

**PROJECT TITLE:- FLIGHT PRICE PREDICTION**

****

**SUBMITTED BY:-**

**VISHNU RATHORE**

**ACKNOWLEDGMENT**

I would like to express my special thanks of gratitude to “**Data trained**” that made this project possible. I would like thank to all **Data trained technical team** for his guidance in building this project. I would also like to thank Data Trained Institution to making me capable of making proper decisions in the field of damascene and Machine learning.

1. Objective

The objective of this article is to predict flight prices given the various parameters. Data used in this article is publicly available at Kaggle. This will be a regression problem since the target or dependent variable is the price (continuous numeric value).

## 2. Introduction

Airline companies use complex algorithms to calculate flight prices given various conditions present at that particular time. These methods take financial, marketing, and various social factors into account to predict flight prices.

Nowadays, the number of people using flights has increased significantly. It is difficult for airlines to maintain prices since prices change dynamically due to different conditions. That’s why we will try to use machine learning to solve this problem. This can help airlines by predicting what prices they can maintain. It can also help customers to predict future flight prices and plan their journey accordingly.

3. Data Used

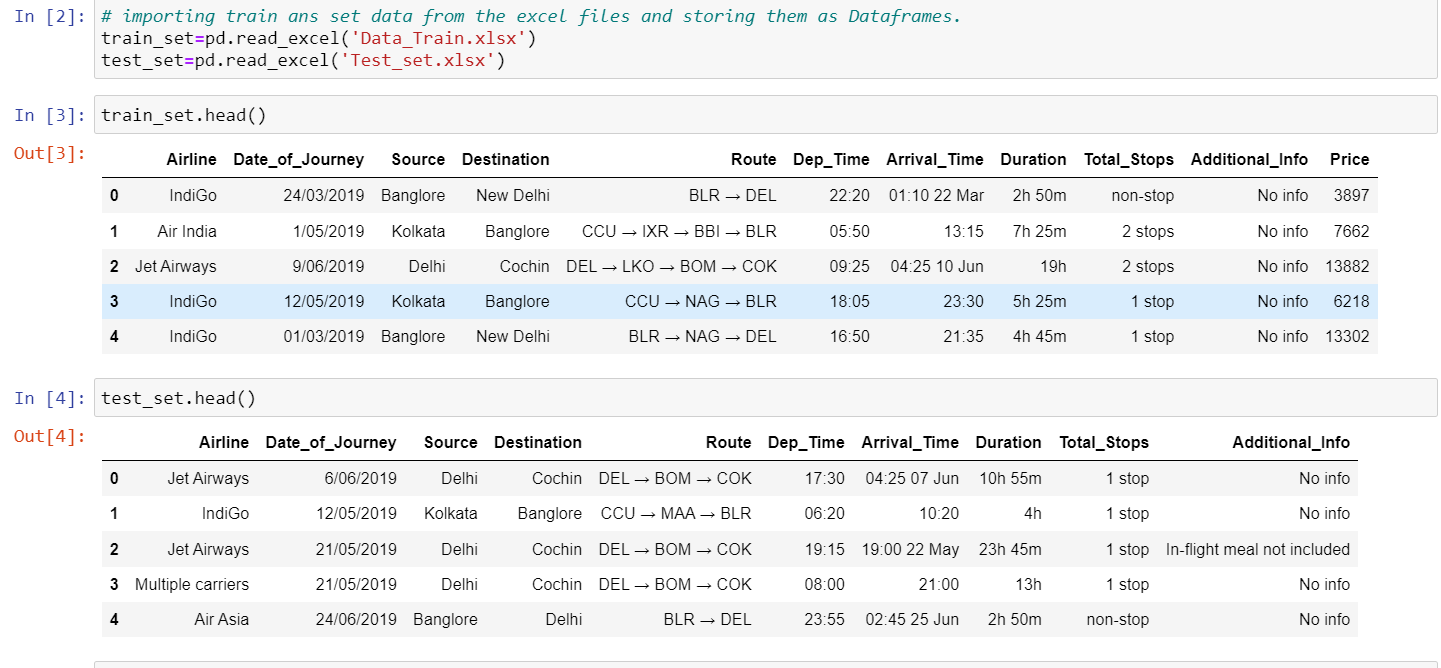
Data was used from Kaggle which is a freely available platform for data scientists and machine learning enthusiasts.

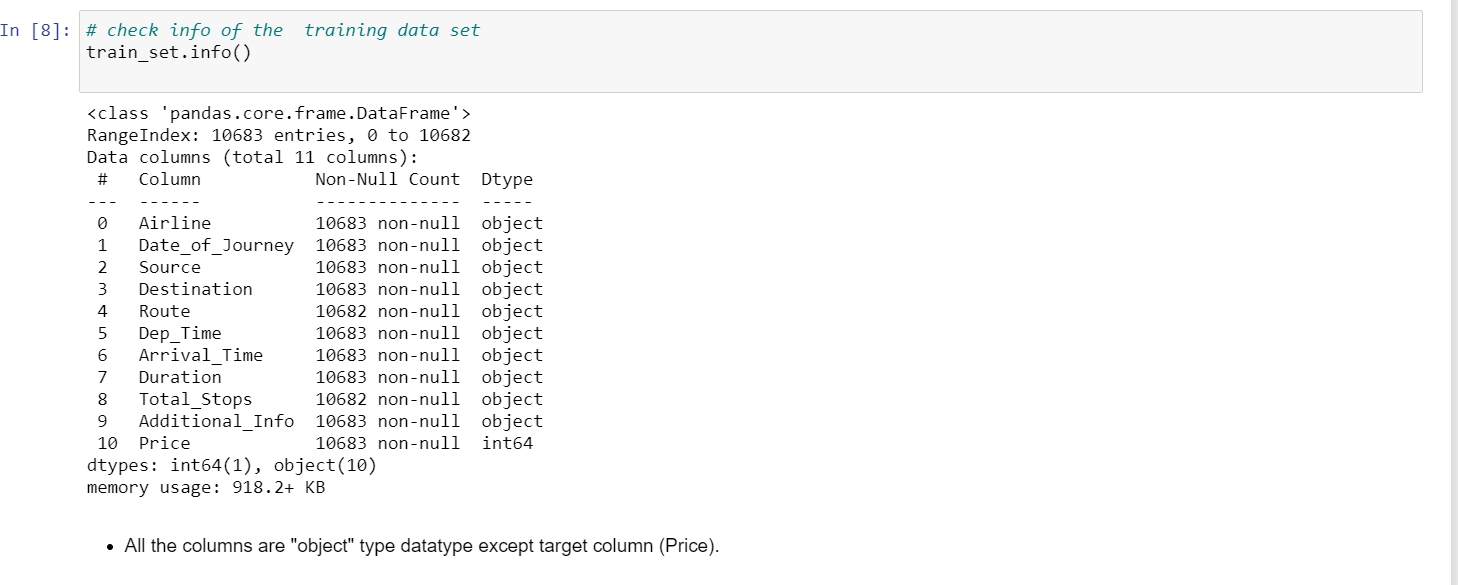
Source: <https://www.kaggle.com/nikhilmittal/flight-fare-prediction-mh>

We are using jupyter-notebook to run Flight Price Prediction task.

## 4. Data Analysis

The procedure of extracting information from given raw data is called data analysis. Here we will use **EDA** module of **DATA PREPROCESSING**  library to do this step.



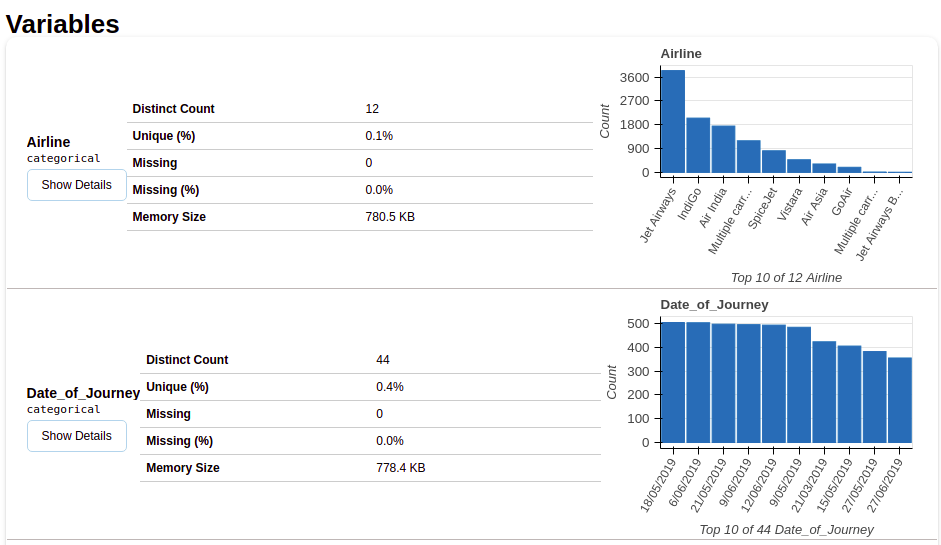


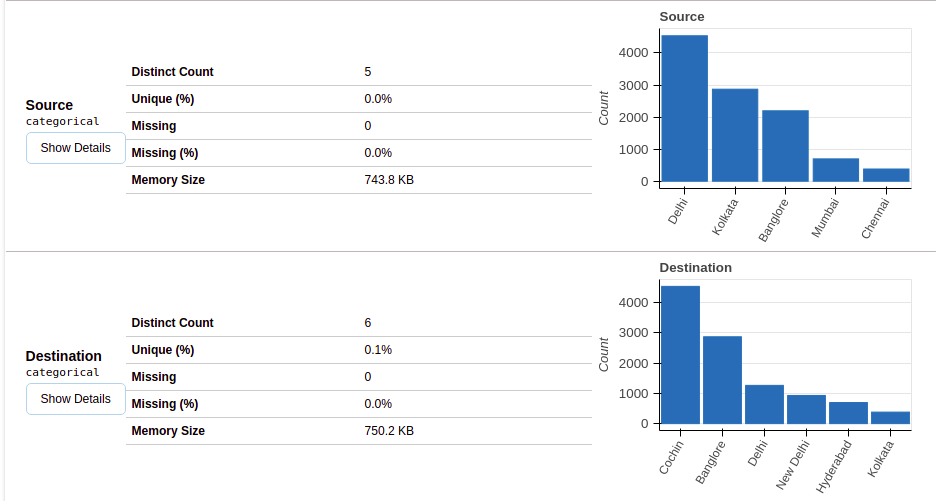
This report contains various sections or tabs.  ‘Overview’ section of this report provides us with all the basic information of the data we are using. For the current data we are using we got the following information:

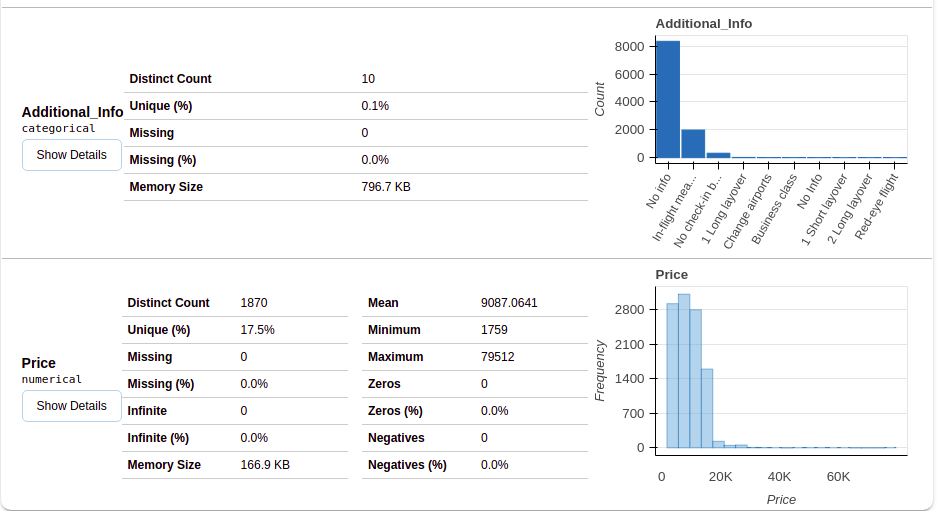
Number of variables = 11  
Number of rows = 10683  
Number of categorical type of feature = 10  
Number of numerical type of feature = 1  
Duplicate rows = 220

### Variables

After you select the variable section you will get information as shown in the below figures.







This section provides the type of each variable along with a detailed description of the variable.

**Conclusion Remarks from Exploratory Data Analysis**

From the data collected and through exploratory data analysis, we can determine the following:

● The trend of flight prices vary over various months and across the holiday.

● There are two groups of airlines: the economical group and the luxurious group. Spice jet,

AirAsia, IndiGo, Go Air are in the economical class, whereas Jet Airways and Air India in the

other. Vistara has a more spread out trend.

● The airfare increases during a holiday season. In our time period, during Diwali the fare

remained high for all the values of days to departure. We have considered holiday season as a

parameter which helped in increasing the accuracy.

● Airfare varies according to the day of the week of travel. It is higher for weekends and Monday

and slightly lower for the other days.

● There are a few times when an offer is run by an airline because of which the prices drop

suddenly. These are difficult to incorporate in our mathematical models, and hence lead to

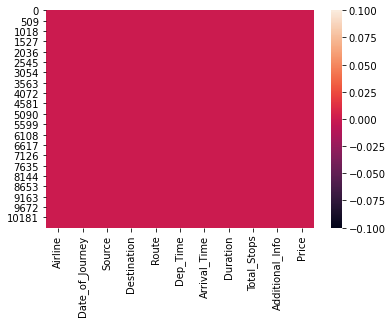
error.

● Along the business routes, we find that the price of flights increases or remains constant as the

days to departure decreases. This is because of the high frequency of the flights, high demand

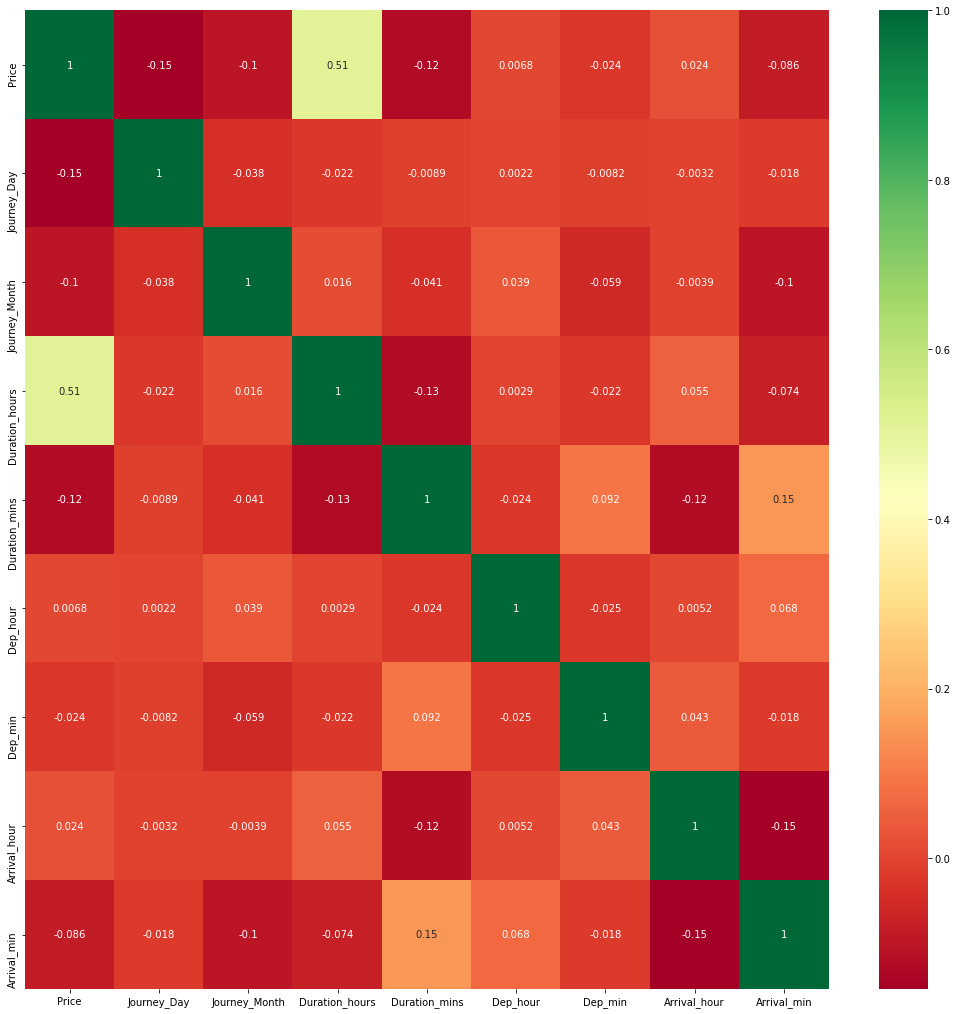
and also could be due to heavy competition

### Heat Map



The heat Map method shows variables having missing values in terms of correlation. Since ‘Route’ and ‘Total Stop’ both are highly correlated, they both have missing values.

# Correlation

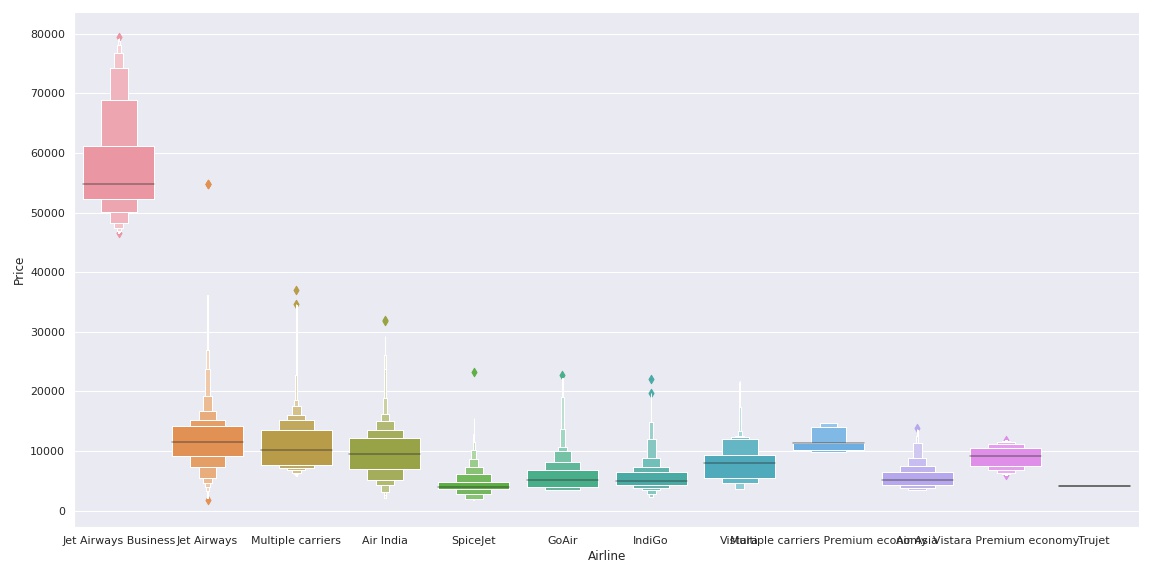


### Handling Categorical Data

Airline, Source, Destination, Route, Total Stops, Additional info are the categorical variables we have in our data. Let’s handle each one by one.

**Airline Variable**

Let’s see how the Airline variable is related to the Price variable.

****

As we can see the name of the airline matters. ‘Jet Airways Business’ has the highest price range. Other airlines price also varies.

Since the **Airline** variable is **Nominal Categorical Data** (There is no order of any kind in airline names) we will use **one-hot encoding** to handle this variable.

## Model Building

## 1.Split dataset into train and test set in order to prediction w.r.t X\_test

## 2.If needed do scaling of data

## 3.Import model

## 4.Fit the data

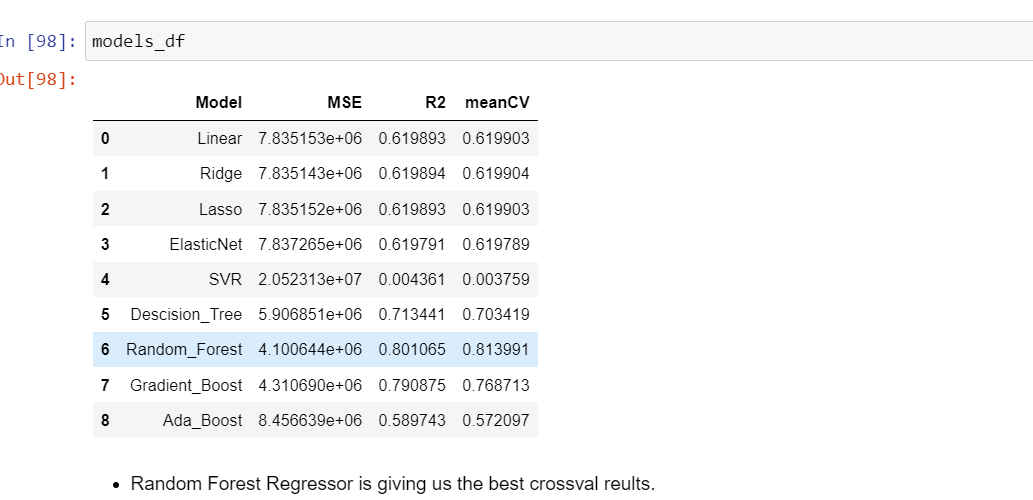
## 5.Predict w.r.t X test

## 6.In regression check RSME Score

## 7.Hyper Parameter Tuning for better score

## 









Conclusion

* Random forest is the best regression model.
* Model accuracy is 81%.
* this model helped me to understand Data  Analysis, Data Preparation, and Model building approaches in a much simpler way.

Reference:-

* Data collection from [www.kaggle.com](http://www.kaggle.com)
* <https://www.analyticsvidhya.com>
* Data trained LMS video.

Future Work

● More routes can be added and the same analysis can be expanded to major airports and travel routes in India.

● The analysis can be done by increasing the data points and increasing the historical data used. That will train the model better giving better accuracies and more savings.

● More rules can be added in the Rule based learning based on our understanding of the industry, also incorporating the offer periods given by the airlines.

● Developing a more user friendly interface for various routes giving more flexibility to the users.