FLIGHT PRICE PREDICTION

Submitted by:

AVINASH PATEL

ACKNOWLEDGMENT

I would like to express my special thanks of gratitude to my teacher as well as my intern company FLIPROBO who gave me this golden opportunity to do this wonderful project on the topic FLIGHT PRICE PREDICTION, which also helped me in doing a lot of Research and I came to know about so many new things I am really thankful to them. Secondly I want to mention some sites which helped me when I got stuck somewhere while completing my projects.

Those sites are :

http://scikit-learn.org

https://www.w3schools.com

https://www.youtube.com

https://www.kaggle.com

INTRODUCTION

- Business Problem Framing

  Our client does not provide us with data. So we have to scrape data from different websites and we have to scrape more than 1500 rows of data.

  So, In order to improve the prediction of price, the client wants some predictions that could help them in further investment and improvement in purchase of flight tickets.

- Conceptual Background of the Domain Problem

  I think we have to get some knowledge on how the price of the tickets gets affected according to other columns and then we have to scrape those data also, to which extent the price gets increased or decreased according to the increase and decrease of other attributes and what are they.

- Review of Literature

  Anyone who has booked a flight ticket knows how unexpectedly the prices vary. The cheapest
  available ticket on a given flight gets more and less expensive over time. This usually happens as
  an attempt to maximize revenue based on -
  1. Time of purchase patterns (making sure last-minute purchases are expensive)
  2. Keeping the flight as full as they want it (raising prices on a flight which is filling up in order
  to reduce sales and hold back inventory for those expensive last-minute expensive purchases)

- Motivation for the Problem Undertaken

  As I am an intern I want to work on as many project as possible for me. So ,I am highly motivated to do this project and to learn new thing and also learn from my mistakes.

Analytical Problem Framing

- Mathematical/ Analytical Modelling of the Problem

  First of all I checked that how many of the column are categorical columns and how many of them are numerical columns. By doing so I find out that all the columns are categorical as all the columns contain some character which makes them categorical.

  Now I started to clean the column individually.

  Here I have applied many techniques like split , replace etc to remove unwanted characters from columns which are supposed to be numerical column.

- Data Sources and their formats

  As no Data is provided by our client, I scraped 2214 rows of data with 9 columns of each. I scraped data from yatra.com

- Data Pre-processing Done

  For Treating Date column: This column has date which is separated by commas so I am going to separate this column into three separate columns named day, month and Dates.

  For Stops column: This column has stops denoted as numbers of stops and for zero stops its denoted as no stops. So, I converted all the stops to there corresponding numbers and for no stops I put 0 as a value.

  Treating Departure time and Arrival Time column: As these are constituted of timing for 24-hour clock. So, I converted these time spans to morning, night, afternoon and evening. I separated from

hour 0 to 6 as night,6 to 12 morning,12-18 afternoon,18-24 evening.

Treating Price column: This column is good but with a single problem that is it contains a rupee symbol Infront of each and every amount so I removed all the rupee symbol and replaced with only the amount.

- Data Inputs- Logic- Output Relationships

  The Remaining Data after all the pre-processing acts as an input data for the Model(in our case Linear Regressor) the already trained model gives us an output of an estimated price.

- Hardware and Software Requirements and Tools Used

  Ram:8GB
  ROM:200MB
  Processor : Ryzen 5 3600X(6 cores 12 threads)
  Tool : Jupiter Notebook
  Language Used: Python

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  First I imported all the basic libraries of python used in the problem solving. Then I read carefully the data and the description given to me. After that I imported the provided csv file into the Jupiter notebook. Now first I checked for any nan values in the dataset after that I checked how many number of rows and columns I am dealing with and I also done some more analysis of the raw data to get a general idea of there max and mins and there inter quartile ranges etc.

  After analysing the data I started cleaning the data as explained above.

  After cleaning I did some visualisation of the data to get some insights from our cleaned data.

  I removed skewness and some excess outliers. On the completion of all these steps I moved on to the model selection part. Here I tested many models and choose one of the model to proceed.

  After I selected one of the models I did hyperparameter tuning and finally I saved my model.

- Testing of Identified Approaches (Algorithms)
  - LinearRegression
  - DecisionTreeRegressor
  - Ridge
  - Lasso
  - RandomForestRegressor
  - KNeighborsRegressor
  - XGBRegressor

- Run and Evaluate selected models

  After running all the models the lowest difference between the accuracy score and the cross validation is for XGBOOST Regressor.

```
LR=LinearRegression()
regress(LR,x,y)

R2: 38.059615695765125
CV Score: 38.86934109722133
Diffrence: 0.8097254014562054
```

```
DTR = DecisionTreeRegressor()
regress(DTR,x,y)

R2: 24.02438111398154
CV Score: 48.20265126754023
Diffrence: 24.178270153558692
```

```
RDG=Ridge()
regress(RDG,x,y)

R2: 37.994191425961766
CV Score: 38.88781653376859
Diffrence: 0.8936251078068267
```

```
LR=LinearRegression()
regress(LR,x,y)

R2: 38.059615695765125
CV Score: 38.86934109722133
Diffrence: 0.8097254014562054
```

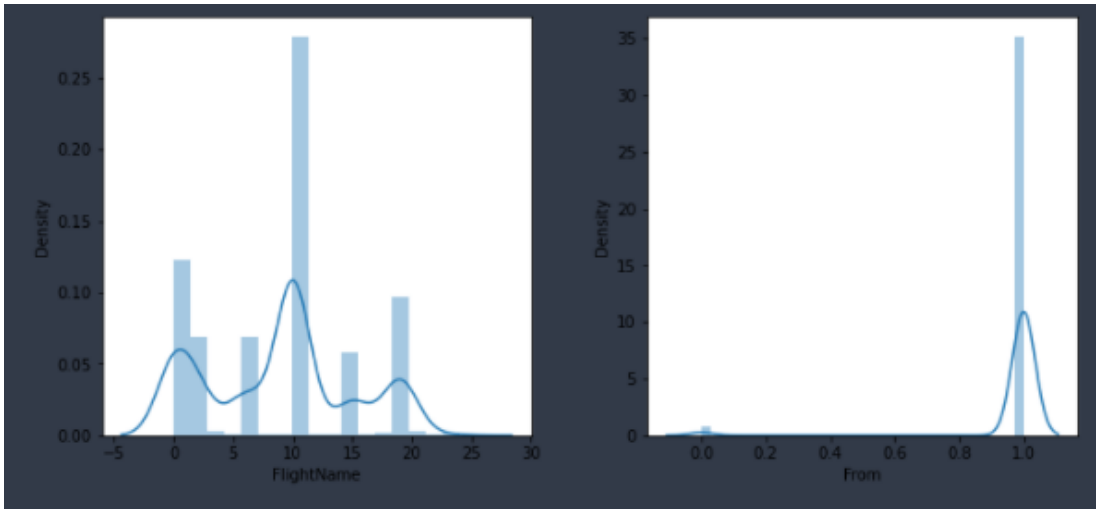
```
RFR=RandomForestRegressor()
regress(RFR,x,y)

R2: 50.17956396734784
CV Score: 67.32845123349266
Diffrence: 17.14888726614482
```

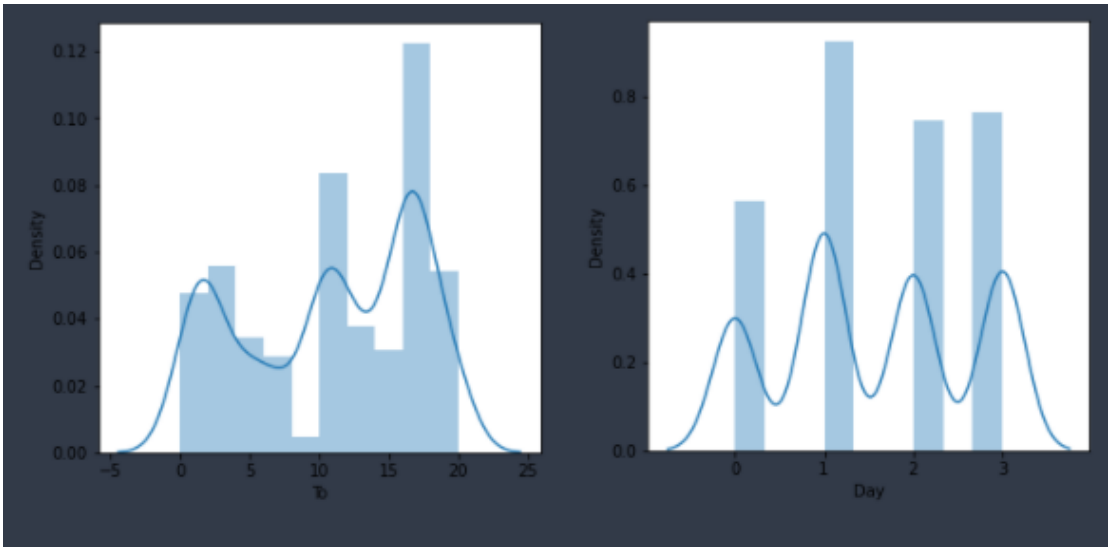
```
KNR=KNeighborsRegressor()
regress(KNR,x,y)

R2: 29.895815814260818
CV Score: 36.78807261685565
Diffrence: 6.892256802594833
```

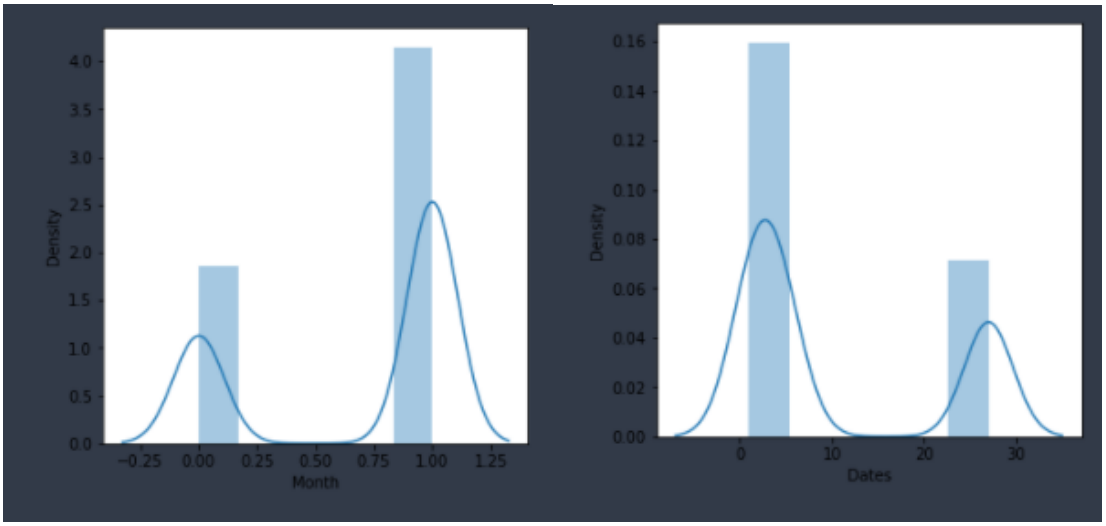
That means Linear Regressor is not overfitted nor underfitted.

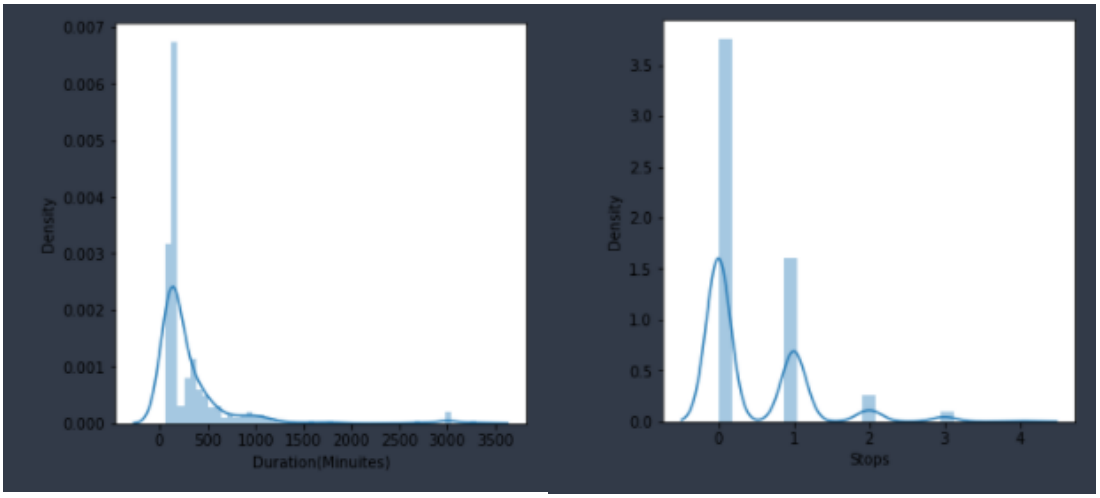
- Visualizations

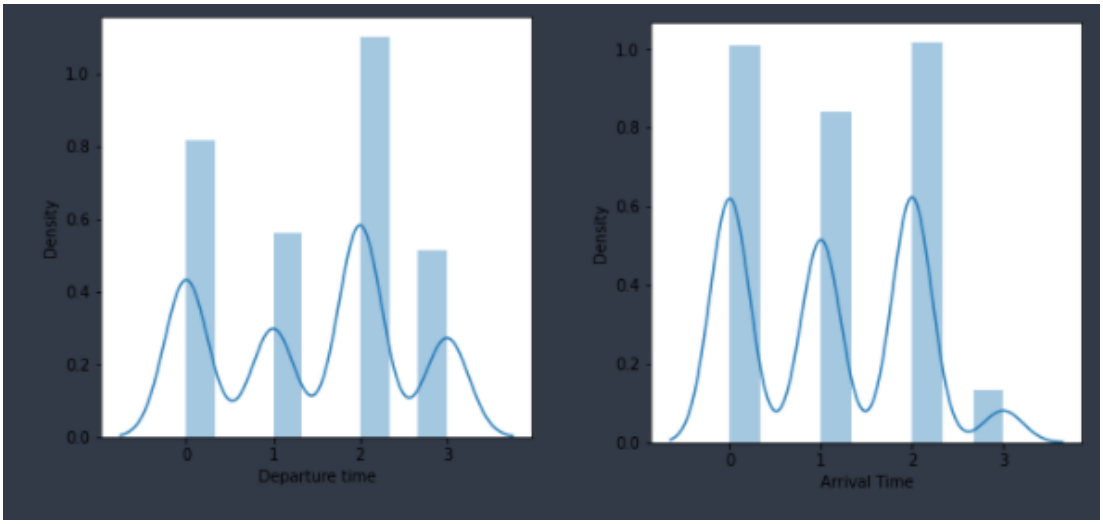
Distplot:

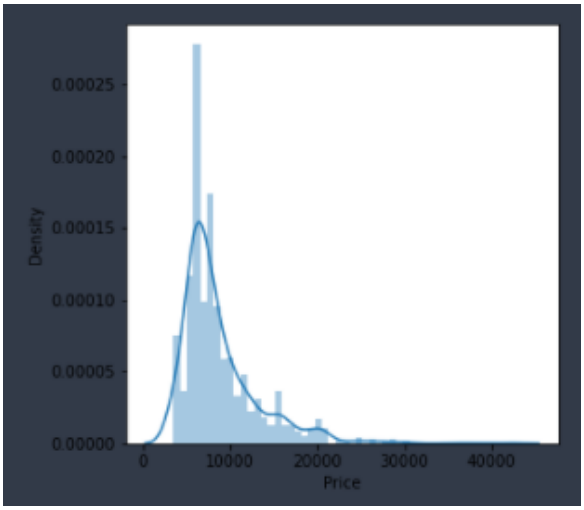



Density
To
Day

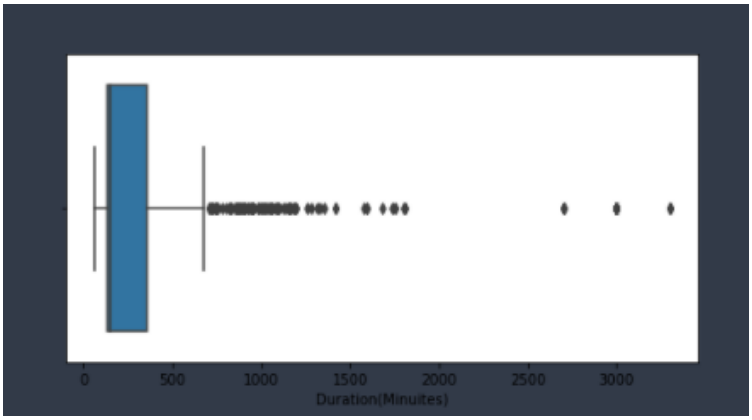

Density
Month
Dates

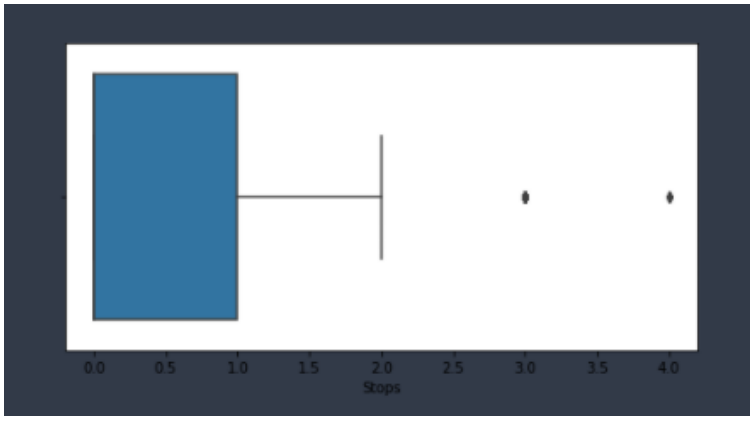

Density
Duration(Minutes)
Stops

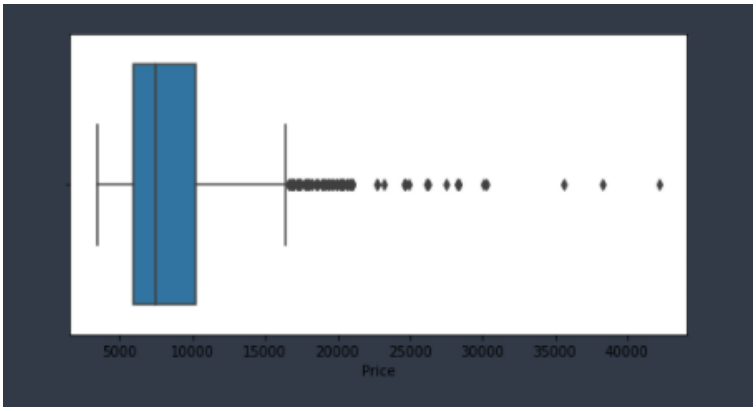

Density
Departure time
Density
Arrival Time


Density
Price

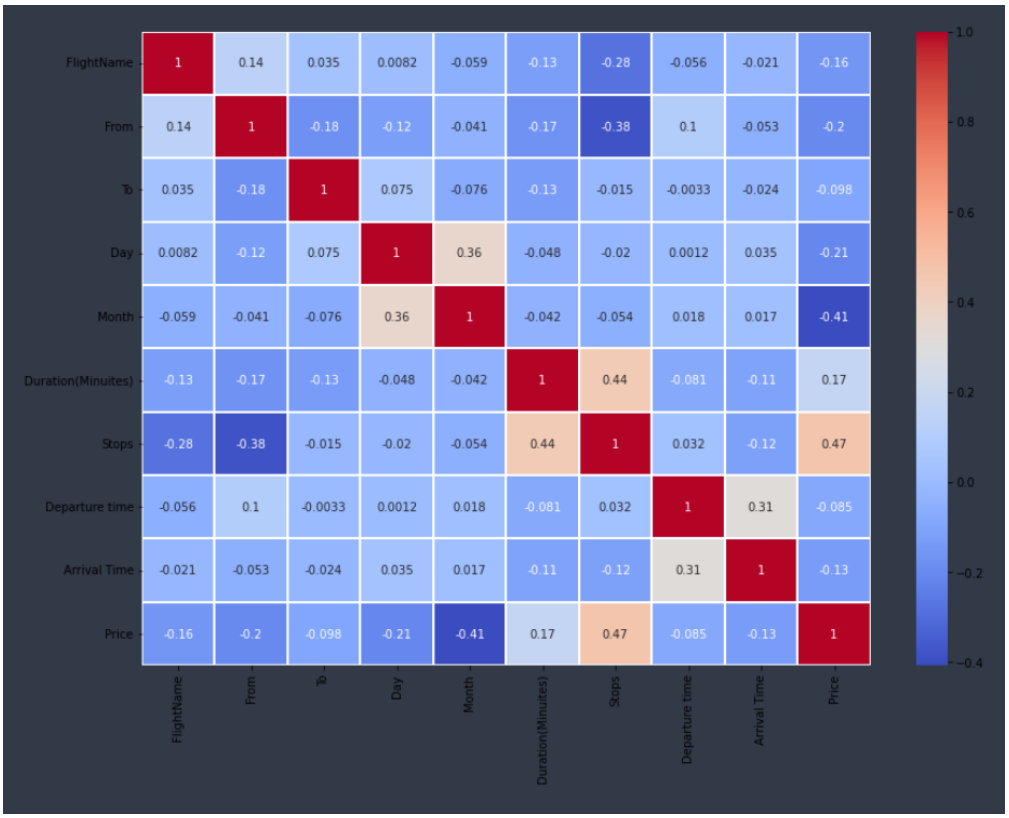
Boxplot:


Duration(Minutes)


0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Stops


5000
10000
15000
20000
25000
30000
35000
40000
Price

Heatmap:


FlightName
From
To
Day
Month
Duration(Minuites)
Stops
Departure time
Arrival Time
Price
1 0.14 0.035 0.0082 -0.059 -0.13 -0.28 -0.056 -0.021 -0.16
0.14 1 -0.18 -0.12 -0.041 -0.17 -0.38 0.1 -0.053 -0.2
0.035 -0.18 1 0.075 -0.076 -0.13 -0.015 -0.0033 -0.024 -0.098
0.0082 -0.12 0.075 1 0.36 -0.048 -0.02 0.0012 0.035 -0.21
-0.059 -0.041 -0.076 0.36 1 -0.042 -0.054 0.018 0.017 -0.41
-0.13 -0.17 -0.13 -0.048 -0.042 1 0.44 -0.081 -0.11 0.17
-0.28 -0.38 -0.015 -0.02 -0.054 0.44 1 0.032 -0.12 0.47
-0.056 0.1 -0.0033 0.0012 0.018 -0.081 0.032 1 0.31 -0.085
-0.021 -0.053 -0.024 0.035 0.017 -0.11 -0.12 0.31 1 -0.13
-0.16 -0.2 -0.098 -0.21 -0.41 0.17 0.47 -0.085 -0.13 1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

- Interpretation of the Results

  After all the pre processing and then visualising the data I find out that although there are many columns to the dataset , many column are useless and the remaining columns do not have a significant correlation with our Target column except Stops and Months column.

CONCLUSION

- Key Findings and Conclusions of the Study

  I find out that few columns are correlated to our target variable that means predicting our target variable is difficult for any model because there is no direct correlation between them. Also, many of the columns have values which are unrealistic. Many of the columns have highly skewed values and many columns also have high amounts of outlies in them.

- Learning Outcomes of the Study in respect of Data Science

  As I have already told that this dataset has a lot of outliers and I have to look into each and every column to get some of the data cleaned. As per the visualisation part I plotted boxplots to see the number of outliers and in which column they exist and in how much amount.

  I have also plotted Distplot to get a general idea of the columns which have skewness to further treat them.

  I plotted heatmap to see which columns are in a correlation with our target variable.

- Limitations of this work and Scope for Future Work

  As I am not able to fully understand the meaning of all the columns and if I could get someone who is already working in the flight or travel company then I could get a better idea of how all the column are working and which outer columns needs to be in the dataset. As per the future scope, yes there is definitely more room to grow as I could get the model efficiency to 38% it could be more on further research.