[Date]

Flight Price Prediction using Machine Learning Techniques

Soham Pawar, Muaaz Patel, Dipali Bhanushali, Khushbu

***Pursuing Masters in Data Science & Business Analytics.***

**Abstract**

# According to a report, India’s civil aviation industry is on a high-growth trajectory. India aims become the third-largest aviation market by 2020 and the largest by 2030. Indian domestic air traffic is expected to cross 100 million passengers by FY2017, compared to 81 million passengers in 2015, as per Centre for Asia Pacific Aviation (CAPA).

According to Google Trends, the search term - "Cheap Air Tickets" is most search in India. Moreover, as the middle-class of India is exposed to air travel, consumers hunting for cheap prices increases

# Introduction

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 this 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 various input features

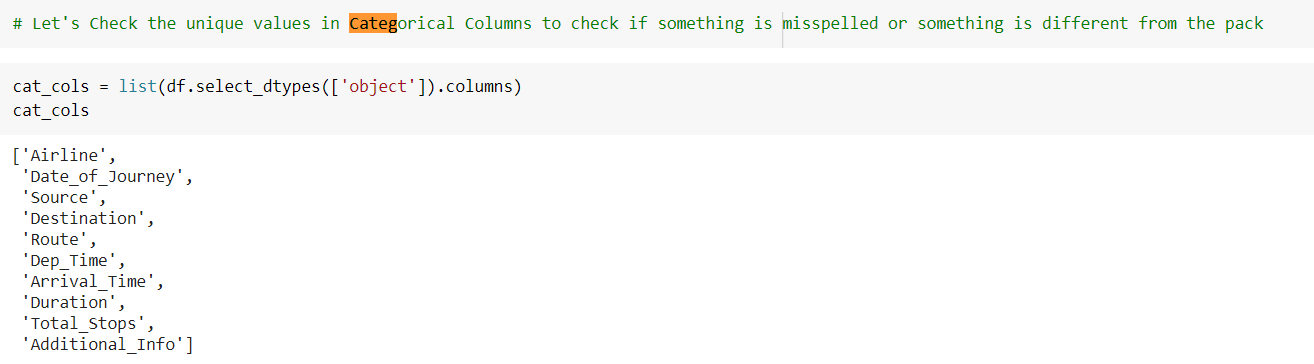
* 1. **Proposed System**

1. **Data Collection**

The dataset for this project can be downloaded from this Kaggle link. Unzip the downloaded zip file and place the "Train.csv" file in your local drive. This is the file that we are going to use to train our machine learning model.

1. **Data Pre-processing**

We now start exploring the columns available in our dataset. The first thing we do is to create a list of categorical columns, and check the unique values present in these columns –



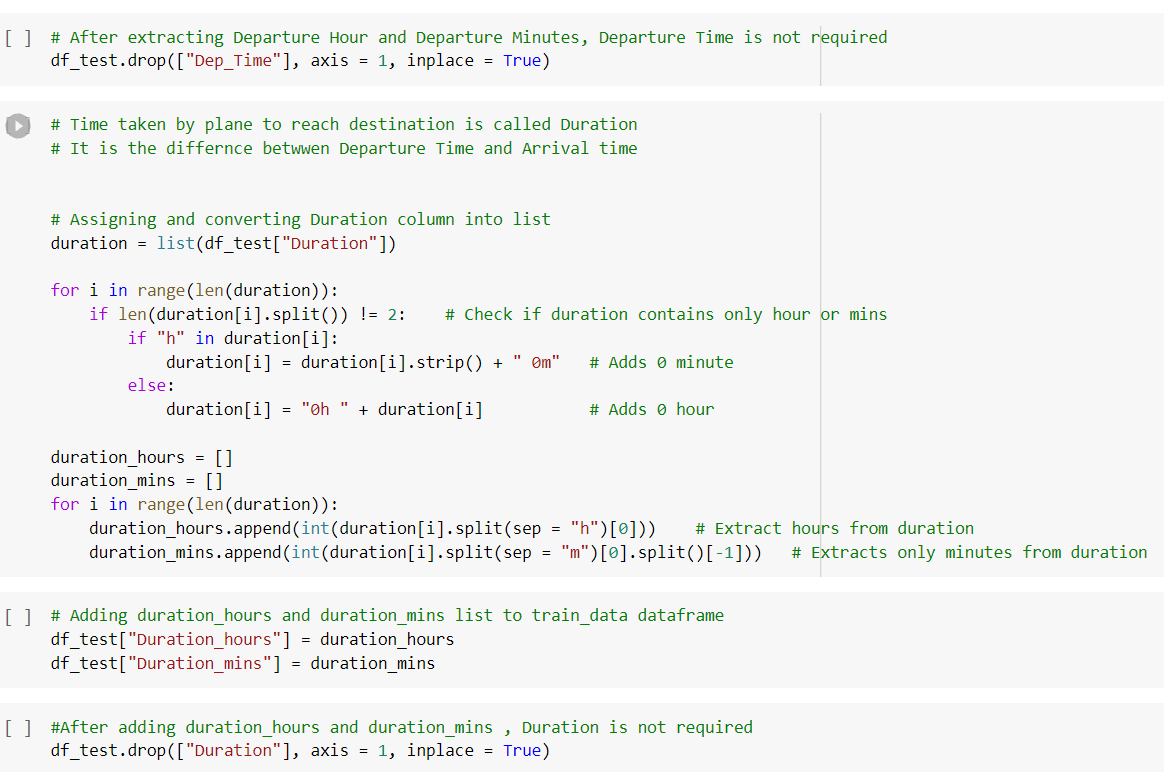
1. **Feature Extraction and Selection**

The data scientist's data has several features that may or may not be relevant to the topic of interest. Also, it may not be in a suitable format. The first and foremost task to the data scientist is to extract the appropriate collection of attributes that preferably suits the learning algorithm. Before processing, it needs to be transformed to prevent relapse problems like overfitting and underfitting as presented. The following Table 1 shows the list of features present in the dataset

|  |  |
| --- | --- |
| **Feature** | **Description** |
| **Airline** | The name of the airline |
| **Date\_of\_Journey** | The date of the journey |
| **Source** | The source from which the service begins |
| **Destination** | The destination where the service ends |
| **Route** | The route taken by the flight to reach the destination |
| **Dep\_Time** | The time when the journey starts from the source |
| **Arrival\_Time** | Time of arrival at the destination |
| **Duration** | Total duration of the flight |
| **Total\_Stops** | Total stops between the source and destination |
| **Additional\_info** | Additional information about the flight |
| **Price** | The price of the ticket. |

**Table 1: List of Feature**

Next, we will analyze the 'Timestamp' category. It represents the exact time the departure time Arrival time duration hence it plays an important role in prediction of price extracting time from the timeframes



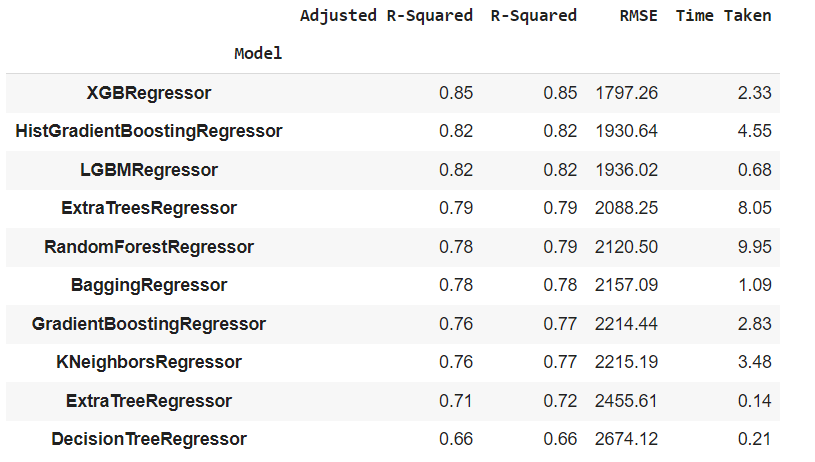
1. **Train and take a look at knowledge Sets** Once the dataset is processed, we want to divide it into two components that are coaching and take a look at the set. We'll take and use the train\_test\_split
   1. **Related Work**

# This study shows that it is feasible to predict the airline ticket price based on historical data. One possible way to increase the accuracy can be combining different models after carefully studying their own performance on each individual bin. Additionally, as the learning curve indicates, adding more features will increase the accuracy of our models. However, limited by the current data source that we have, we are unable to extract more information of a particular flight. XGBoost has tried to be a good model and give accurate information.

# Machine Learning algorithms are applied on the dataset to predict the dynamic fare of flights. This gives the predicted values of flight fare to get a flight ticket at minimum cost. Data is collected from the websites which sell the flight tickets so only limited information can be accessed. The values of R-squared obtained from the algorithm give the accuracy of the model. In the future, if more data could be accessed such as the current availability of seats, the predicted results will be more accurate.

# MCLP has many successful applications, including credit card portfolio management, credit card risk analysis, firm bankruptcy prediction, network intrusion detection, medical diagnosis and prognosis, and classification of HIV-1 mediated neuronal dendritic and synaptic damage. Multi-Criteria Linear Programming Regression (MCLPR) was firstly introduced by Zhang, which converted a classification problem to a regression one. The data can be separated into two groups to move it downward and upward by parameter and then classified by hyperplane to construct a regression model. The excellence of MCLPR is its ability to fix the ill-posed condition with a limited amount of sample, handling non-linear relationships by kernel function, and giving the global solution if it exists. MCLPR has already proved its performance in many real-life datasets.

# 2.1. Experimental Results



**2.2 Conclusion**

We further proceed to test the object that we saved using pickle and the model can be reused. We used XGBRegressor as it can approximately accurate best predicted price from the model.

This study shows that it is feasible to predict the airline ticket price based on historical data. One possible way to increase the accuracy can be combining different models after carefully studying their own performance on each individual bin. Additionally, as the learning curve indicates, adding more features will increase the accuracy of our models. However, limited by the current data source that we have, we are unable to extract more information of a particular flight