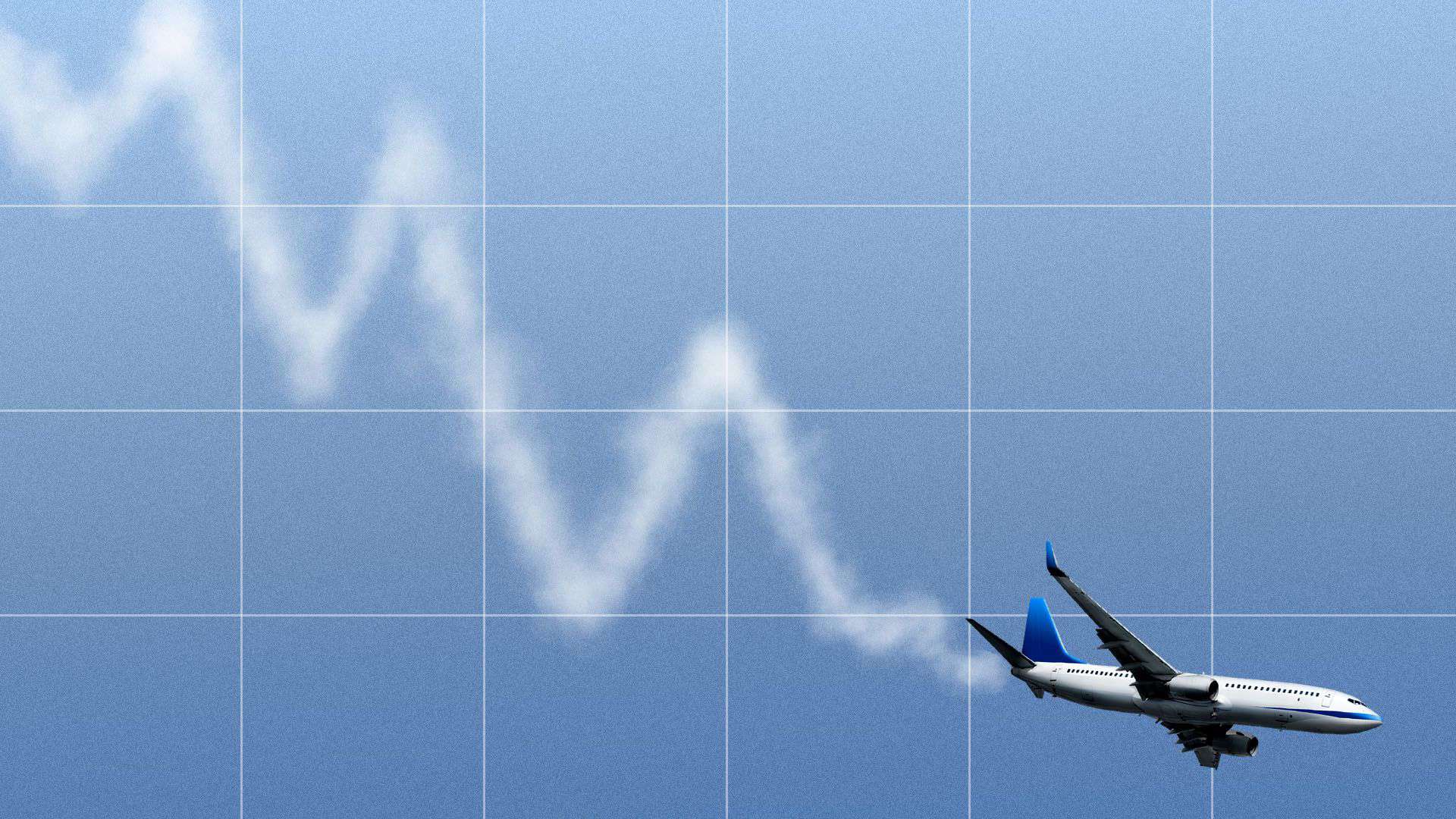
**Flight Price Prediction**

This is my first article regarding my data science project. The Article is about predicting the flight price.

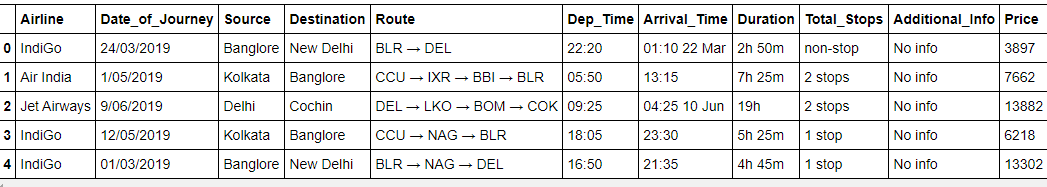
Problem Definition

Predicting the flight price nowadays is so hard. If you Google any flight price or if you check any airline's website to book a ticket, the first thing we will see the ticket price. Flight prices are unpredictable. There will be a lot of difference if you check the flight price today and tomorrow for the same flight and same route. As a data scientist, we know that with provided right data anything can be predicted. We will apply the same formula to predict the flight price.



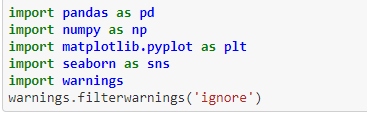
Data Description

We got two data sets, the train data set and the test data set. The train dataset contains the airline name, date of journey, source, destination, departure time, arrival time, duration of the journey, total stops, additional information, and most important thing price. The test data set contains all the information which is shown in the train data set except the price column (which should be predicted). With various machine learning algorithms, we will train the model to predict the flight prices for the given test data.



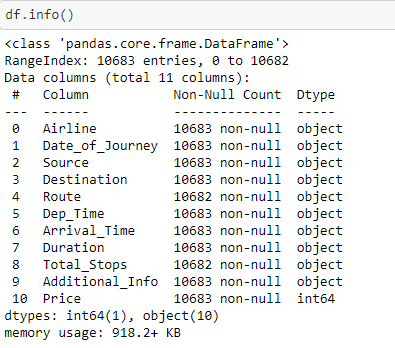
I used Python coding for this problem.

Import required libraries

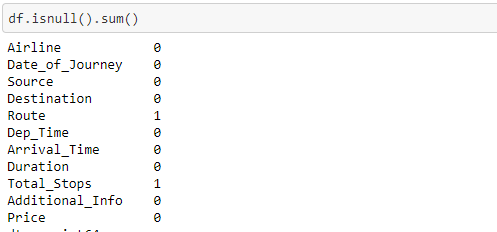


Checking data type and missing values

Checking the data type and null values in the given data set is very important to do data analysis and the EDA part.



Missing values:



We can see that in the columns Route and Total stops there is a single null value in each. Since it is a single null value without disturbing the given data it is better to drop the null values instead of replacing null values with other values.

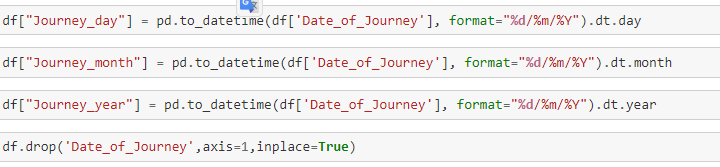
C:\Users\Naniii\Desktop\Capture 7.PNG

Feature Engineering

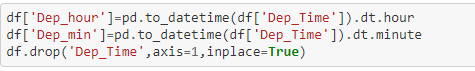
We can see that in the given data set the columns Date\_of\_Journey, Dep\_Time, Arrival\_Time, Duration are should be in exact time and date format, but with the df.info() it shows these columns are in object format.

So I want to extract the data as day, month, year, hour, and minute. It will be helpful for me to understand the data and to analyse and to make the model to learn the data perfectly. Here we use "pandas.to\_datetime()" which will help me to extract the data and to get separate columns for date, month , year, hour and minute.

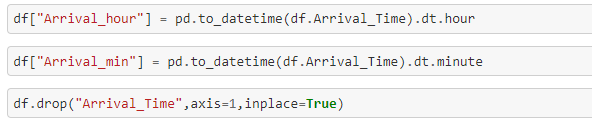
Date\_of\_Journey



Dep\_Time



Arrival\_Time

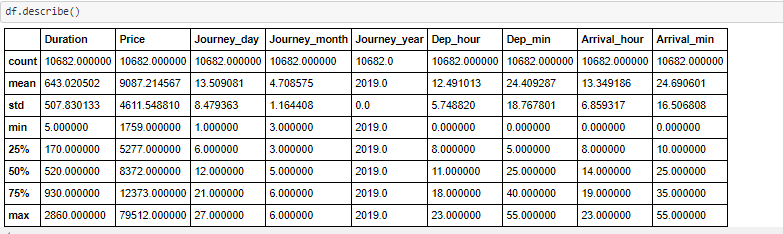


Duration

In this column, the data is in the format of 'h'(hour) and 'm'(minute), (eg:5h 25m). So I went on replacing the hours as minutes. So, that I can change the data to minutes instead of 'h' and 'm'.

C:\Users\Naniii\Desktop\Capture 14.PNG

Descriptive Statistics.



We can see that the column Journey\_Year contains only 2019. The whole data set is from 2019. So, this column is not more effective to predict. So we will remove this column. Also, the column Route is not in order and the column Route depends on the columns "Source, Destination, and Total Stops”. So, I want to remove the 'Route' column to remove the mess in the data.

C:\Users\Naniii\Desktop\Capture 21.PNG

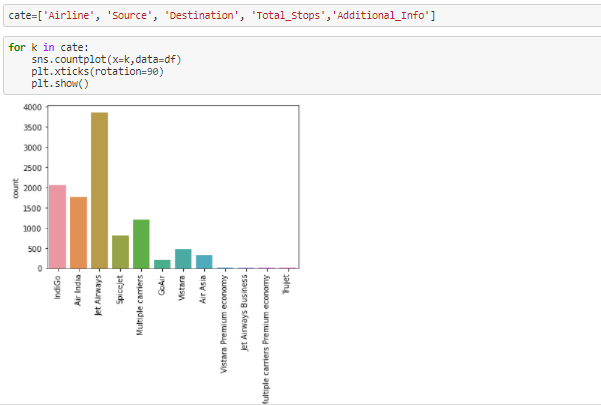
And

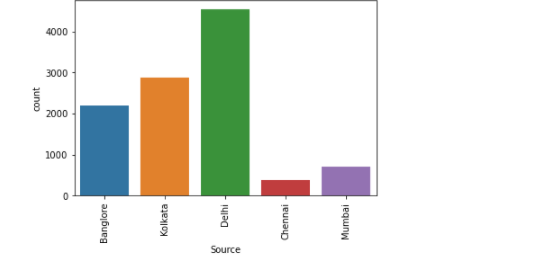
C:\Users\Naniii\Desktop\Capture 24.PNG

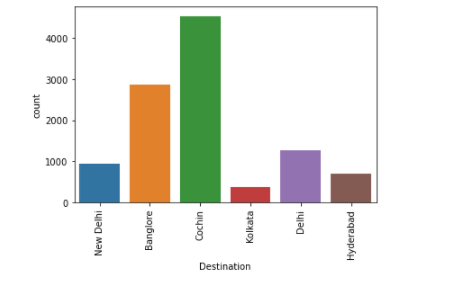
EDA

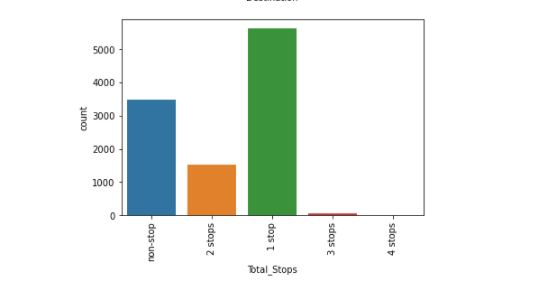
Analysis of categorical data.

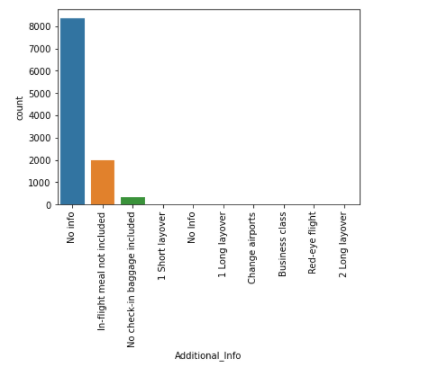
Here I separated the columns which have categorical data for analysis purpose. I used for loop to get all the categorical analysis at one go.











Observations from the above count plots:

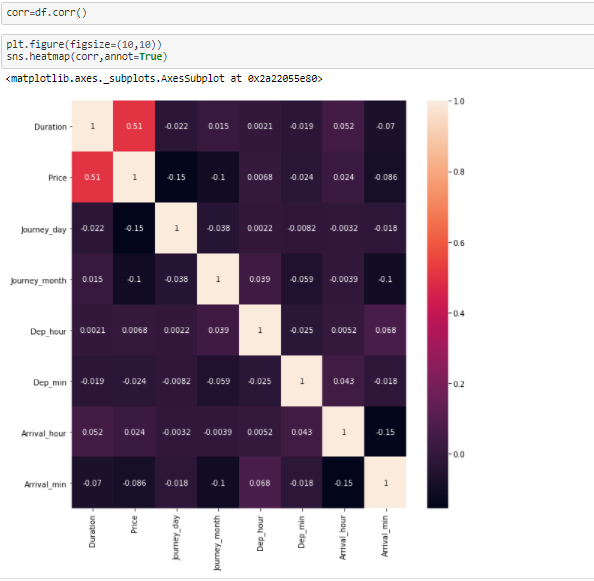
In the given data set the more information is from:

* Airlines: Jet Airways
* Source: Delhi
* Destination: Cochin
* Total\_stops: 1\_stop

Also we can see that  the column Additional\_Info has 80% of data with ‘No\_info’. If there is 80% no information in the column, I think it is not useful. So, I want to remove this column.

C:\Users\Naniii\Desktop\Capture 36.PNG

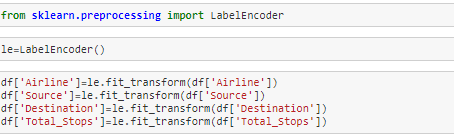
Numerical Analysis with the help of correlation matrix and Heatmap



From the above heat map, we can see that Duration is highly positively correlated with flight prices, where journey day and month are negatively correlated.

Prepare categorical Variables for model building:

We need numerical data to train the model. So we will do an encoding process to change the categorical data into numeric data.



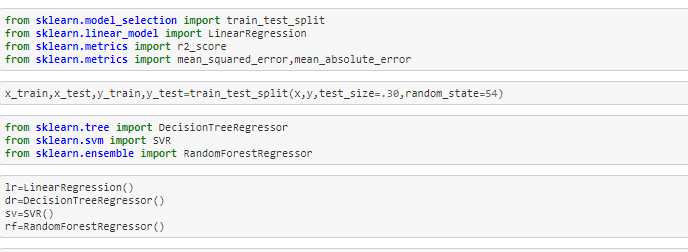
Splitting independent and dependent variables.

C:\Users\Naniii\Desktop\Capture 39.PNG

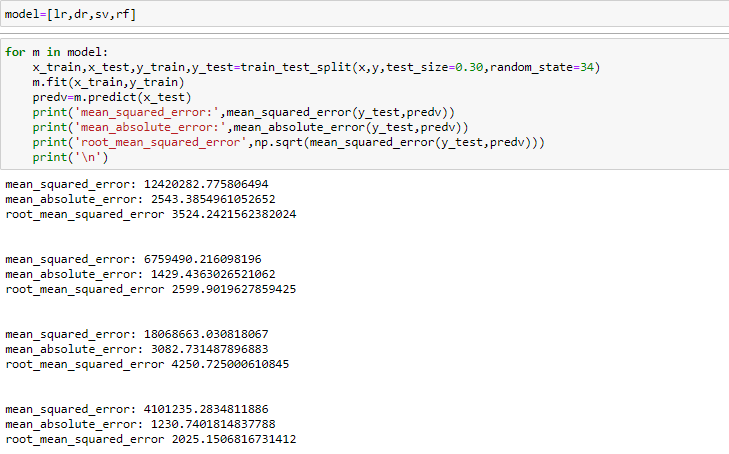
 Building Machine Learning Models.

The goal in this step is to train different machine learning algorithms and check which algorithm is performing good. Since the dependent variable is continuous we will use regression models. The final goal is to get less RMSE(Root Mean Squared Error). The less RMSE (fit the model) the better the model.

Required Machine Learning Libraries



With the for loop we will train all these 4 models.



We can see that the RandomForestRegressor is providing less 'RMSE'. So we will do hyper parameter tuning for better results.

Hyper Parameter Tunning.



We got less 'RMSE' hyper parameter tuning. Also, the r2\_score is 94%. So, this is our best model.

Save this model for future flight price prediction:

We can save the model either with pickle or joblib. Here I used joblib to save the model.

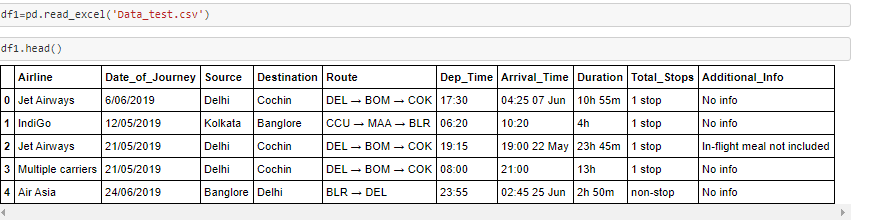
Save the model:



Now the task is to predict the flight price with the given test data set of 2671 rows. Do the whole EDA process which we did for the train data set.

# Test data.

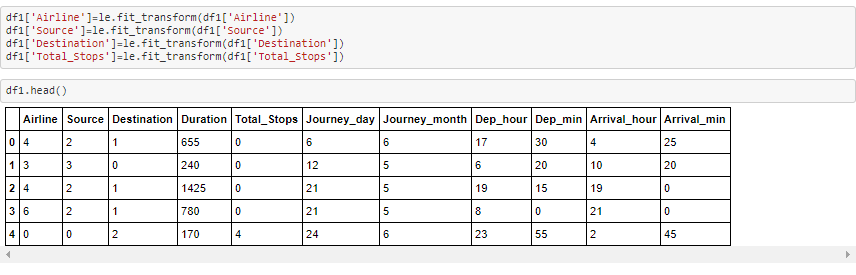
We loaded train dataset as "df", now we will load the test data set as "df1"



Feature Engineering:

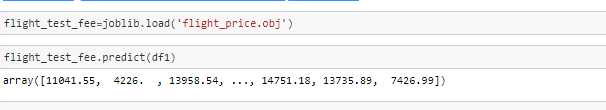


Label Encoding:



Now we will load the best saved model to pass this test data to predict flight prices.

# Loading the saved model to predict.



We got the predicted prices for the given test data.

Conclusion.

In this type of problem Feature Engineering is most crucial. You can see how we have handled the categorical and numerical data and also how we build different Machine learning models on the same dataset. We also check the RMSE score of each model so that we can understand how it should perform in our test dataset. At last, You can also further improve the Model by tuning different parameters which are being used in the model. We also learn how to pass the test data to the best-saved model to predict flight prices. For beginners, this article will help to understand how to train the model and how to pass the test data to the best save model in step by step process.

Final Words.

Any reviews or feedback regarding article and way of problem solving is greatly appreciated, as a data science student it will help me to learn more. You can send your feedback at manjunathreddy.in@gmail.com.