**Flight Price Prediction**

**1.Problem Definition:**

Flight ticket prices can be something hard to guess, today we might see a price, check out the price of the same flight tomorrow, and it will be a different story.

To solve the problem, we have been provided with prices of flight tickets for various airlines between the months of March and June of 2019 and between various cities, using which we aim to build a model which predicts the prices of the flights using input features.

**2.Data Analysis:**

We have 2 datasets here ---Training set and test set.

The training set contains the features, along with the prices of the flights. It contains 10683

records,10 input features and 1 output column that is 'Price'.

1.The Arrival time column has dates attached along with, which we will need to separate.

2.The Duration is in a string format, which we will need to convert to integer type.

3.The total stops also has text 'stops' added along with the number of stops, and certain column as

'non-stop' which we will need to convert to integer types.

4.We observe that we have all the columns as 'object' data types, and our target column 'Price' is of

integer type.

5.We have 1 missing value in Route column and 1 missing value in Total stops column. We will drop

both the missing values.

**3. EDA Concluding Remark:**

1.Airline column has 12 unique values-'IndiGo' ,'Air India' ,Jet Airways' ,SpiceJet' ,Multiple

carriers', 'Go Air', 'Vistara', 'Air Asia', Vistara Premium economy', Jet Airways Business', 'Multiple

carrier’s Premium economy', Trujet'.

2.Source column has 5 unique values-'Bangalore', 'Kolkata', Chennai', 'Delhi', 'Mumbai'.

3.Destination column has 6 unique values -'New Delhi', Bangalore', 'Cochin', 'Kolkata', 'Delhi',

'Hyderabad'.

4.Additional info has 10 unique values - 'No info', 'In-flight meal not included', 'No check-in baggage

included','1 Short layover', 'No Info', '1 Long layover', 'Change airports', Business class', 'Red-eye

flight','2 Long layover'.

**4.Pre-Processing Pipeline:**

1.Jet Airways is the most preferred airline with the highest row count, followed by Indigo and Air

India.

2.Count for Vistara Premium economy, Trujet, Multiple carries premium economy and Jet airways

business is quite low.

3.Majority of the flights take off from Delhi.

4.Chennai has the minimum count of flight take-offs.

5.Maximum flights land in Cochin.

6.Kolkata has the lowest count of receiving the flights.

7.Majority of the flights have stops as 1, flights with 3 and 4 stops are quite low.

8.The price column contains the minimum value as 1759 and maximum value as 79512.

9.Jet Airways business class has the highest prices between 50k-80k.

10.Outliers are present in Total hours, Total stops and price. We will not remove outliers from total

stops since price is impacted by number of stops.

11.Skewness present in Travel\_ hours column. We use log transform method to remove skewness.

12.To bring the data to a common scale, since there are certain columns with very small values and

some columns with high values. We will use standard scaler.

**5. Building Machine Learning Models:**

Linear regression model gives us 62% score, Lasso model gives us 61% R2 score and random

forest gives us 80% score. Random Forest model gives us the best accuracy.

**6. Concluding Remarks:**

We saved the model using pickle we got R2 score for test data is 81%, meaning that we are

actually, able to predict values quite near to actual prices, for majority of the rows and have

achieved the required task successfully.