Technical Report I Proposal For Flight Prices Prediction

Mercy Mulambia

January 4, 2024

Problem description

Air travel has become a necessary part of the modern way of life as more and more people search for faster ways to get from one place to another. The dataset used in this study was obtained from Kaggle.com, an open-source website (Bathwal, 2022). 'Ease My Trip' is an online platform that allows travellers to purchase airline tickets, therefore prospective travellers use it to do so. The 'Ease My Trip' website provided the flight booking dataset, which will be analysed. The cost of airline tickets might fluctuate regularly depending on several factors, including the destination, length, and timing of the flights, in addition to certain events and occasions like holidays and festivities.

By predicting pricing patterns over time, determining the right prices for certain routes, and enabling the airline to compare upcoming prices with those of competitors, they can gain a competitive edge. These are some ways that airlines might optimise their pricing strategy. This technique will help people understand the trends that prices follow and will also provide them with an estimated price value that they can use to save money before buying airline tickets.

The article 'Flight Fare Prediction Using Machine Learning', investigated whether machine learning techniques can be applied to predict airline ticket costs (Sarao and Samanta, 2022). The study explores several supervised learning methods, such

as Support Vector Machines, Logistic Regression, Classification Tree, Naive Bayes, and SoftMax Regression. Another paper is 'Using Spark Machine Learning Models to Perform Predictive Analysis on Flight Ticket Pricing Data' (Wong et al., 2023). Comparing machine learning algorithms for predicting airline ticket prices was the aim of this paper's investigation.

Goal

This project aims to predict flight prices and execute some statistical hypothesis tests to extract useful information. Here are some hypothesis tests that will be executed;

- 'Does the price vary with Airlines?'
- 'How does the ticket price vary between Economy and Business class?'
- 'How does the price change with the change in Source or Destination?'
- 'Does ticket price change based on the departure time and arrival time?' (Bathwal, 2022)

The Gradient Boost Tree algorithm demonstrated the highest accuracy for price prediction, out of the four regression algorithms tested in the second paper (Wong et al., 2023). The two predictive models that will be conducted in this study are the Gradient Boost Tree algorithm and Random Forest Regressor (Pedregosa et al., 2011). Using the two algorithms for this study, the results will be compared to the 'Wong et al.,' paper, to see how well it predicts flight prices.

Description of your data

Two data sets were gathered: one set concerned economy class tickets and the other concerned business class tickets. The website yielded a total of 300261 unique flight booking possibilities. 50 days of data collection were conducted in 2022, from February 11 to March 31. The data sets were merged to create a massive dataset.

The dataset includes flight booking possibilities for travel between India's top 6 metropolises, sourced from the website EaseMyTrip. The cleaned dataset contains 11 features and 300261 data points (Bathwal, 2022).

The dataset still requires some changes before beginning the analysis, such as getting rid of the column called 'Unnamed: 0'. Another requirement is to change the string variables into integers for example, the column called 'Class', would have either economy or business class. This would have to be changed to 0 or 1, 0 representing economy and 1 representing business class. Several columns require this change to begin the analysis.

Methodology

A Tentative Data Analytic Pipeline:

1. Define Objectives
2. Data Collection/ Data Cleaning
3. Exploratory Data Analysis
4. Model Selection/ Model Training
5. Model Evaluation/ Interpret Results

Python and Apache Spark will be used to analyse this dataset and predict the results. The main two predictive models are the Gradient Boost Tree algorithm and Random Forest Regressor. Decision Tree Regression will be considered as well as AdaBoost Regressor.

References

Bathwal, S. (2022), 'Flight price prediction', *https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction/data* .

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011), 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research* **12**, 2825–2830.

Sarao, P. and Samanta, P. (2022), 'Flight fare prediction using machine learning', *Available at SSRN 4269263* .

Wong, P., Thant, P., Yadav, P., Antaliya, R. and Woo, J. (2023), 'Using spark machine learning models to perform predictive analysis on flight ticket pricing data', *arXiv preprint arXiv:2310.07787* .