PRML Bonus Project
Flight Ticket Price Prediction
Final Summary Report

By Mitarth Arora (B20EE096)

- **Importing the dataset:**

a.) Using pd.read_csv() using sep=”,” and names of column as per given in the database

- **Pre-Processing and Cleaning:**

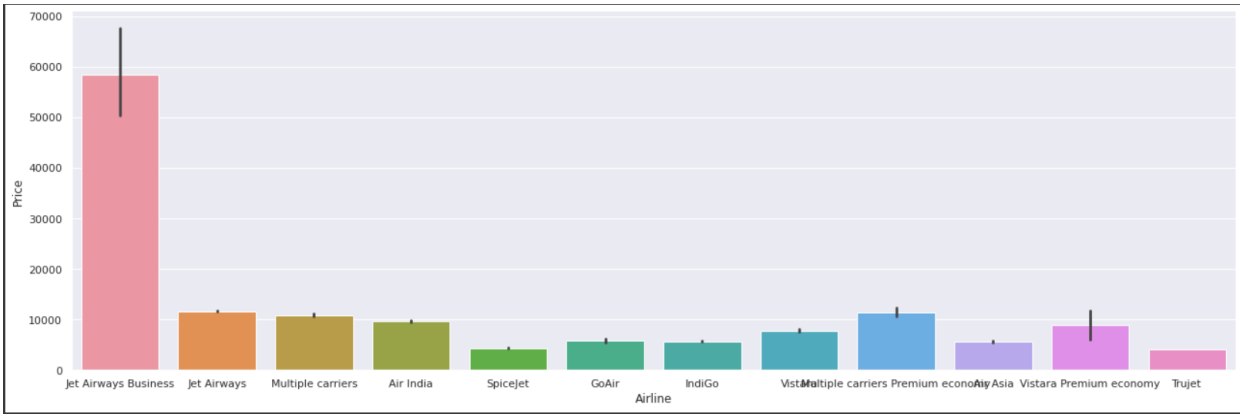
a.) The following things were done for it’s pre-processing :-

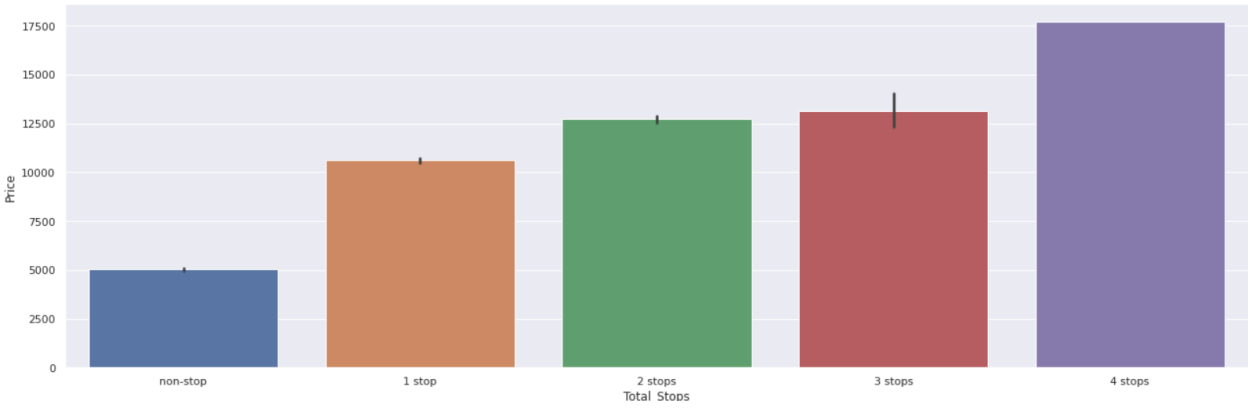
a.1) We checked if there were any NaN values and dropped them since there were less of them.
a.2) After that we encoded the following features ’Airline’, ’Source’, ’Destination’, ’Additional Info’ and ’Total Stops’ using the Label Encoder.
a.3) Then the following general data pre-processing was carried out in the code.

1. Interpreting Date_of_Journey
2. Making sense of Dep_Time and Arrival_Time
3. Using the Duration feature
4. Handling categorical data
5. Train test split

- **Data Visualization and Data Analysis:**

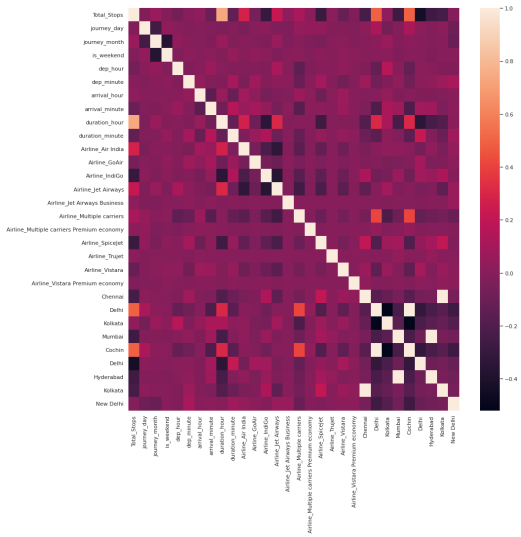
a.) Various plots and numerical data were used out in my code file for comparative analysis and getting the visual understanding of the data . few of them are shown below:





b.) The above plots and data out in code shows the relative importance of different features and by looking at the plot we get to know that 'Total Stops' has the most importance among all.

c.) The heat map in code (also shown below) suggests that the features are neither too strongly nor too weakly correlated.

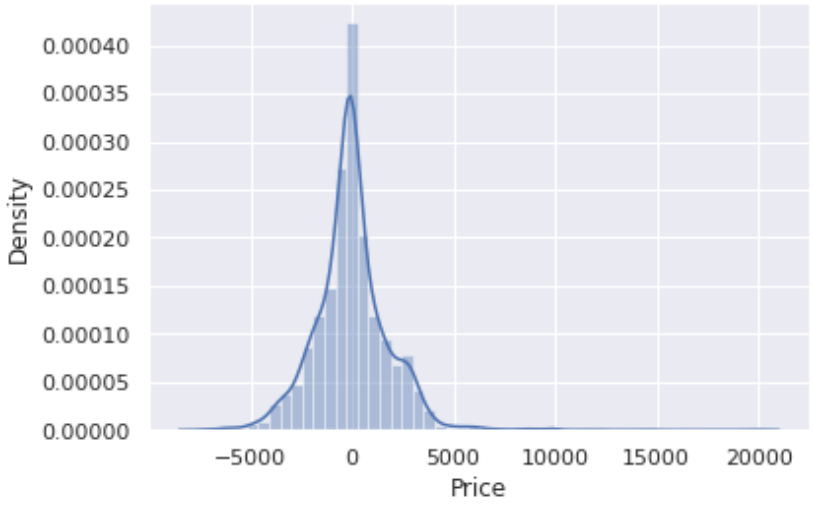


- **Applying Various Models:**

  a.) I have applied different Models in VS code for predicting the price of the flight and their performance scores are given below in the table.

Model	R2 Score	MAE	MSE	RMSE
Random Forest Regressor	0.7581	1281.5499	3553301.041	1885.0201
Decision Tree regressor	0.6843	1385.844	6181030.294	2486.167
Logistic regression	0.4377	2486.06	12438636.10	3526.84
Light GBM regressor	0.8251	1235.59	3306228.080	1818.303
Gradient Boosting Regressor	0.6733	1493.756	4335135.787	2082.098

b.) The plots of distribution in difference between the actual and predicted price are plotted in the code as well as shown below for some of the models like Random Forest and many others said above.



- **Model Optimization:**

  a.) Grid Search: We applied the concept of Grid search to find the best value of hyper-parameters like max depth and n estimators for various models and saw a significant increase in the R2 scores.

- **Comparison of Models:**

  a.) We applied different models and then comparing it using various scoring metrics , I got the following results:

  ----------Final Model used : Light GBM Regressor ----------

  r2 score : 0.8251155478869581
  MAE: 1235.5930765889534
  MSE: 3306228.080372286
  RMSE: 1818.3036271129984

- **Deployment of the Model:**

  a.) As a result, machine learning research usually focuses on optimizing and testing some parameters, but more criteria are needed for deployment in public policy settings. There has been relatively little attention paid to the distinction between technical and non-technical deployments.
  The real benefit and impact of machine learning models depends, however, on effective implementation.

  b.) Following a thorough analysis of different models and techniques I have decided to go with Multi Layer Perceptron (A Deep Learning Technique) as the model that will be used to predict flight prices based on user input, which is outside the scope of this course project, but I will try implementing it in the future so that the whole system is more robust.

End of the Report