Flight Price Prediction in India
Section 01 - Group 4

Team Members

- Emre Karataş 22001641
- Gökay Özbay 21901888
- Mustafa Kütükcü 21902972
- Ege Berkay Karagenç 22002799
- Erim Berke Salman 22001964

Description of the Dataset

This dataset contains information about flight booking options from the website Easemytrip for flight travel between India's five metro cities. The dataset can be found on Kaggle as "*Flight Price Prediction*" [1], and we will use a file named 'Clean_Dataset.csv' among the three CSV files. It consists of 300261 flight entries, and the data includes airline, flight, source and destination city, departure and arrival time, stops, class, duration, days left, and price [1].

Problem Definition

Firstly, we will initiate our project with an exploratory data analysis (EDA) to understand the nuances and complexities of the flight booking dataset. Given that we are dealing with various features ranging from categorical to continuous variables, we might perform feature engineering techniques to capture nonlinear relationships among variables.

Our primary objective is to predict flight prices accurately. To achieve this, we plan to start with simpler machine learning models, specifically Linear Regression, to establish a starter point for prediction. This will serve as our fundamental understanding of how individual features relate to flight prices. We may extend to more complex models depending on the complex relationships we understood during EDA.

If the dataset shows high collinearity or a linear model doesn't sufficiently capture the data's complexity, we will explore regularisation techniques like Ridge and Lasso Regression. We also plan to implement ensemble methods like Gradient Boosting, XGBoost, and the Random Forest algorithm to capture complicated data patterns that simpler models might miss. We consider implementing Feed Forward Neural Network (FNN) or Recurrent Neural Network (RNN) as deep learning models since we have time series data.

To evaluate the performance of our models, we will use metrics commonly used in regression tasks, such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared values. These metrics will give us a clear understanding of how well our models perform and where they might be lacking.

Lastly, we plan to split the dataset into 80% for training and 10% for testing and validation.

References

[1] S. Bathwal, "Flight price prediction," Kaggle, https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction/data (accessed Oct. 16, 2023).