

FLIGHT PRICE PREDICTION

Submitted by:

RAVI PRAKASH BAJPAI

**ACKNOWLEDGMENT**

For this project data has been collected through web scraping. Data from different website has been scraped,these are the following websites

Makemytrip.com

easemytrip.com

[www.cleartrip.com](http://www.cleartrip.com)

tickets.paytm.com

I have also taken help from many other websites ,some of them are mentioned below

<https://towardsdatascience.com/>

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

<https://www.analyticsvidhya.com/>

https://www.kaggle.com/

**INTRODUCTION**

This project Flight price prediction is mainly built to predict the price of the air ticket for any future date.

Generally it is seen that price for air ticket raise and fall and is unpredicatable to determine the price for any future date. Variation in the price of air ticket depend on many factors such as flight name(air asia,indigo,indian air line,kingfisher), date,time to travel and many others,so considering these factors this project is built to determine near about price for any air ticket.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

After scraping the data it can be seen that the given data set problem is regression type,as it is seen that label is price of air ticket and it is continous, so it is regression problem. After analysing the other features of the data set it can clearly seen that most are them are categorical data.many features are object data type,these are converted into numeric form.

* Data Sources and their formats

Data for the given project has been scraped through different website dealing with air tickets and tourism, different source and destination for different date has been select for one adult passenger and accordingly data has been scraped using selenium. Some of website has been mentioned below and juypter notebook with dataset in csv file has been uploaded.

Makemytrip.com

easemytrip.com

[www.cleartrip.com](http://www.cleartrip.com)

tickets.paytm.com

* Data Preprocessing Done

It is analysed from data set that label is continous data and nearly all the features are categorical data.it can also seen that data is uncleaned, for cleaning the data following steps has been followed

1. Extracting the day and month from the date.
2. Droping the unnecessary features from the data set(unnamed: 01).
3. Extracting the hour and minute from departure and arrival features.
4. From time of travel extracting the no. of hour to travel.
5. Converting the "no. of stop" from string data type to numeric by using replace function.
6. Many features are object data type converting them into numeric form by applying label encoder.

* Data Inputs- Logic- Output Relationships

In the given data set all the object data type features has been converted into numeric.

After ploting the heat map it is analysed that features are not very closely related with each other. their is no multicolilnearity present in the data set. It is also observed that feature are not closely related to label in the given data set.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

From the given data set it is observed that problem can be solved using regresssion algorithm. It is seen that label of the given problem is in continous form. So all the regression algorithm are used in the given problem.

* Testing of Identified Approaches (Algorithms)

Following algorithm is used for training and testing the

model

Random Forest Regressor

Ada Boost Regressor

XGB Regressor

LinearRegression

* Run and Evaluate selected models

After applying the different algorithm following result has been observed.

Algorithm used R 2 score

Random Forest Regressor .79

Ada Boost Regressor .33

XGB Regressor .73

LinearRegression .40

Considering the random forest regressor algorithm for the given model as it is showing the best result for the given problem.

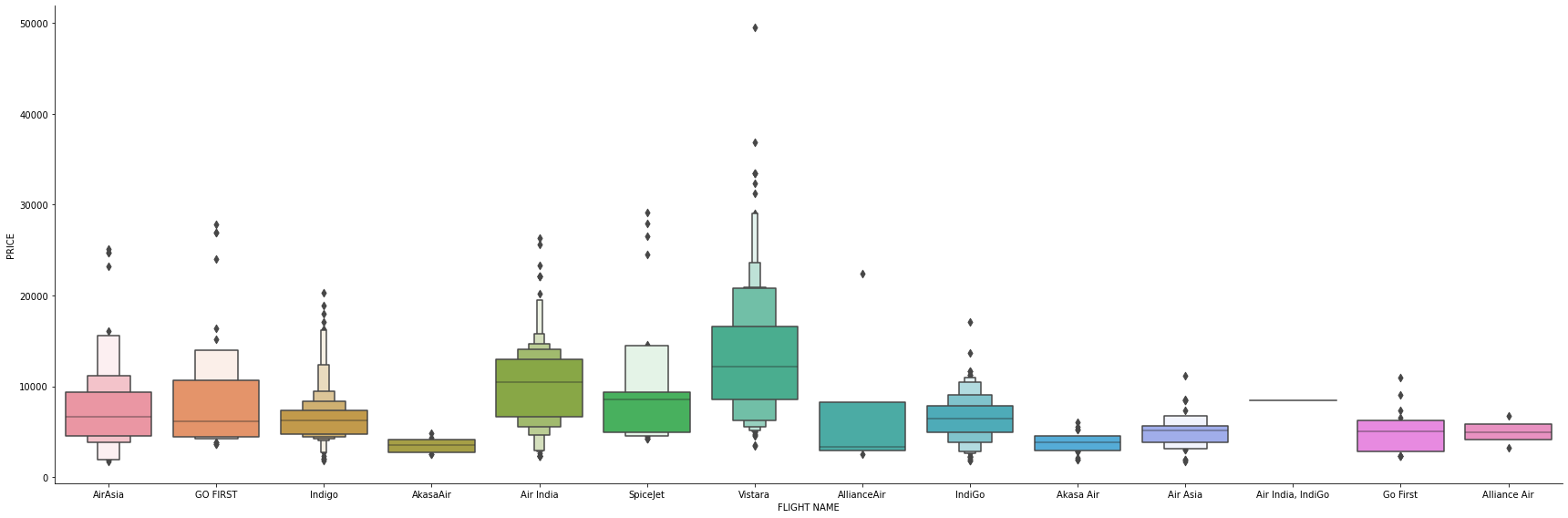
* Key Metrics for success in solving problem under consideration

R square score has been used as a metrics for the given model.

* Visualizations

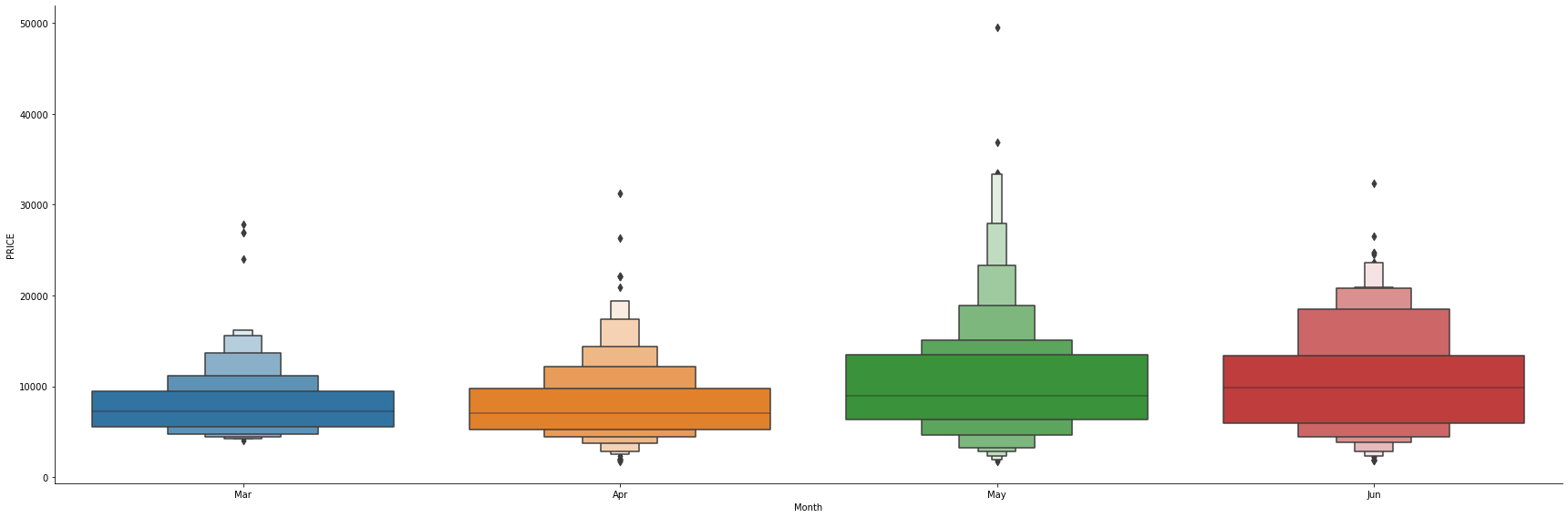
Different plot has been ploted to show the relationship between them and also analying some useful results

1. Plotting Price vs Airline plot



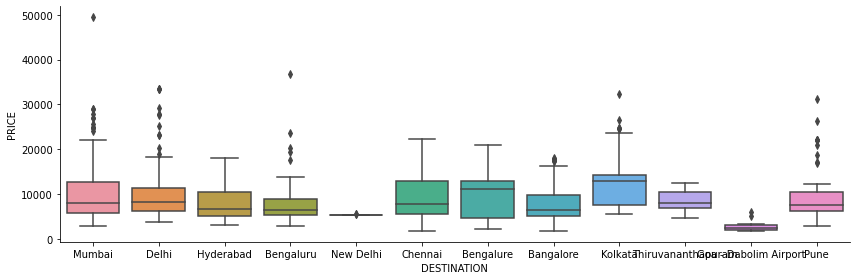
From above plot it can be concluded that vistara airline is expensive and akasaair is cheapest.

1. Plot box for price and month



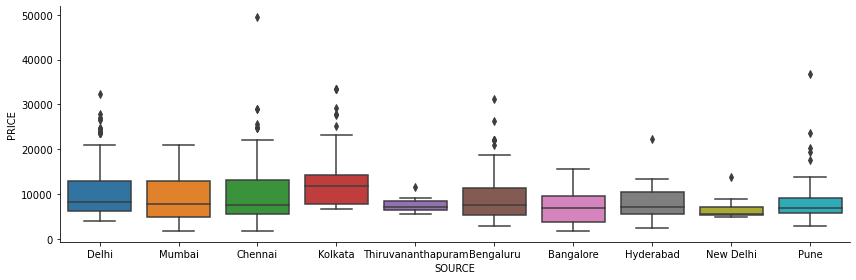
From above plot it can be seen that in month of may price is cheapest and also highest

1. Plotting Box plot for Price vs Destination



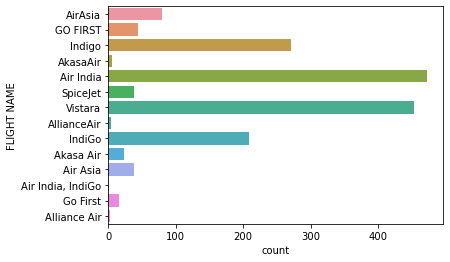
From above result it can be seen that pune have maximum outliers.

1. Plotting Box plot for Price vs Source



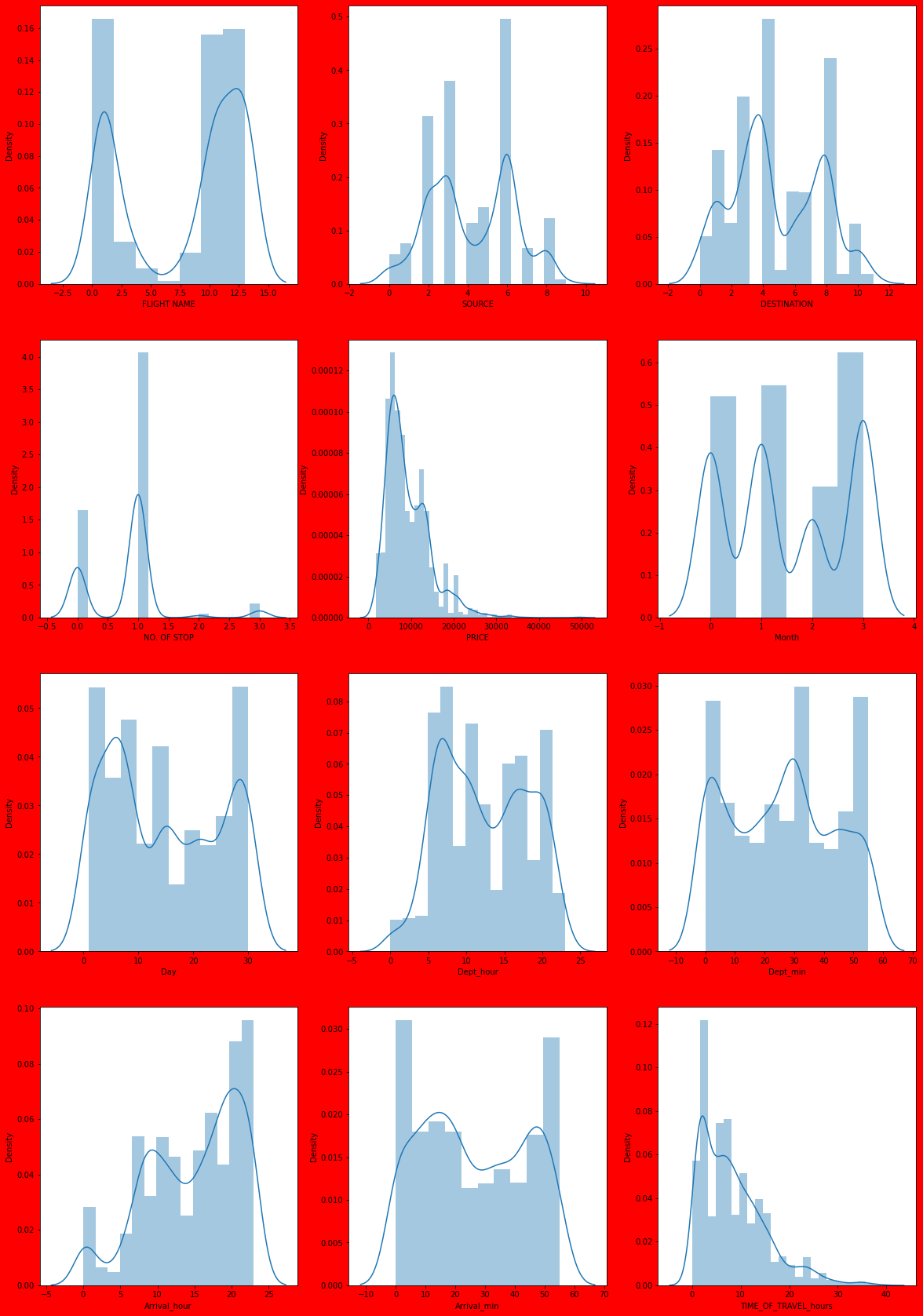
From above result it is seen that delhi and pune have maximum outliers as source.

1. Count of no of flight for different airlines

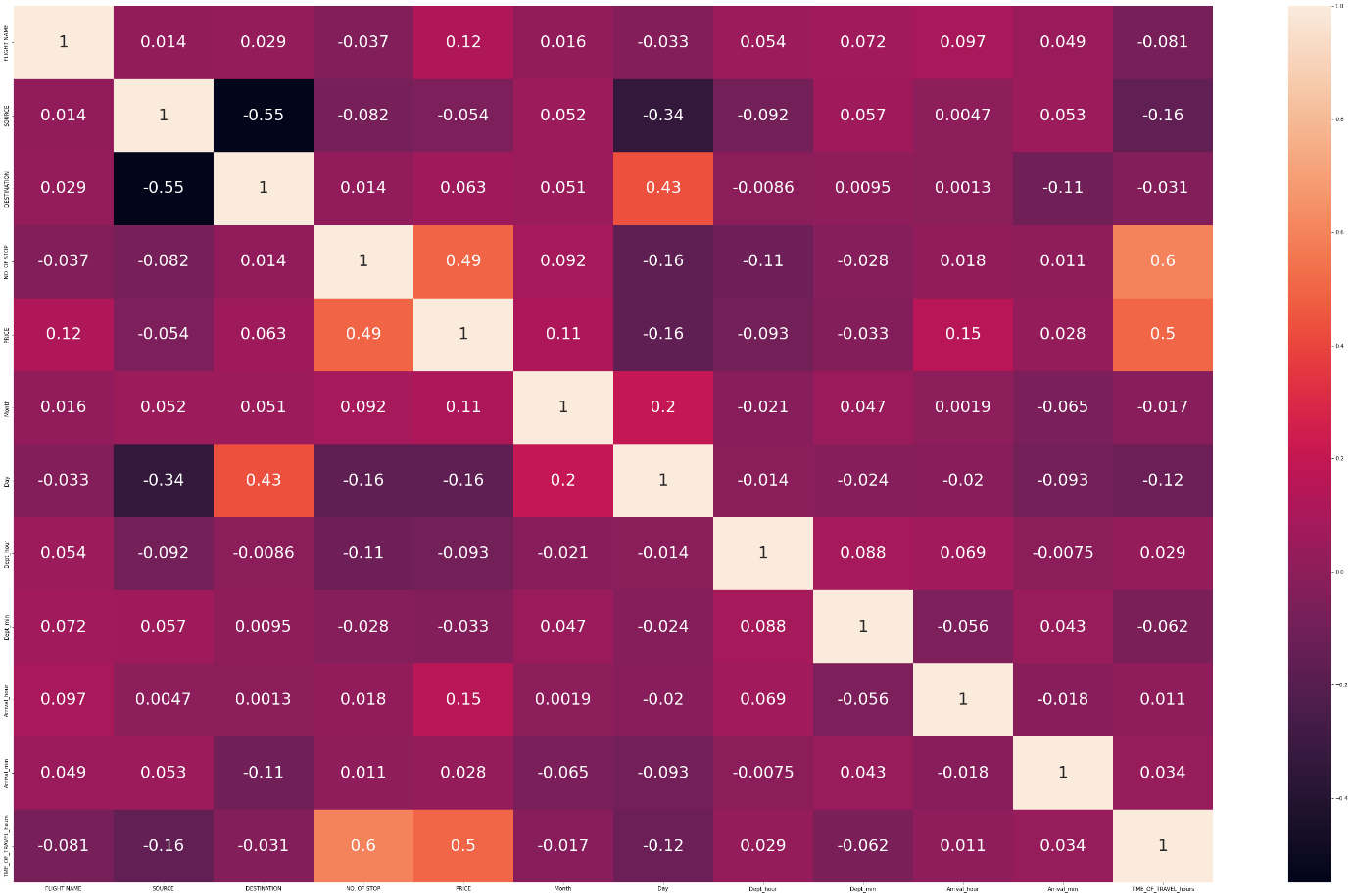


1. Displot and heat map has been used for the given model

Through displot we can analyse that features in the given data set is in normalized form and forming bell shape curved.



1. Through heat map it is observed that features are not tightly co related with each other and features are not closely related to label in the given data set.



* Interpretation of the Results

Considering the result of random forest regressor that is nearly 80 % . Applying hyper parameter tuning after random forest regressor,result remains same after tuning.

**CONCLUSION**

it can be concluded that through this project air ticket for any future date can be predicted and accordingly traveller can buy the ticket. it can be concluded that price of air ticket mainly vary due to two factors first how much time before ticket was booked, larger the time cheaper the ticket. Secondly from which flight passenger is traveling that is name of the flight.