FLIP ROBO

Flight Price Prediction

Submitted by:

Nirav Mehta

ACKNOWLEDGMENT

I scraped Google Flights for the information. There are statistics between various cities from various dates.

INTRODUCTION

- Business Problem Framing

Finding the best time to buy plane tickets can be difficult for passengers because they lack the knowledge to predict future price changes. Flight ticket costs are quite dynamic by nature. In this project, our main goals were to identify the underlying trends affecting travel costs in India using scraped data and to recommend the ideal window of time for purchasing a ticket by using a regression model to forecast flight costs.

We gathered information for this research from 3 routes in 3 cities all over India. 6127 data points total from data collected over a 2-month period in Mumbai-Delhi, Delhi-Mumbai, and Bangalore-Delhi.

We gathered information for this research from 3 routes in 3 cities all over India. 6127 data points total from data collected over a 2-month period in Mumbai-Delhi, Delhi-Mumbai, and Bangalore-Delhi.

- Conceptual Background of the Domain Problem

Anyone who has purchased a plane ticket is aware of how costs may change suddenly. Airlines employ highly technical, quasi-academic "revenue management" or "yield management" strategies. The

The lowest priced ticket for a specific date fluctuates in price over time. Typically, this occurs as an effort to increase profits based on -

Purchase timing trends (making sure last-minute purchases are expensive)

Increasing costs on a flight that is filling up in order to keep it as full as possible Reduce sales and keep goods on hand for those pricey, last-minute purchases.

Therefore, if we could provide travellers with information about the best time to purchase their plane tickets based on historical data and also show them various trends in the airline sector, we might assist them in saving money on their travel expenses. It would be a practical application of data analysis, statistics, and machine learning methods to address a recurring issue for passengers.

- Motivation for the Problem Undertaken

Air travel is one of the most important ways of transportation in the fast-paced world of today. Contrary to other forms of transportation, flight prices are extremely volatile. In this project, we are attempting to address this issue by assisting travellers in finding the best deals on tickets while taking into account their needs and projecting price changes based on those needs. This will enable a customer to make an informed choice regarding the itinerary of his flights.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

According to a study of the deleted data, the target variable in the dataset has continuous values and the majority of the attributes are categorical. Consequently, it is a regression issue.

• Data Sources and their formats

The Data is Scrapped using Selenium from Google Flights.

We gathered information for this research from 3 routes in 3 cities all over India. 6127 data points total from data collected over a 2-month period in Mumbai-Delhi, Delhi-Mumbai, and Bangalore-Delhi.

• Data Preprocessing Done

a) Dataset loading
b) Eliminating redundant values
c) Using feature engineering, removing years from Year
d) Our dataset's structure, with 1.37 million rows and 14 columns
e) Examining statistical indicators
f) Our dataset contains no Null values.
g) OnHotEncoding is used to label categorical items.
h) Taking Away Outliers
i) Eliminating skewness
j) Correlating several features and dropping the city.

• Data Inputs- Logic- Output Relationships

The input data made available aids in understanding the numerous elements that make up the factors that affect flight costs. The target variables may vary if any specific feature changes.

- State the set of assumptions (if any) related to the problem under consideration

This information pertains to three major locations in India, and consequently, no one would choose a trip that required an international layover, cost a lot of money, or took too long. We have eliminated the flights with international layovers based on this supposition.

- Hardware and Software Requirements and Tools Used

Hardware Used: 8Gb+ of RAM, 128 GB or more (256 GB recommended), 1TB+ of free hard drive space.

Software's and Tool's Used: Jupyter Notebook, NumPy, Pandas, Matplotlib, Seaborn, Plotly, Scikit Learn

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

There are no null values among the 6124 data points in the data set. We know that this is a regression problem because the dependent feature is a continuous variable. Due to the fact that all other features are categorical variables, which are unsuitable for either removing outliers or accounting for skewness, I discovered outliers in 2 features when I analysed the dataset. The IQR method was used to correct the outliers. According to studies, there were several elements that had no bearing on prices, thus they were removed. Data visualisation was used to visually depict the cleaned and transformed data. The creation of a model for the data set was the final and most crucial step.

- Testing of Identified Approaches (Algorithms)

The Algorithm's used are as follows:

- LinearRegression :-

```
o  RMSE 2143.9600568725723
o  MAE 1545.3499662416486
o  r2_score : 68.81989409136125
o  cv_score : -289.3490613879849
o  Difference between r2_score and cv is  358.1689554793461
```

- Ridge :-

```
o  RMSE 2143.9338375490843
o  MAE 1545.0936247491961
o  r2_score : 68.82065671404321
o  cv_score : -288.065846182562
o  Difference between r2_score and cv is  356.8865028966052
```

- Lasso:-

```
o  RMSE 2143.9529005172953
o  MAE 1545.5874682833712
o  r2_score : 68.82010224406663
o  cv_score : -289.05028065462324
o  Difference between r2_score and cv is  357.87038289868985
```

- ElasticNet:-

```
o  RMSE 2377.146424435179
o  MAE 1795.4017850020005
o  r2_score : 61.668477148464085
o  cv_score : -859.2838379662171
o  Difference between r2_score and cv is  920.9523151146811
```

- DecisionTreeRegressor:-

```
o  RMSE 845.4064805278203
o  MAE 175.83426443202978
o  r2_score : 95.15185778006821
```

```
o  cv_score : 62.454291920030904
o  Difference between r2_score and cv is  32.69756586003731
```

- RandomForestRegressor:-

```
o  RMSE 812.2905922837805
o  MAE 341.24446513552664
o  r2_score : 95.52423729767784
o  cv_score : 59.438710929944996
o  Difference between r2_score and cv is  36.08552636773285
```

- GradientBoostingRegressor:-

```
o  RMSE 1890.1151111452452
o  MAE 1330.7736777159635
o  r2_score : 75.76624490807336
o  cv_score : -170.0793391249583
o  Difference between r2_score and cv is  245.84558403303166
```

- Run and Evaluate selected models

The following techniques are utilised for fitting the train and test datasets and hyper parameter tuning: a. GradientBoostingRegressor b. CatBoostRegressor c. Ridge Regressor

- Key Metrics for success in solving problem under consideration

The following Key Metrics were utilised to solve the issue: R2 Score, Cross-Validation Score, MSE, and RMSE

- Interpretation of the Results

With a cross-validation difference of 2.83, we have gotten a reasonably decent accuracy of 82 percent.

CONCLUSION

- Key Findings and Conclusions of the Study

- ✓ The prices decrease with further booking and travel dates.
- ✓ Fridays, Saturdays, and Sundays have relatively higher flight prices.
- ✓ While Vistara is the most popular airline despite having higher tickets, Air Asia is recommended for those looking for the lowest costs.

- Learning Outcomes of the Study in respect of Data Science

There weren't many outliers in the dataset. The only issue was that the system needed more training and analysis data. Data cleansing was crucial to get accurate predictions because the data we scraped was raw and misplaced. Duration, airline, and the time difference between the booking and departure dates were the most crucial elements in deciding the price. We used the gradient boosting approach since it produced the best outcomes.

- Limitations of this work and Scope for Future Work

There is a discrepancy between the actual prices and the predicted prices, which could result in under fitting. This can be the result of training a model with fewer requirements or features. Therefore, if a better performance model is required in the future, the majority of the parameters will need to be scrapped in order for the model to produce better results.