

Flight Price Prediction Project

Submitted by:

Harshil Goradia

**ACKNOWLEDGMENT**

The data of flight for month of October and November was extracted from ‘MAKEMYTRIP.COM’. The data extracted from the website contains Airline name, destination, source, price and few other features that will be useful for predicting flight price.

**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. Time of purchase of ticket affects the price of the ticket. People who book tickets at last minutes tend to pay more than average sometimes 2X or 3X price of ticket.

* Conceptual Background of the Domain Problem

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

Now-a-days the prices of the flight ticket prices vary every time you check. To determine factors affecting the changes in prices and predicting the fare prices.

**Analytical Problem Framing**

* Data Sources and their formats

The data that is used in the project has been scarped from makemytrip.com for the month of October and November. The data was not in the form of table, later it was organised and placed in a table format.

* Data Preprocessing Done

Feature Engineering was done to make the data more understandable. SimpleImputer with strategy = ‘most frequent’ was used to replace the null values. The data type of the dataset was fixed to float later and it was scaled using label encoder for normalizing the distribution.

* Hardware and Software Requirements and Tools Used

1. Jupyter Notebook - The whole project was done in jupyter notebook which provides python environment.
2. Selenium – Selenium was used for web scarping to extra data from makemytrip.com. This data was later used for analysis.
3. Pandas and Numpy - These libraries which are most important for an predictive analysis project. These libraries provide a function of shaping the data, importing the data, and viewing the data.
4. Matplotlib and Seaborn – These libraries were used for visualising the data in the entire project. All the visualization done in the project were done using either one of these libraries.
5. Sklearn – This library was the most used library as compared to other libraries. This library has been used in pre-processing for imputing numerical null values and for scaling the dataset. Later it was widely used for importing algorithms for machine learning and predicting.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

To solve the problem and train dataset I feature engineered all the required columns to understand the data better and make it more efficient for machine learning. Later co-relation was also checked before visualizing the columns of the table. We produced dummies for Arrival column. The dataset was scaled using label encoder before using machine learning models.

* Testing of Identified Approaches (Algorithms

1. Linear Regression
2. Random Forest Regression
3. Decision Tree Regression
4. Ada Boost Regression
5. KNeighbors Regression

* Run and evaluate selected models

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated

This is highlighted as this the most accurate model.

Mean absolute error: 963.7780745614035

Mean squared error: 2714382.525653188

Root Mean Squared error: 1647.5383229695108

Accuracy: 0.7774157593785039

Text

Description automatically generated

Graphical user interface, text

Description automatically generated

Text

Description automatically generated

* Visualizations

Graphical user interface, application, Teams

Description automatically generated

Here we can observe that Month and Day are the most co-related columns.

Chart, histogram

Description automatically generated

Indigo company has the highest number of flight counts.

Chart, bar chart

Description automatically generated

Vistara-Indigo is the most expensive combination airline. Vistara alone is also expensive than other airlines. Cheapest is the combination of Air India-Go First.

Chart, bar chart

Description automatically generated

Flights which have 3 stops is the most expensive. Flights with no (N) stops are the cheapest.

Chart, bar chart

Description automatically generated

Flights which depart at 8am-9am and 11am-12noon are the most expensive. Flights at 5 o’clock in the evening is the cheapest flights.

Chart, bar chart

Description automatically generated

The flights arriving in the middle of the day are most expensive flights.

Chart, scatter chart

Description automatically generated

The red line passes through most possible points from actual and predicted points.

* Interpretation of the Results

From visualization it was found that Month and Day were the only highly co-related features. Indigo is most busy airline however Vistara is the most expensive. Flights that depart at 8 am and 11 am or arrive in the middle of a day are the most expensive. After modelling it was observed that RandomForestRegressor was the best performing model with 0.777 accuracy score, followed by AdaBoostRegressor with 0.770 accuracy score. The worst performing model was LinearRegression with >0.47 accuracy score.

**CONCLUSION**

* Key Findings and Conclusions of the Study

RandomForestRegressor has the highest accuracy score of 0.777. This means that more than 77% percent of the predicted prices of the flights were accurate.

* Learning Outcomes of the Study in respect of Data Science

Visualization plays important role in analysis of a dataset. Visualization guides data scientists to a proper direction for analysing the data. In this data set looking at the count of various features the following path was chosen. Cleaning data increases the quantity of the data, but it affects the quality of the data. In the dataset of this project cleaning the data was important to make it efficient and remove the null values using simpleimputer. Various algorithms were used to predict the price of the flights out of which RandomForestRegressor was best performing and LinearRegression was worst performing.