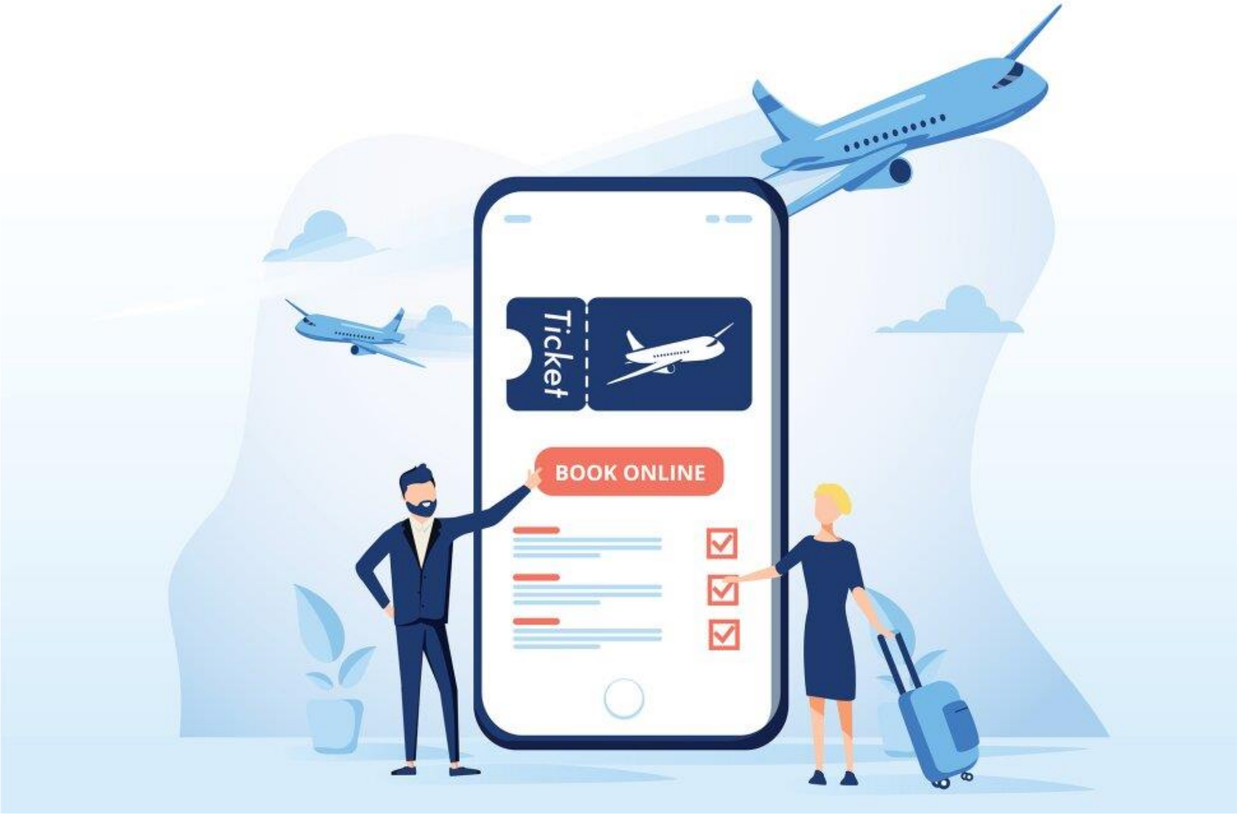
FLIP ROBO



FLIGHT PRICE PREDICTION PROJECT

Submitted by:

JASMINE KAUR

ACKNOWLEDGMENT

This project was my first experience in terms of collecting the data from a site where prices change every day and you have to work very diligently in order to get the most out of it. It was a fun activity to build a price prediction model on the data collected by my own efforts. I applied all machine learning skills and techniques learned at datatrained. Also, the continuous guidance provided by the flip robo technologies team was very helpful in the process of this project. I would like to say special thanks to my SME, Ms Khushboo Garg, whose constant help make this project possible.

INTRODUCTION

- Business Problem Framing

  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) So, you have to work on a project where you collect data of flight fares with other features and work to make a model to predict fares of flights.

- Motivation for the Problem Undertaken

  Collecting the data is one of the most important task in the process of building a project that predicts the price of flights. There are various data sources that provides details of different routes, times, etc. on a daily basis. But the ability to bring those resources together and then to build a project that has the ability to predict the fares was exciting as well as challenging at the same time.

  Due to its everyday usage we were motivated to develop such a project that can help us predict the flight prices that changes very often. We consider different parameters like date, duration , location etc. to build this project.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

There are 3 phases of this project:

1. *Data collection*: We scraped data using selenium web scraping technique. We scraped airline name, date of journey, source, destination, departure time, arrival time, duration, total stops and price from the yatra.com website. We computed the data in a tabular form to make further analysis.
2. *Data analysis*: After collecting the data, we have cleaned the data. In this process we checked for null values, skewed data, outliers. We have performed feature engineering in order to get most out of the data. We have checked for multicollinearity to make sure that there is no relationship between the independent variables. We performed EDA to get useful insights from the data. Lastly, we encoded the categorical data into numerical data using label encoder.
3. *Model building:* After analysing the data, we have applied four regression models and then applied cross validation to check for multicollinearity. The best model comes out to be KNearest Neighbors Regression Model.

- Data Sources and their formats

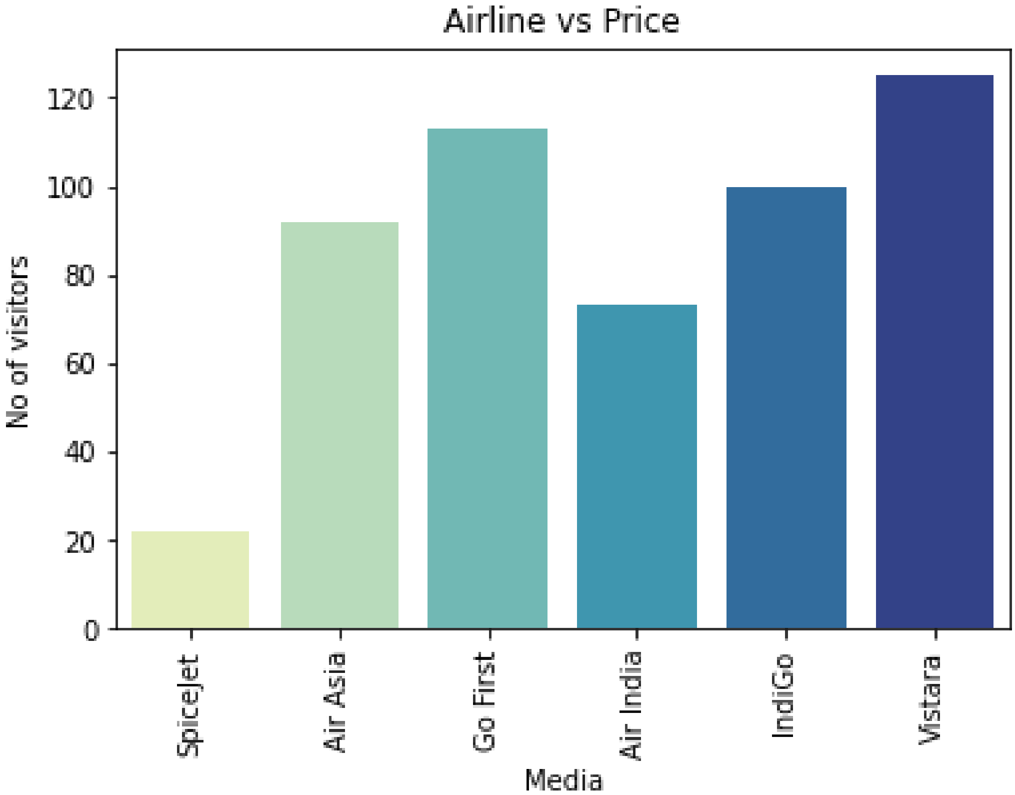
We have collected data from yatra.com using selenium web scraping technique.

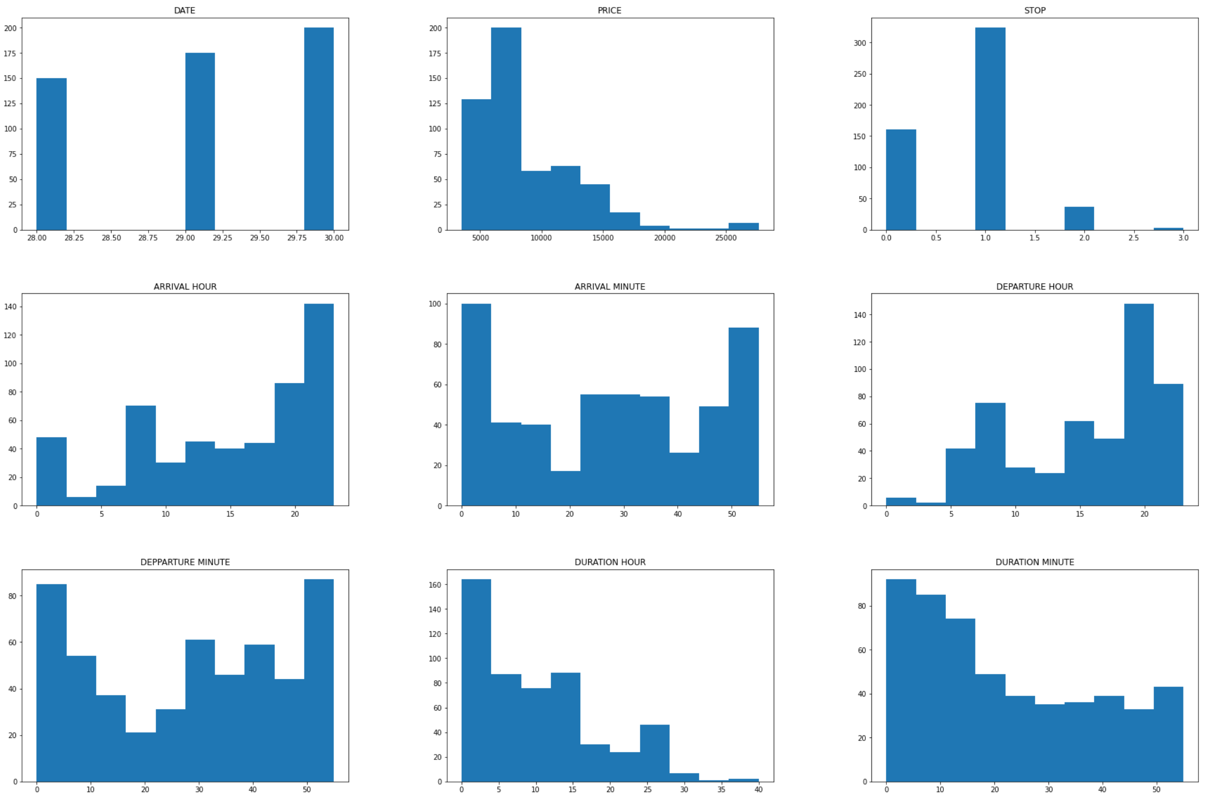
- Data Preprocessing Done

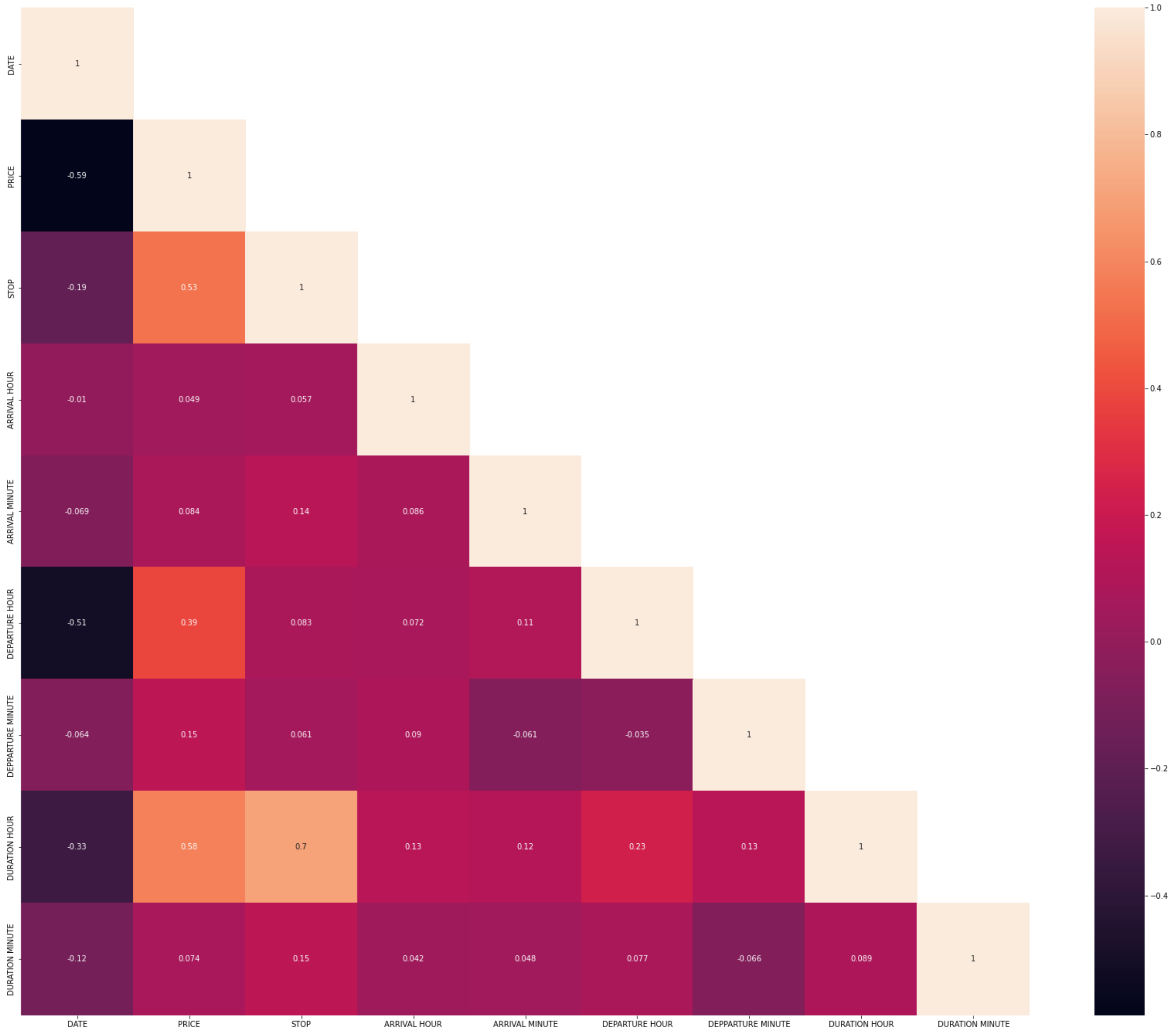
1. NULL VALUES: There were no null values in the data.
2. Converted the object data type into the integer data type using label encoder.
3. Removed unnecessary symbols or units from the data.
4. Checked skewness and outliers and dealt with it.
5. There was no multicollinearity as VIF was less 5 for all the variables.

- Data Inputs- Logic- Output Relationships

  The output variable is the price of the flight. Since the price is a continuous, thus is a regression problem. The input variables are airline name, source, destination, departure time, arrival time, data, month, year, duration, total stops. These input features work as independent variables and help in predicting the price of the car.




DATE
0
25
50
75
100
125
150
175
200
28.00
28.25
28.50
28.75
29.00
29.25
29.50
29.75
30.00
PRICE
5000
10000
15000
20000
25000
STOP
0.0
0.5
1.0
1.5
2.0
2.5
3.0
300
250
ARRIVAL HOUR
ARRIVAL MINUTE
DEPARTURE HOUR
DEPPARTURE MINUTE
DURATION HOUR
DURATION MINUTE
20
40
60
80
100
120
140
160
0
5
10
15
20
25
30
35
40
50



- Hardware and Software Requirements and Tools Used

The hardware and software requirements along with the tools, libraries and packages used are:

1. MS Excel

2. Google colab

3. Pandas, Numpy

4. Visualization tools

5. Github

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

1. Boxplot for summarizing variations and checking outliers
2. Histogram are used to check outliers
3. Correlation matrix was used to analyse the correlation between the features.
4. Standard scaler was used to standardize the data.
5. We have split the data using train test split.

- Testing of Identified Approaches (Algorithms)

The algorithms used for the training and testing are:

1. Random Forest Regressor
2. Linear Regression
3. Gradient Boosting Regressor
4. KNeaighbors Regressor

- Run and Evaluate selected models

MODELLING ALGORITHM	R2 SCORE
LINEAR REGRESSION	0.53
RANDOM FOREST REGRESSION	0.89
KNEAREST NEIGHBORS REGRESSION	0.52
GRADIENT BOOSTING REGRESSOR	0.85

CROSS VALIDATION

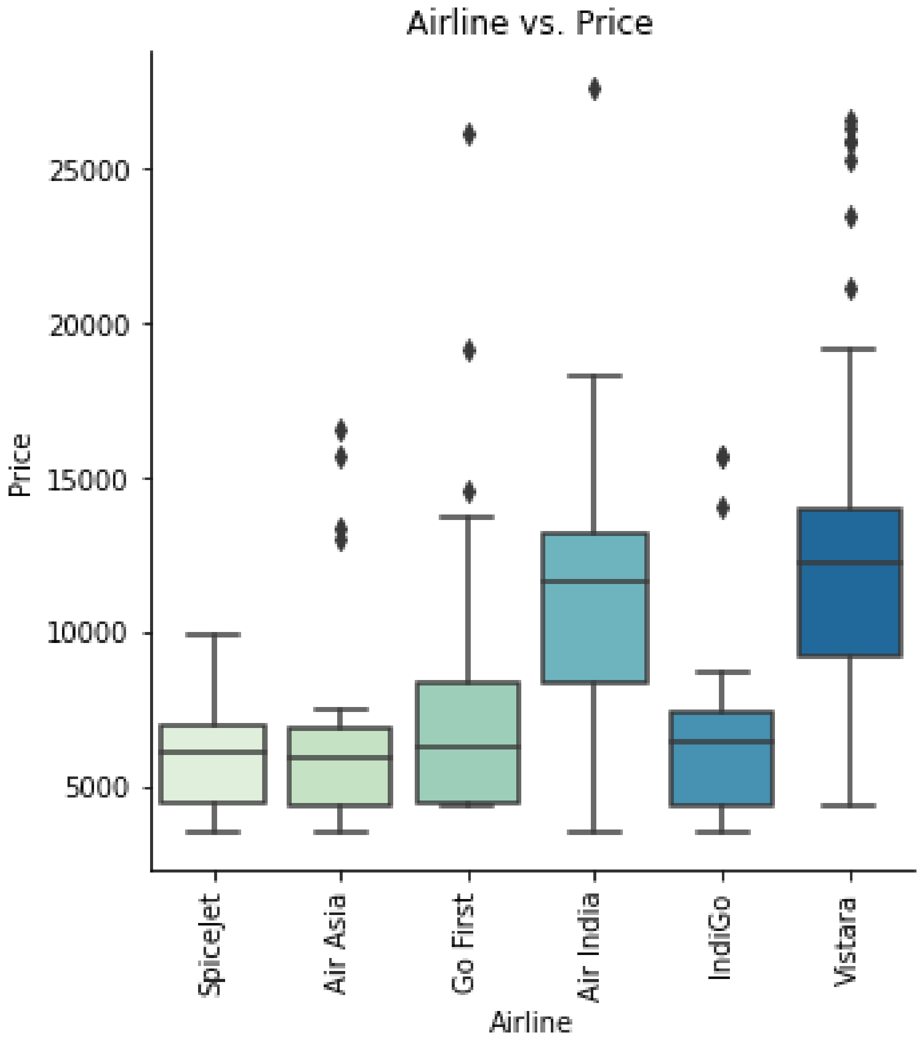
MODELLING ALGORITHM	CV SCORE	
LINEAR REGRESSION	0.40	
RANDOM FOREST REGRESSION	0.53	
KNEAREST NEIGHBORS	0.48	

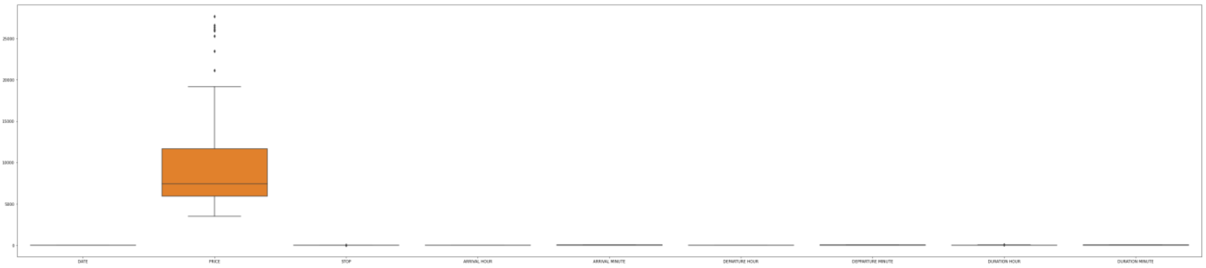
REGRESSION		
GRADIENT BOOSTING REGRESSOR	0.57	

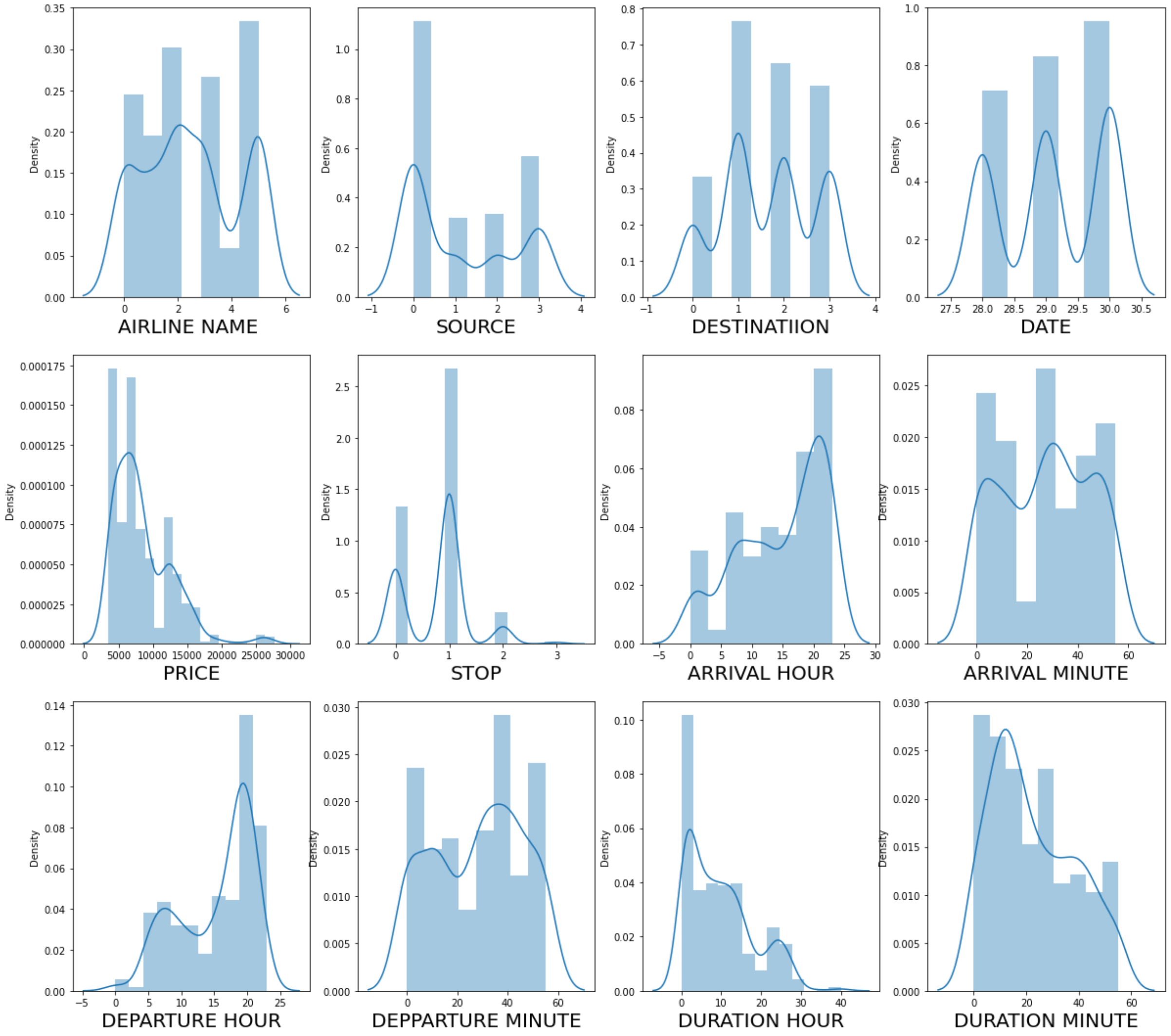
- Key Metrics for success in solving problem under consideration

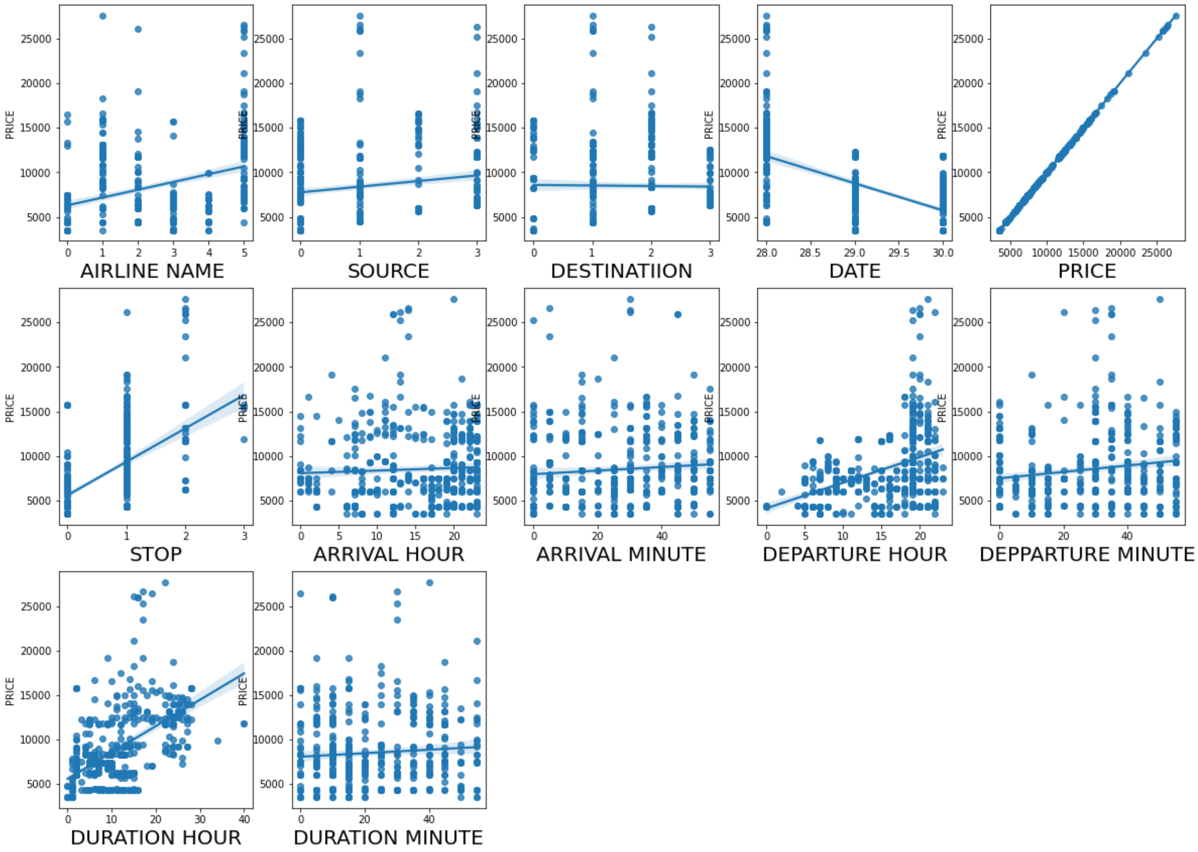
  The key metrics used is r2 score. R-squared is a statistical measure that represents the proportion of the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

- Visualizations




25000
20000
15000
10000
5000
0
DATE
PRICE
STOP
ARRIVAL HOUR
ARRIVAL MINUTE
DEPARTURE HOUR
DEPARTURE MINUTE
DURATION HOUR
DURATION MINUTE


Density
AIRLINE NAME
SOURCE
DESTINATIION
DATE
PRICE
STOP
ARRIVAL HOUR
ARRIVAL MINUTE
DEPARTURE HOUR
DEPPARTURE MINUTE
DURATION HOUR
DURATION MINUTE
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.000000
0.000025
0.000050
0.000075
0.000100
0.000125
0.000150
0.000175
0
5000
10000
15000
20000
25000
30000
27.5
28.0
28.5
29.0
29.5
30.0
30.5



- Interpretation of the Results

  After comparing the r2 score and the cross validation score, we have reached at a conclusion that the KNearest Neighbors Regressor is the best model. We have applied hyperparameter tuning and then we have saved the model.

CONCLUSION

- Key Findings and Conclusions of the Study

  We have got insights about collecting the data and then using to solve real life problems.

  Performing data cleaning, data pre-processing are difficult but interesting steps in the process if reaching the end goal.

Machine learning models help us in predicting the dependent variable with the help of independent variables.
We have finalised KNearest Neighbors as the best model and saved it in pkl format.

- ## Learning Outcomes of the Study in respect of Data Science

Collecting data is one of the most difficult steps. As it is also the most important step in building a model, we need to be very careful in scraping it.

Scraping data is a time consuming step, so we need to pay attention to every small detail.

Handling unclean data and cleaning the data is the base of any model. We should perform this task with care.

- ## Limitations of this work and Scope for Future Work

We could have scrape more data but due to limited time, we are able to scrap this much data only.