Flight Price Webscraping & Prediction

Submitted By :

ASHUTOSH CHAUDHARY

ACKNOWLEDGMENT

I would like to express my gratitude towards my internship mentor Ms. Srishti Maan for helping me in completion of the project.

BUSINESS PROBLEM

From my understanding the problem is about wscraping data from websites of used cars to make dataset & then making prediction on flights ticket prices.

OBJECTIVE FOR PROBLEM UNDERTAKEN

We have to study every feature's behaviour present in the dataset with regards to target feature (Price of flight tickets) & make observation from its behaviour that how will it affect the target feature to build a model which would perform quite good on dataset to make prediction of flight ticket prices.

ANALYTICAL PROBLEM FRAMING

- **Origin of dataset & data types of every features**

We have to create dataset ourselves by webscraping the data from various websites & then convert it into dataset. After doing this we convert the dataset to csv file & then we need to import it using various libraries. Also features present in the dataset are both continuous & categorical data types.

- **Mathematical/Analytical modelling of the problem**

For visualization we only use two plots all the times that were countplot & boxplot & for model building we use Linear Regression, Decision Tree Regression, Random Forest Regression, Bagging Regression, AdaBooster Regression, GradientBooster Regression & Support Vector Regression models to opt best out of them to work on dataset.

- **Assumptions related to problem statement**

No assumptions were made while working on the dataset

- **Libraries & Tools used**

We used numpy, pandas, matplotlib.pyplot, seaborn, sklearn, pickle & warning libraries for this task.

STEPS TAKEN FOR THE TASK

1. **Importing Libraries for the task**

   Numpy, pandas, matplotlib.pyplot, seaborn, sklearn, pickle & warnings were imported for task to get completed.

2. **Importing Dataset using libraries**

   Imported the dataset using pandas library in jupyter notebook.

3. **Checking Dimension of dataset**

   By checking dimension of dataset we get to know that it contains 1568 rows & 8 columns.

4. **Checking Description of dataset**

   From description we find the mean, min value, max value, etc of every column which contains continuous data in them & from observation we concluded that only target feature was containing the continuous data in it.

5. **Checking for presence of null values in dataset**

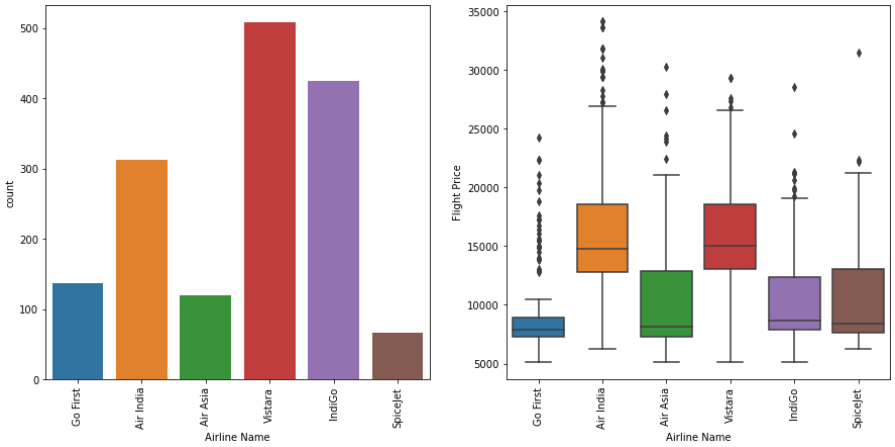
   We checked for the presence of null values in dataset & we get to know that no null value was present in any column of dataset.

6. Identifying Target variable

We have created the dataset ourselves, so we know that target variable is named Flight Price.

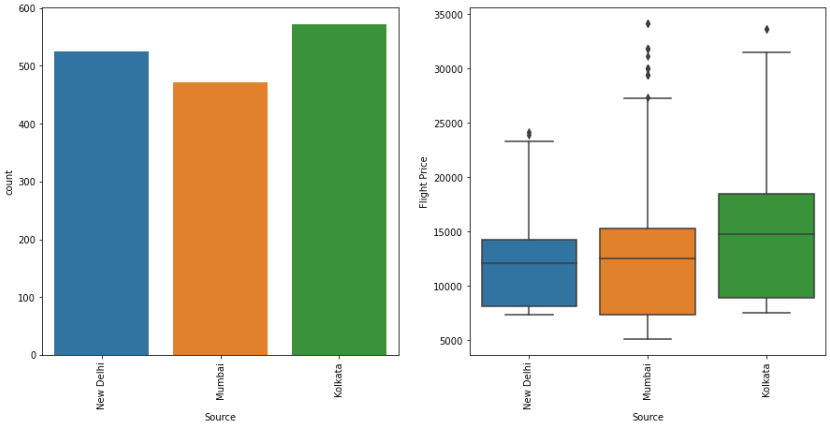
7. Performing EDA on whole dataset

Airline Name



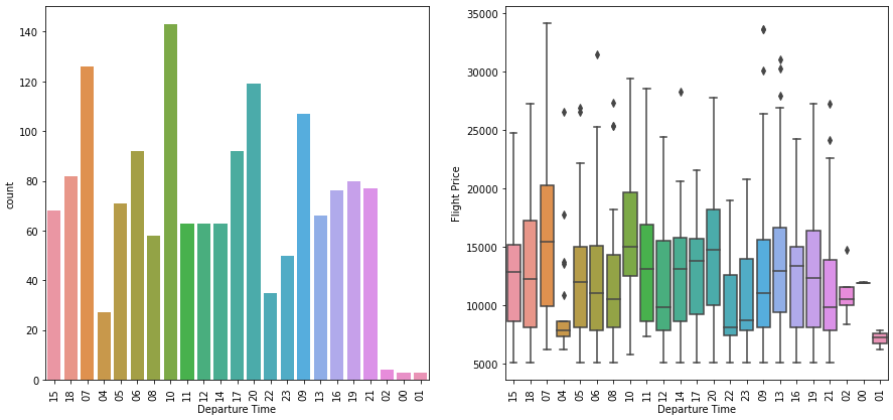
From visualization we concluded that most passengers are traveling from Vistara airlines & outliers of Flight Price are present in every category of Airline Name.

Source



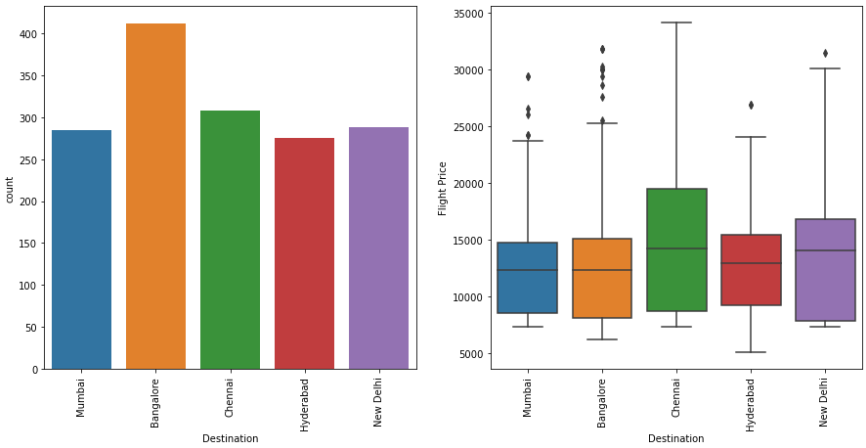
From visualization we get to know that most passengers boarding the flights are from Kolkata city & outliers of Flight Price are present in every category of Source.

Departure Time



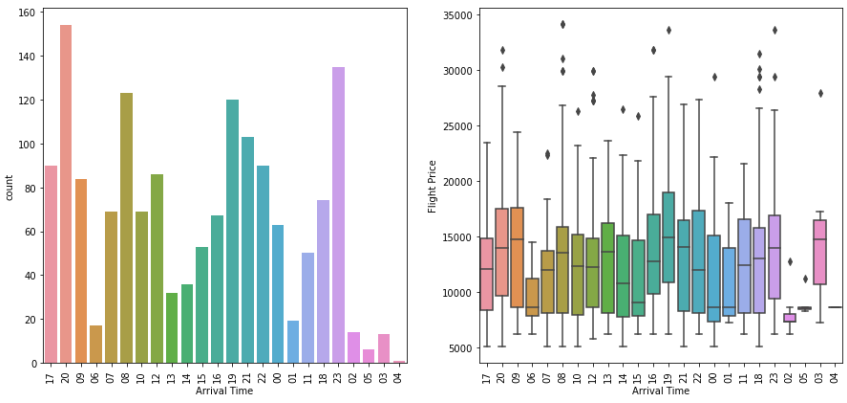
From visualization we concluded that most passenger's flight departure time is category 10 of Departure Time & outliers of Flight Price are present in 9 categories out of total 24 categories of Departure Time.

Destination



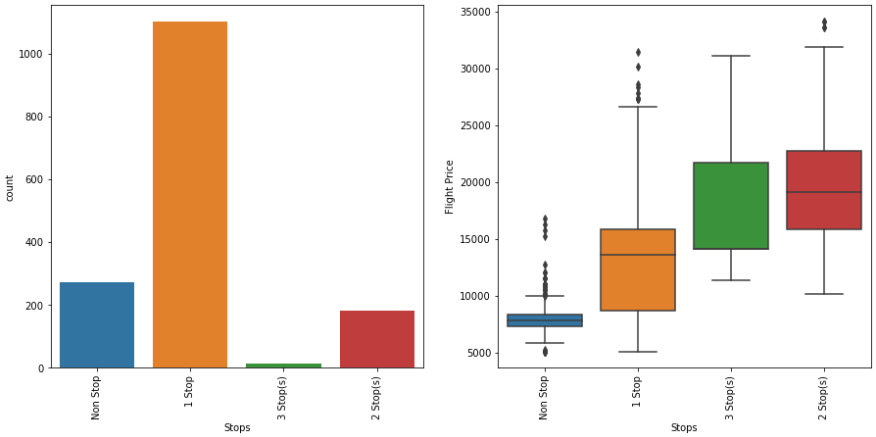
From visualization we concluded that most passengers are traveling to Bangalore city & outliers of Flight Price are present in 4 categories out of total 5 categories of Destination.

Arrival Time



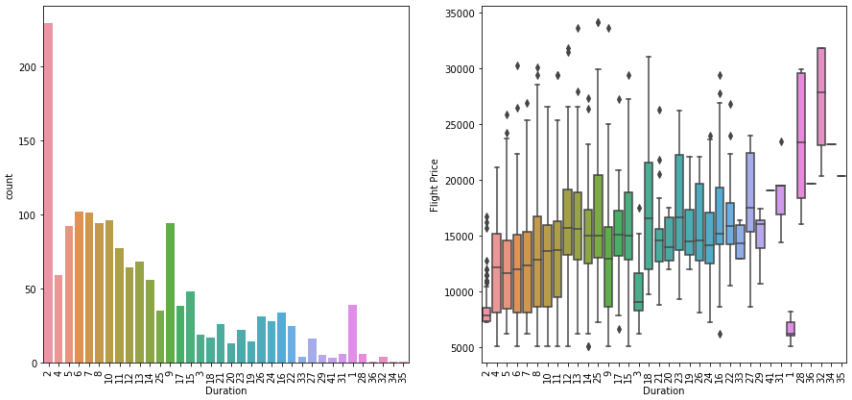
From visualization we concluded that most flights arrival time is category 20 of Arrival Time & outliers of Flight Price are present in 15 categories out of total 24 categories of Arrival Time.

Stops



From visualization we concluded that most flights flying are having 1 stop in between & outliers of Flight Price are present in 3 categories out of total 4 categories of Stops.

Duration



From visualization we concluded that most flights duration are of 2 hours & outliers of Flight Price are present in 19 categories out of total 36 categories of Duration.

8. Plotting Heatmap

We plotted heatmap to check correalation of every column with target column & we concluded that following (Source, Arrival Time, Stops, Duration) columns have high positive correlation with target column & also we dropped column named Destination to reduce multicollinearity.

9. Splitting & Scaling data

Then we splitted data into X & y in which y contains target variable & X contains other variable. Then we further splitted X & y into X_train, X_test, y_train, y_test using train_test_split. After that we scaled X_train & X_test set using StandardScalar to remove outliers if present any in columns.

10. Applying GridSearchCV

We chose 7 regression models (Linear Regression, Decision Tree Regression, Random Forest Regression, Bagging regression, AdaBoosting Regression, GradientBoosting Regression, Support Vector Regression) to check performance of them on the dataset to find out the best suited model for the dataset. Then one by one we applied GridSearchCV on each model to find best hyper parameter tuning while model is working on dataset. After applying GridSearchCV on every model we concluded that GradientBoosting Regression model is giving best test r2_score which was 0.6912 with hyper parameter combination (learning_rate : 0.1, max_depth : 4, min_samples_leaf : 3)

11. Applying Model on dataset

Chose GradientBoosting Regression Model & applied it on dataset to make prediction after tuning hyper parameters. After making prediction we made dataframe which contained actual & predicted values side by side.

12. Applying metrices

After making predictions we checked performance of model through various metrices :

- R2_score : 0.6912
- Root mean squared error : 3068.48
- Mean absolute error : 2206.99

As we observed metrices were giving quite satisfactory readings to conclude that model is performing quite well on dataset.

13. Saving Model

After checking performance of model we saved it using pickle library.

CONCLUSION

In the end we can conclude that flight ticket prices are mostly getting affected by Stops & Duration. So this means the long hours of flights have high ticket prices & flights with more stops in between have high ticket prices. Also Airlines Name are also affecting the flight ticket prices which means the trusted & high quality brand airlines have high price tickets & low quality brands have cheap tickets.