Flight Price Prediction

Problem Statement

Flight price ticket costs may be one thing exhausting to guess, these days. we’d see a price, verify the value of identical flight tomorrow and it'll be a special story. We’d have often detected traveller’s expression that flight ticket prices are thus unpredictable. Here you'll be given prices of flight tickets for numerous airlines between the months of March and June of 2019 and between various cities.

Size of coaching set: 10683 records

Size of check set: 2671 records

FEATURES:

Airline: The name of the airline.

Date\_of\_Journey: The date of the journey

Source: The supply from that the service begins.

Destination: The destination wherever the service ends.

Route: The route taken by the flight to succeed in the destination.

Dep\_Time: The time once the journey starts from the source.

Arrival\_Time: Time of arrival at the destination.

Duration: Total length of the flight.

Total\_Stops: Total stops between the source and destination.

extra\_Info: Additional info regarding the flight

Price: the value of the price ticket

Problem Definition

Customers have always found flight ticket pricing changes puzzling and difficult to forecast. Airlines utilize dynamic pricing strategies to increase demand for their seats and increase revenue. As a result, consumers are having difficulty finding low-cost tickets. The closely associated terms with their airline rates, such as commercial, financial, social issues, and marketing, are taken into account when implementing these dynamic tactics. Airlines are doing all possible to maintain high revenue and profit margins. Airline costs frequently fluctuate unpredictably, with flight tickets tomorrow differing from those today. The system is tough because each aircraft has a restricted number of seats to sell. If there is a high demand for flight tickets, the price will rise; nevertheless, if there are unsold seats, the price of air tickets may fall as a result of the loss of revenue. Machine Learning is a great technique to learn from historical flight pricing data and construct logic based on the data provided to meet the difficulty of predicting travel rates. The Random Forest Regressor and Extra Tree Regressor will be used to help us predict flight prices depending on a range of parameters, such as data extraction, data processing, and data interpretation.

Data Analysis

Data was gathered from a variety of sources. The comprehensive information on the flight ticket is obtained from an online data source (Github.com). This data, which is in the form of a csv record, was taken from the website. The information containing input features and its target variable required for data analysis is contained in the file. To improve the accuracy of the results, we retrieved new characteristics from the existing variables. To make data analysis easier, features like "Arrival Time," "Arrival Date," "Arrival Month," "Day," "Month," and "Year" are generated.

Cleaning and Data Preparation

There is a lot of work to be done with all of the acquired data. After gathering information, all irrelevant data "features" such as "Arrival Date" and "Arrival Month" are eliminated, as are duplicate features and incorrect characteristics (such as "Additional information" and "Year"). The missing values in variables like "Route" and "Total Stops" in the data set are resolved using the Simple Imputer function. Data transformation is one of the important phases in data preparation; hence, data in string format is converted to float data type utilising encoding techniques such as OrdinalEncorder (). For example, the data type "Airline" is a string rather than an integer. This is done in order to create a final data frame from the original data and do thorough data analysis. This is the most critical and time-consuming procedure in all ML projects.

Visualizations for data analysis

Data preparation is breaking down information, recognising patterns, and then using various machine learning algorithms. In this example, we partition our data into three data frames based on the type of data: nominal, ordinal, and continuous.

Nominal Data Visualizations

For nominal/categorical data, we utilise cat plot since it shows the frequency of the columns. While examining data, the following observations were made.

* Jet Airways, Indigo, and Air India have the most flights, while True Jet and Vistara Premium Economy have the fewest.
* The majority of flights depart from Delhi, with fewer departing from Chennai.
* The majority of flights land at Cochin Airport, with the fewest landing at Kolkata Airport; the majority of flights have only one stop in between, and the majority of them are also non-stop; however, only a few flights have three or four stops.
* The flights typically travel throughout the months of June, May, and March.

Visualization of Values of a Continuous Type

As stated in the observation:

* Except for the price columns, the data is widely dispersed in other columns.
* The data in the price columns is right skewed, but it is a goal variable.
* The maximum number of flights are flying between the dates 3 and 7 in the day column.

Conclusions of the EDA

We receive all the important variables and significant information required for developing an ML model after executing all the data transformations, integration, and cleaning. The data set ends up with 11 variables and 10,683 records. Important features utilised for analysis are included in the final data set:

1. Airlines
2. Origin
3. Destination
4. Route
5. Deep Time
6. Arrival Time
7. Duration
8. Total Stops
9. Price
10. Months
11. Day

Remarks on the Final Observation

The standard deviation in the "Route," "Deep Time," "Arrival Time," "Duration," and "Price" columns is two, indicating that the values in these columns are widely dispersed and far from the mean.

The standard deviation of the other columns is too large, indicating that the data is distributed normally and that there are fewer chances of skewness.

The target variable ("Price") has a minimum and maximum price of 1759 and 79512, respectively.

The range is excessively broad.

The "Total Stops" column is the most inversely correlated, meaning that the more stops, the lower the flight costs.

"Route" is the most positively associated variable.

The variables "Route," "Arrival Time," "Source," "Month," and "Deep Time" are all favourably connected with the target variable, while "Airline," "Destination," "Duration," "Day," and "Total Stops" are all adversely correlated.

Pipeline for pre-processing

We didn't detect much skewness in our data because the skewness range is threshold +/-0.5. Only the variables "Airline," "Destination," "Price," and "Month" displayed skewness, with all input characteristics being of the object data type and "Price" being the target variable. As a result, there is no skewness in the data.

Detecting and Cleaning Outliers

Outliers are detected using statistical approaches like factor analysis.

The z-score tool is used to find outliers in this scenario. Only the "Price" column, which is the Target Variable, had outliers that could not be detected as outliers. As a result, the data does not contain any outliers.

Validation

Behave to normalise the data and get the values within a specific range for the model to interpret data in order to develop and train ML models. Because the data's value ranges are wide, we utilised the Standard Scaler () approach to normalise it. The Standard Scaler method returns all values in the datasets that fall between 0 and 1. This will aid the learning of data via machine learning algorithms. We utilise regression methods because the target variable has a variety of values.

Building Machine Learning Models

The main purpose here is to create a model that will aid in measuring the performance of better and more refined algorithms. We compared and tested different Regression and Ensemble Techniques to determine which algorithm performs better then stacked them all together to see how the model predicts.

Regression using Random Forests

Random Forest Regression is a Regression task-solving supervised machine learning technique. The correlation between two continuous variables is determined using a simple Random Forest Regression technique. Prediction inaccuracy is kept to a bare minimum.

The mean absolute error (MAE) for the model after applying Random Forest Regression was found to be 1.19 percent.

The mean squared error (MSE) is 3.52%.

RMSE (Root Mean Square Error):1.87 percent

We utilised Randomized Search CV to find the optimal parameters to increase the model's performance.

Final Thoughts

Random Forest Regression has an r2 value of 80.07 percent.

Using Ensemble Techniques to this regression Models to create stability in the model's performance, we acquire rf score of 81 percent.

Following the identical steps as for the training file (the entire EDA process), we used the best saved model from the training file to forecast the analysis of the testing file.