EE769 - Introduction to Machine Learning

Project Report

Flight Price Prediction

Anukool Vikram - 210040019
Rahul Agarwal - 210040120
Deepika Jandu - 210040043

Project Guide
Prof. Amit Sethi

Introduction

Flight price prediction is a crucial task in the airline industry as it helps customers plan their travel budgets and assists airlines in revenue management. This project focuses on predicting flight prices using machine learning techniques, particularly leveraging the random forest regression algorithm.

Problem Statement

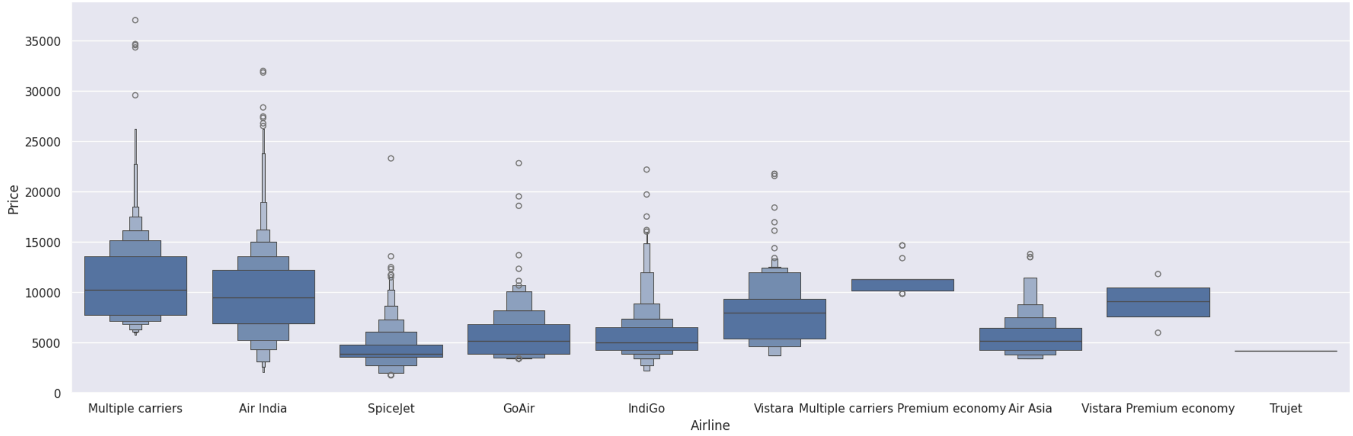
The primary objective of this project is to develop a model that accurately predicts flight prices based on various input features such as airline, source, destination, departure time, arrival time, duration, and other factors.

Dataset Description

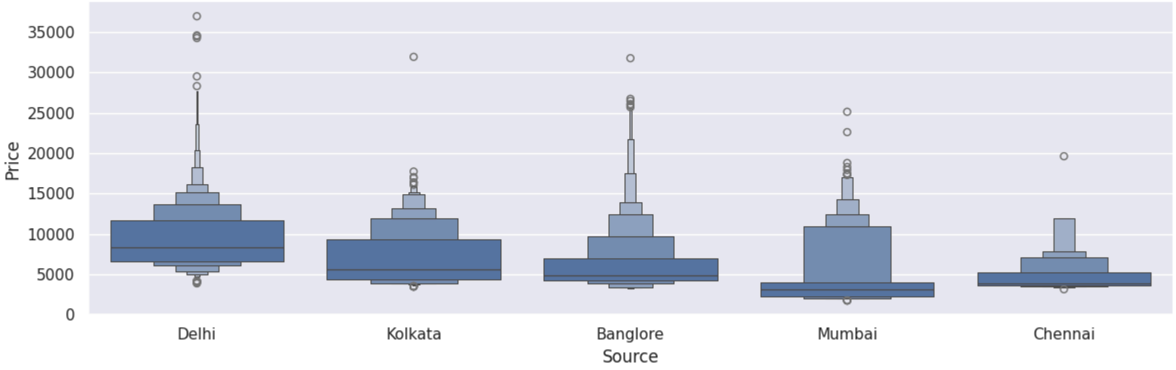
The dataset used in this project contains information about flight prices along with several features such as airline, source, destination, departure time, arrival time, duration, and additional information. The dataset is split into training and testing sets for model development and evaluation.

Exploratory Data Analysis (EDA)

- The EDA phase involves understanding the dataset's structure, identifying missing values, and exploring the distribution of features.
- Date features like 'Date_of_Journey', 'Dep_Time', and 'Arrival_Time' are converted to datetime objects for further analysis.
- Categorical features like 'Airline', 'Source', and 'Destination' are analyzed for their impact on flight prices using visualizations such as box plots.



Box plot showing the relationship between different airlines and flight prices.



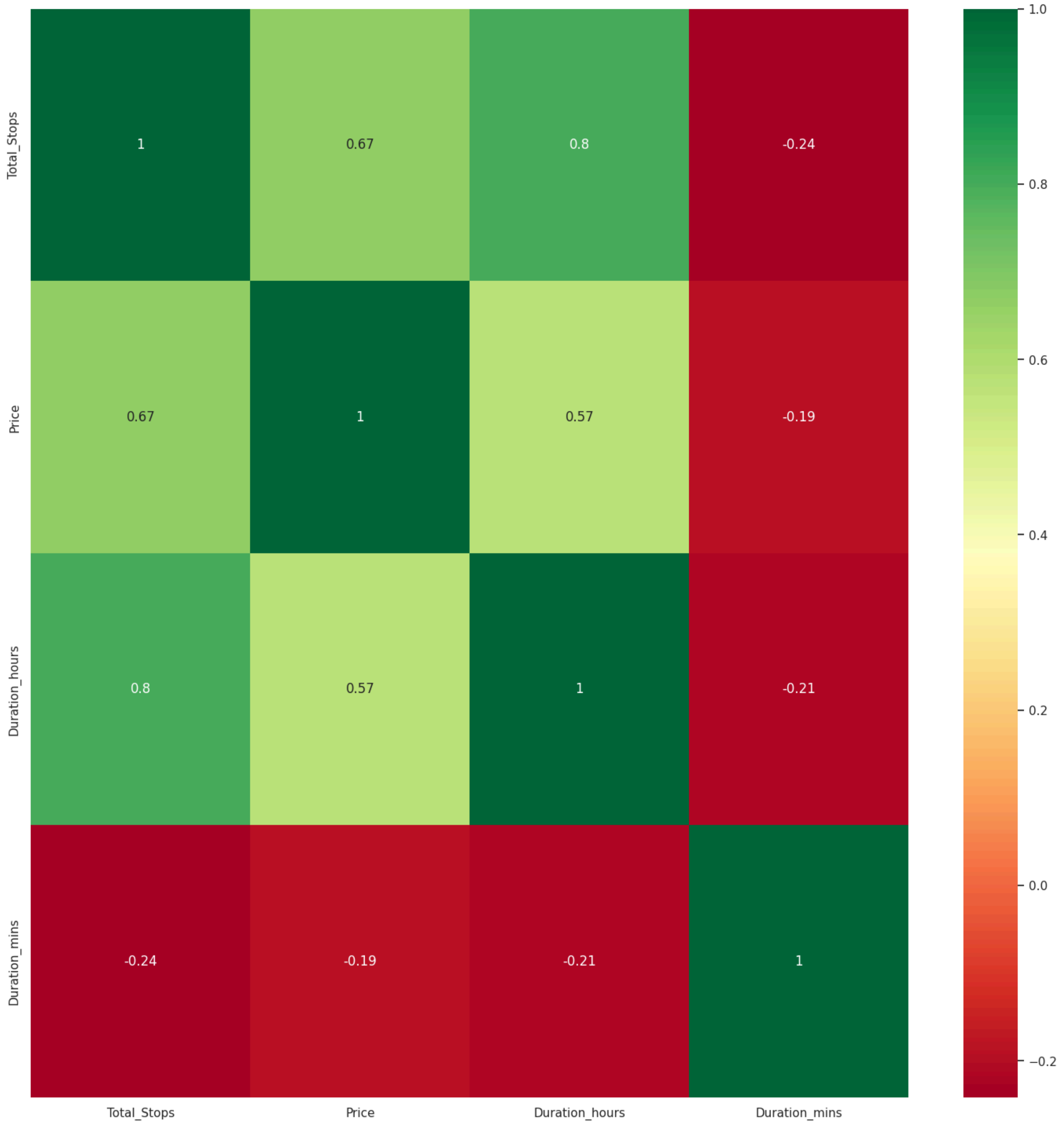
Box plot showing the distribution of flight prices based on the source of the flight.

Data Preprocessing

- Categorical variables are encoded using techniques like One-Hot Encoding and Label Encoding to prepare the data for model training.
- Features like 'Route' and 'Additional_Info' are dropped as they contain redundant or irrelevant information.
- Duration information is processed to extract hours and minutes separately for better feature representation.

Feature Selection

- Feature importance is determined using the ExtraTreesRegressor algorithm to identify the most influential features for predicting flight prices.



Correlation plot showing the relationship between different features and flight prices.

Model Development

- The random forest regression algorithm is chosen for modeling due to its ability to handle non-linear relationships and handle categorical data effectively.
- The model is trained on the training dataset and evaluated on the testing dataset using metrics like Mean Absolute Error (MAE), Mean

Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared.

Hyperparameter Tuning

- RandomizedSearchCV is employed to search for the optimal hyperparameters for the random forest regressor.
- Hyperparameters like the number of estimators, maximum depth, and minimum samples split are tuned to improve model performance.

Model Evaluation

- The tuned model is evaluated using the testing dataset, and performance metrics are calculated to assess its predictive accuracy.
- Visualizations such as density plots and scatter plots are used to analyze the distribution of predicted versus actual flight prices.

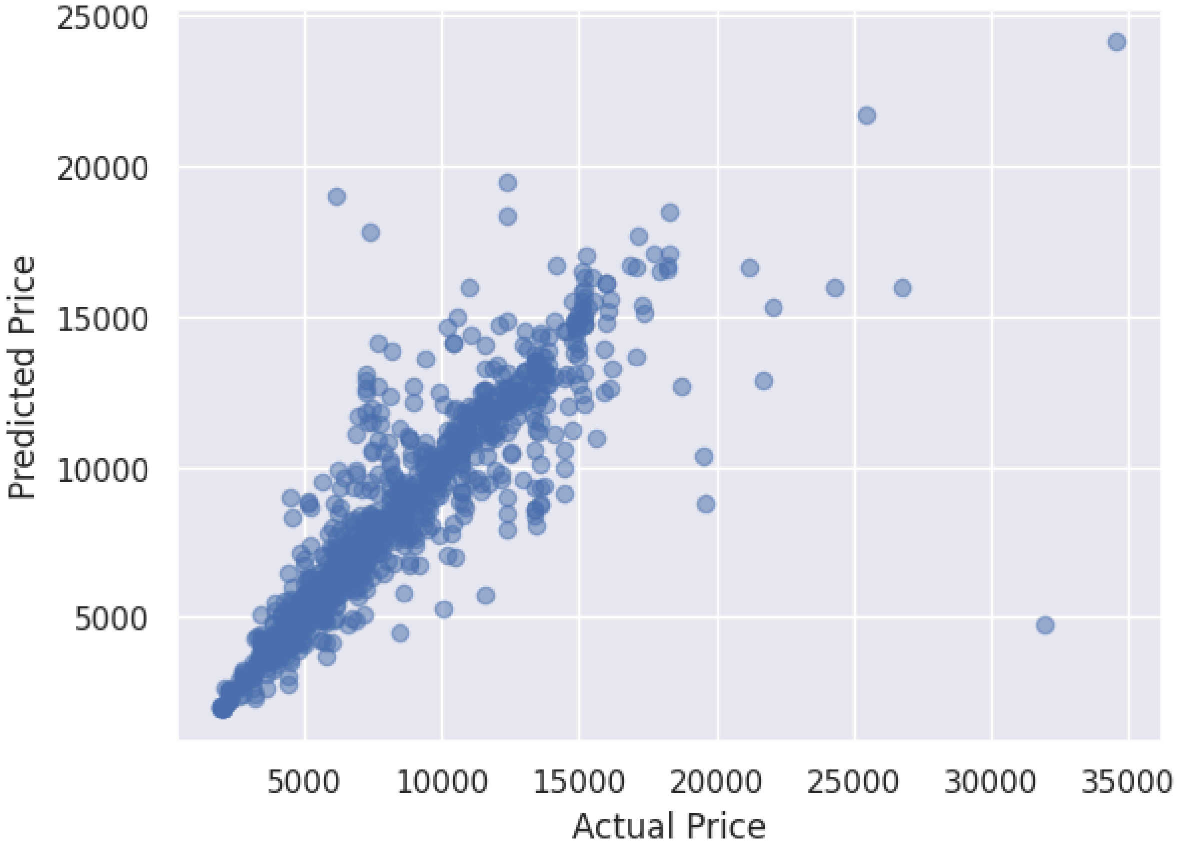
Conclusion

- The developed random forest regression model demonstrates promising performance in predicting flight prices.
- The model's accuracy and reliability can be further improved through fine-tuning of hyperparameters and incorporating additional relevant features.
- Overall, the project provides valuable insights into flight price prediction using machine learning techniques, which can benefit both customers and airline companies in planning and decision-making processes.

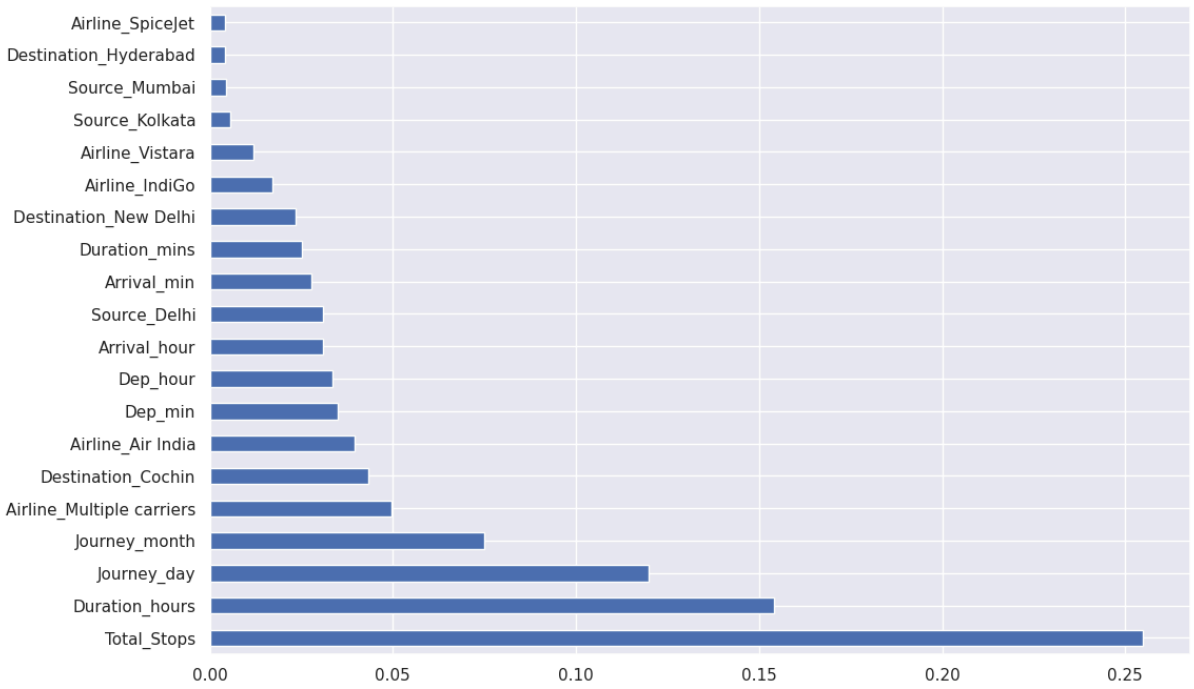
Future Scope

- Future enhancements to the project could involve exploring other machine learning algorithms, such as gradient boosting or neural networks, to compare their performance with the random forest approach.

- Incorporating real-time data streams and external factors like weather conditions and economic indicators could enhance the model's predictive capabilities.
- Deployment of the trained model into production environments for real-world applications, such as online flight booking platforms, could be pursued to provide users with accurate and dynamic pricing information.



Scatter plot showing the distribution of predicted versus actual flight prices for model evaluation.



Feature importance plot highlighting the most influential features for predicting flight prices.