Flight Price Prediction Project

1. Table of Contents

1. Introduction

The objective of this project is to predict flight prices using machine learning algorithms, specifically Random Forest and Decision Tree classifiers. Accurate flight price predictions can help airlines optimize their pricing strategies and help travelers plan their trips more efficiently.

2. Data Preparation

2.1 Data Loading

The dataset was loaded and basic information about the dataset was displayed to understand its structure and contents.

2.2 Exploratory Data Analysis (EDA)

EDA was performed to understand the dataset better. Key steps included checking for missing values, understanding the distribution of data, and visualizing important features.

Key Findings from EDA:

Some columns had a significant number of missing values.

Certain features like 'Route' and 'Total_Stops' were redundant.

'Additional_Info' column had too many values labeled as 'No info', making it less useful.

2.3 Data Cleaning and Preprocessing

Based on EDA findings, data cleaning and preprocessing steps were performed:

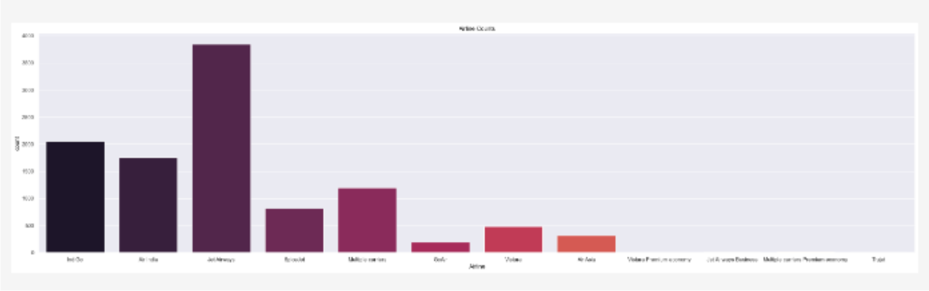
Handling Missing Values: Dropped or filled missing values appropriately.

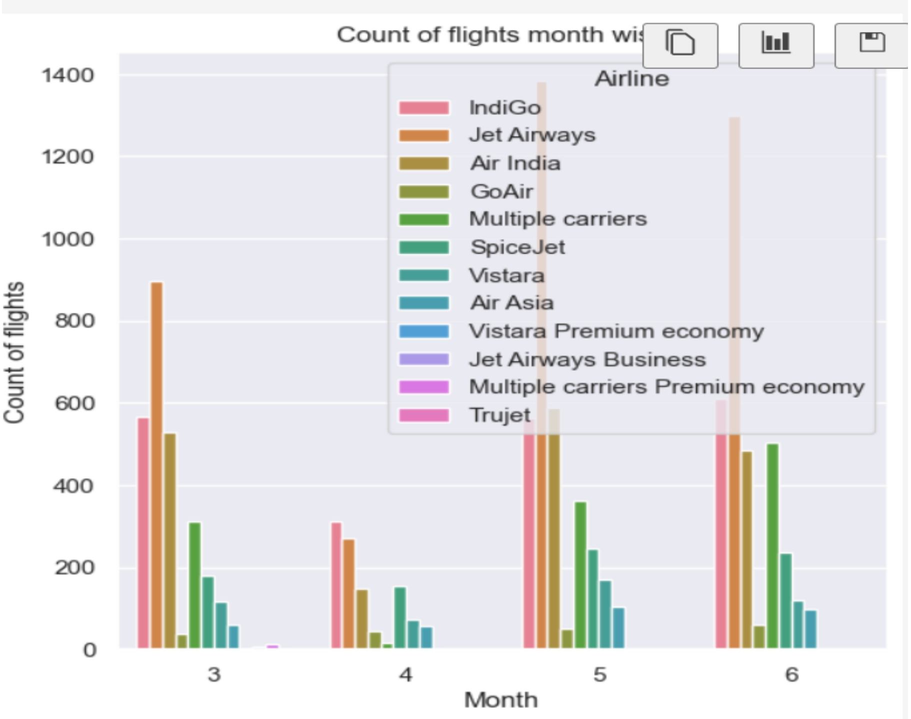
Encoding Categorical Variables: Converted categorical variables into numerical values.

Feature Selection: Selected important features that contribute significantly to the target variable.

Plot of Airplanes Count

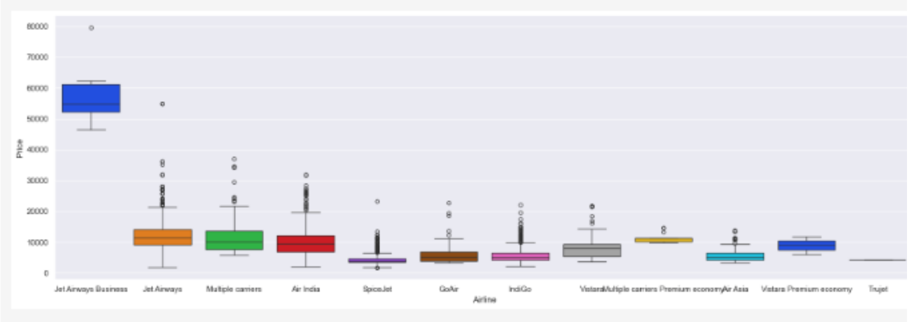
1. Jet Airways have the higher count
2. Then Indigo
3. Then Air India





Plot of Price vs Airplanes

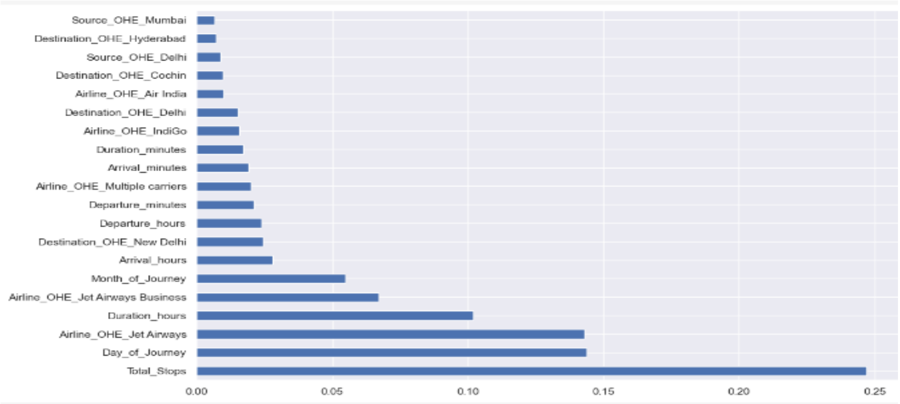
1. Jet Airways Business have the higher price.
2. Then after all flights have almost same price.



Important Features

We have use the ExtraTreesRegressor to find out the important features.

1. Total stops is very important feature.
2. Then after Date of Journey and Jet Airways



3. Model Training

3.1 Random Forest

A Random Forest regressor was trained using the prepared dataset. Random Forest is an ensemble method that builds multiple decision trees and merges them to get a more accurate and stable prediction.

3.2 Decision Tree

A Decision Tree regressor was also trained for comparison. Decision Trees split the data into subsets based on the most significant feature, aiming to create the most homogenous branches.

4. Hyperparameter Tuning

4.1 Randomized Search CV

Hyperparameter tuning was performed using Randomized Search CV to find the best parameters for the Random Forest model. This process involves searching across different hyperparameters to find the combination that results in the best performance.

4.2 Parameters Tuned:

Number of trees in the forest (n_estimators)

Maximum depth of the tree (max_depth)

Minimum number of samples required to split a node (min_samples_split)

Minimum number of samples required at each leaf node (min_samples_leaf)

Number of features considered for splitting at each node (max_features)

5. Model Evaluation

The performance of both models was evaluated using various metrics:

Mean Absolute Error (MAE): Measures the average magnitude of errors in a set of predictions, without considering their direction.

Mean Squared Error (MSE): Measures the average of the squares of the errors, which gives more weight to larger errors.

Root Mean Squared Error (RMSE): Square root of MSE, which is in the same units as the target variable.

6. Predictions

Using the trained models, predictions were made on the test dataset. The results were compared to evaluate the accuracy and effectiveness of each model.

7. Accuracy Table

The table below shows the performance metrics for both the Random Forest and Decision Tree models:

Metric	Random Forest	Decision Tree
Mean Absolute Error	1175.847	380.299
Mean Squared Error	4383434.006	736856.890
Root Mean Squared Error	2093.665	858.403
R^2 Score	0.7967	0.9658

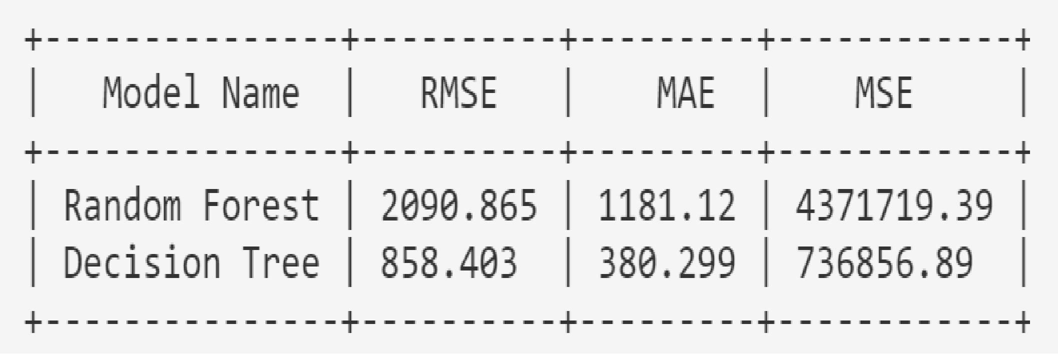
Mean Absolute Error (MAE): The average magnitude of errors in a set of predictions, without considering their direction. Lower is better.

Mean Squared Error (MSE): The average of the squares of the errors. Lower is better.

Root Mean Squared Error (RMSE): Square root of MSE, which is in the same units as the target variable. Lower is better.

R^2 Score: Proportion of the variance in the dependent variable that is predictable from the independent variables. Higher is better, with 1 being a perfect score.

8. Comparision



```
+---------------+----------+---------+------------+
|   Model Name  |   RMSE   |   MAE   |    MSE     |
+---------------+----------+---------+------------+
| Random Forest | 2090.865 | 1181.12 | 4371719.39 |
| Decision Tree | 858.403  | 380.299 | 736856.89  |
+---------------+----------+---------+------------+
```

9. Conclusion

The project successfully demonstrated the use of Random Forest and Decision Tree regressor in predicting flight prices. Hyperparameter tuning significantly improved the performance of the Random Forest model. Future work could involve exploring more advanced algorithms and additional features to further enhance prediction accuracy.

10. References

https://www.analyticsvidhya.com/blog/2022/01/flight-fare-prediction-using-machine-learning/