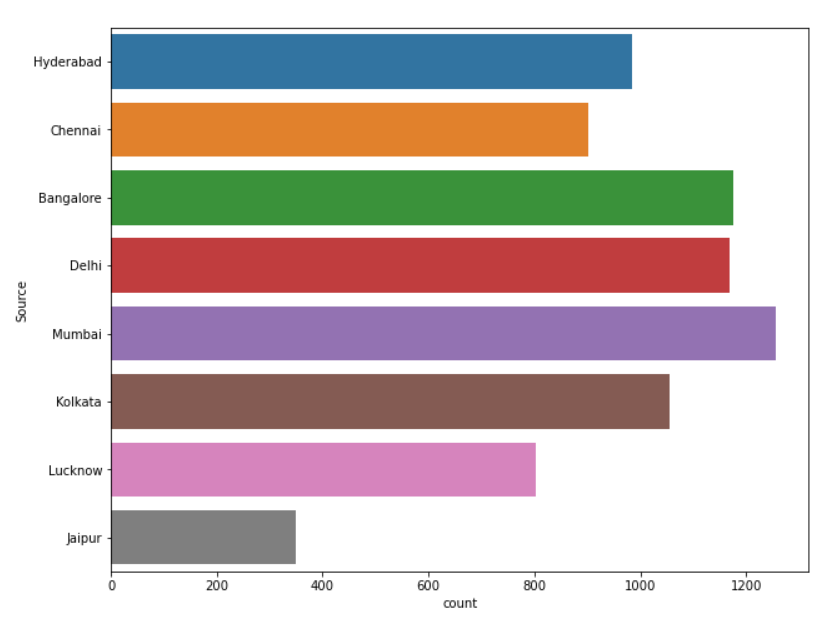
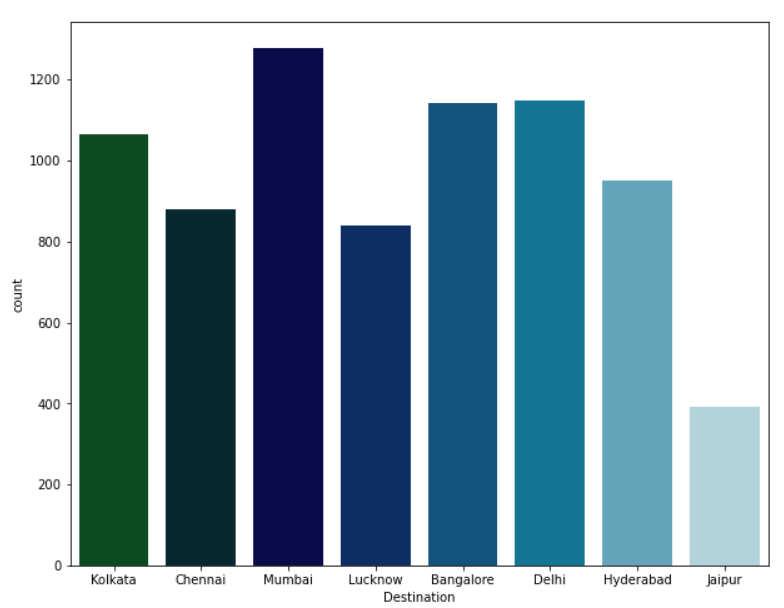


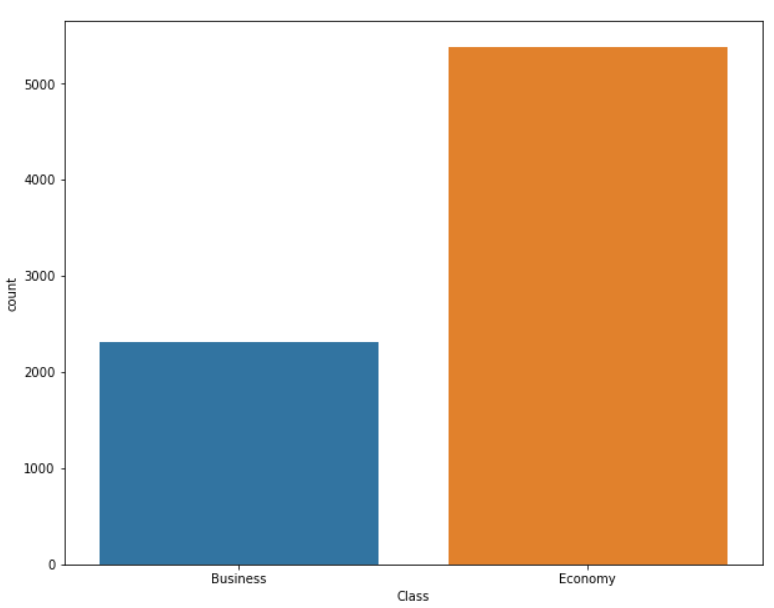


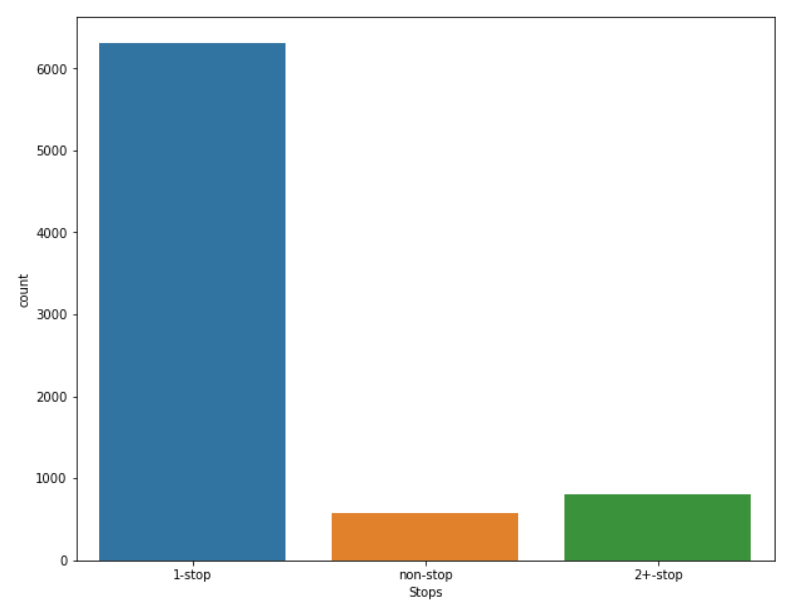




**Visualization**







1) Vistara having a larget number of flights(3218) as compare to other airways.

2) Air india(2392) and Indigo(1314) airways standing over the 2nd and 3rd position and having good flights as comapre to others.

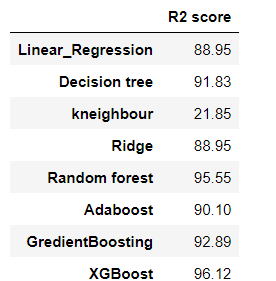
3) Spicejet(228), Air asia(373) and Go first(196) airways have a less number of flights as compare to others.

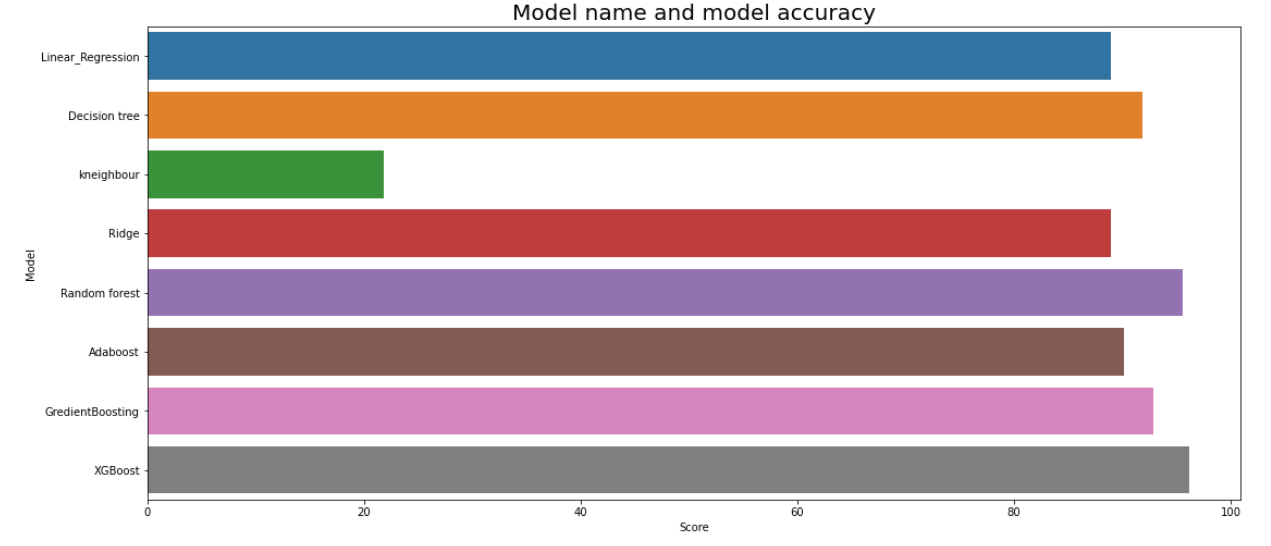
4) Destination- Mumbai(1279), Banglore(1143) and Delhi(1147) cities are the most preferable destination by the peoples.

5) Delhi, Kolkata ,Mumbai , Chennai and Banglore are the main source of flights but mumbai(1255),banglore(1176) and delhi(1169) is the most preferable source for all the airways.

6) When we see the total stops columns, Most of the flights takes one stopage(6314) and there are many flights also which takes zero stopage(576). Number of 2 stopage is very less.

**CONCLUSION**





## Best Model[¶](http://localhost:8888/notebooks/Flight_price(Model_building).ipynb#Best-Model)

### We are chosing Random forest regressor as a best model, Because this model has least diffrence between model acuracy score and cross validation score and also its accuracy is highest as compare to others.That is why we are choosing RFR

### Model accuracy is : 96.17

### Cross\_validation score : 95.54

### Diffrence : 0.62

**About XG boost regressor**

XGBoost stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems.

XGBoost is a powerful approach for building supervised regression models. The validity of this statement can be inferred by knowing about its (XGBoost) objective function and base learners. The objective function contains loss function and a regularization term.

In this paper, we built serveral regression models to predict the pricevarious flights . We eveluated and compared each model to determine the one with highest performance. We also looked at how some models rank the features according to their importance. In this paper, we followed the data science process starting with getting the data, then cleaning and preprocessing the data, followed by exploring the data and building models, then evaluating the results and communicating them with visualizations.

**Flig**

**ht Price Prediction**