

Flight Fare Prediction using Machine Learning Approaches

Ridima Verma

PhD Research Scholar

School of Engineering and Applied Sciences, Ahmedabad University
Gujarat, India.

Abstract—Flight fare prediction is an important problem in the airline industry, as it affects the decisions of both consumers and providers. In recent years, machine learning approaches have been widely adopted to address this challenge. In this study, a comprehensive review of various machine learning techniques is presented, including linear regression, decision trees, random forests, and neural networks, for flight fare prediction. I have discussed the advantages and limitations of each approach and compared their performance on a real-world data set. The results demonstrate that machine learning approaches can significantly improve the accuracy of flight fare prediction compared to traditional methods. Moreover, I have shown that ensemble methods, such as random forests, can further enhance the performance of individual models. The findings provide useful insights for airline companies and customers to make informed decisions about ticket pricing and purchasing.

Index Terms—Flight fare prediction, Machine learning, Regression, Feature engineering, Data Analysis, Accuracy

I. INTRODUCTION

Most individuals have likely experienced purchasing an airplane ticket and observing how the cost changes based on demand. Airline companies use intricate techniques to determine flight prices, which tend to rise over time. The pricing strategy often varies based on factors such as the time of day or the season. While airlines aim to maximize profits, customers seek lower fares, and therefore, they usually purchase tickets well in advance of the departure date. Many studies have been conducted to predict airline ticket prices. Ginni et al. [1] used Partial Least Square Regression to forecast the optimal time to buy airline tickets with 75.3% accuracy. Huang et al. [2] predicted ticket sales income using Artificial Neural Networks and Genetic Algorithms, resulting in a mean absolute percentage error of 9.11%. M. Papadakis et al. [3] attempted to predict ticket prices using the Ripple Down Rule Learner, logistic regression, and Support Vector Machine. Their accuracy rates were 74.5%, 69.9%, and 69.4%, respectively. Janssen et al. [4] developed a best-fit model to provide unbiased information to travelers on whether to purchase a ticket or wait for a better price, which resulted in a prediction of "true bargains" using linear quantile mixed

models. However, this research only applies to one type of ticket for economy class and only on one-way flights from John F Kennedy Airport to San Francisco Airport.

Yang and colleagues [5] utilized various machine learning models, including Linear Regression, Naive Bayes, SoftMax regression, and SVM, to predict prices based on over 9,000 data points with six attributes. SVM had the highest accuracy at 80.6%, while LR had the lowest training error rate at approximately 22.9%. Lantseva and team [6] used eight different algorithms to forecast prices, with Bagging regression achieving the best accuracy of 87%, followed by Random Forest at 85%. United States Airlines [7-8] developed a model that considered market demand, operating expenses, distance, and airport status to predict airline selection and route changes. The model focused on maximizing profits and was better at predicting market exit than entry. Boruaah et al. [9] proposed a model using Bayesian estimation and Kalman filtering to calculate ticket prices for selected aircraft. William et al. [10] used ARMA and random forest regression, while Chen et al. [11] employed incremental learning to predict ticket prices. Juhar et al. [12] presented a regression model that calculated the cost per kilometer, with advanced ticket purchases being more cost-effective for overseas flights but not necessarily for domestic flights.

II. METHODOLOGY

To develop a new airline pricing prediction model, extensive data collection is necessary, which includes information such as airline names, flight sources and destinations, and routes. Two datasets with different attributes are used to build the model, both of which are in CSV format and contain details about relevant features. The most crucial step in this process is selecting the features that will impact the flight pricing model. To create an effective machine learning model, data gathering, preparation, model selection, training, hyperparameter tuning, and prediction are all essential steps. By gathering data, previous patterns can be analyzed to identify recurring trends. Data preparation involves acquiring the dataset, importing necessary libraries, identifying and managing missing data, encoding category data, and dividing the dataset. Model selection involves choosing the most appropriate model to use. Finally, the collected data is split into two subgroups: the training set,

which is used to develop the model, and the test set, which is used to evaluate the model's performance.

A. Overview of Data

The data set includes 30,000 data points and features as airline flight source city departure time stops arrival time destination city class duration days left price

B. Cleaning and Pre-processing of Data

Data preparation is one of the most challenging aspects of machine learning. Raw data must be transformed into a format that is suitable for modeling. Often, the raw data cannot be used directly because machine learning algorithms require numerical data. Furthermore, certain machine-learning approaches have restrictions on the type of data that can be used. Therefore, the data may need to be updated to account for statistical noise and errors, and some variables may need to be altered or encoded before they can be used in a machine-learning approach. Additionally, complex nonlinear relationships can be created using this information.

To prepare the data set for modeling, all the null or missing values were removed, and converted the data types from item sorts to numerical types. Also, the splitting of some attributes was done to make them more useful and extracted hours, minutes, days, and months using the appropriate syntax. Some of the data in the data set were nominal categorical data and ordinal specific data. To handle this, we used one hot encoding and label encoding to transform the data. All of the steps were completed to ensure that the data is suitable for training and testing in accordance with the requirements of the model. Overall, data preparation is a crucial and time-consuming process in machine learning that must be performed meticulously to ensure the accuracy and validity of the results.

III. MACHINE LEARNING MODELS

Machine learning is a highly-discussed area within computer science and engineering, and has a wide range of applications across multiple industries. It equips machines with intelligence through the use of tools, strategies, and algorithms. The modeling tools available through machine learning are particularly powerful and can be trained using a learning process that involves collecting data that pertains to a specific problem and continuously adapting to new, previously unknown data. These introduce a few techniques for predicting aircraft ticket pricing. Approaches such as linear regression, lasso regression, decision tree, and decision tree regressor are implemented on the data set.

IV. CONCLUSION

In conclusion, this report explored the use of various regression techniques to predict flight fares. The methods

| Approach | Accuracy |
|-------------------------|----------|
| Linear Regression | 89.375% |
| Lasso Regression | 89.374% |
| Decision Tree | 91.293% |
| Decision Tree Regressor | 92.053% |

Fig. 1. Model-Accuracy Table

investigated were linear regression, lasso regression, decision tree, and decision tree regressor. The data used in this study consisted of a combination of factors such as departure time, arrival time, airline, route, and number of stops, among others.

The results of the analysis revealed that all four methods produced reasonable predictions of flight fares. However, some methods performed better than others. For instance, the lasso regression model was able to identify the most important features in predicting flight fares and produced a more accurate prediction than the other models.

The results of this study demonstrate the potential of regression techniques for predicting flight fares. Depending on the specific needs of the user, a combination of these techniques could be used to produce the most accurate predictions. However, further research could be conducted to determine the best combination of methods and predictors for more precise predictions.

V. REFERENCES

- [1] W. Groves and M. Gini, "A regression model for predicting optimal purchase timing for airline tickets," Technical Report 11-025, University of Minnesota, Minneapolis, 2011.
- [2] H.-C. Huang, "A hybrid neural network prediction model of air ticket sales," *Telkomnika Indonesian Journal of Electrical Engineering*, vol. 11, no. 11, pp. 6413–6419, 2013.
- [3] M. Papadakis, "Predicting Airfare Prices," 2014. Clerk Maxwell.
- [4] T. Janssen, "A linear quantile mixed regression model for prediction of airline ticket prices," in *A Treatise on Electricity and Magnetism 3rd ed.*, vol. 2, 2014, pp. 68–73.
- [5] R. Ren, Y. Yang and S. Yuan, "Prediction of the airline ticket price," Technical Report, Stanford University, 2014.
- [6] Lantseva, Anastasia, Mukhina, Ksenia, Nikishova, Anna, Ivanov, Sergey, Knyazkov and Konstantin, "Data-driven Modeling of Airlines Pricing," *Procedia Computer Science*, vol. 66, pp. 267–276, 2015.
- [7] Sha, Z., Moolchandani, K., Panchal, J., Delaurentis, D. (2015). Modeling airline's decisions on route selection using discrete choice models - data and supplementary material.
- [8] Sha, Z., Moolchandani, K., Panchal, J., DeLaurentis, D. (2016). Modeling airlines' decisions on city-pair route selection using discrete choice models. *Journal of Air Transportation*, 24, 1–11.

- [9] A. Boruah, K. Baruah, B. Das, M. Das, and N. Gohain, "A Bayesian Approach for Flight Fare Prediction Based on Kalman Filter," in Progress in Advanced Computing and Intelligent Engineering, Singapore, 2019, pp. 191-203
- [10] William Groves and Maria Gini, "A regression model for predicting optimal purchase timing for airline tickets.," Technical report, The University of Minnesota, Minneapolis, USA, Report number 11-025, 2018.
- [11] Yiwei Chen and F. Vivek Farias, " Robust Dynamic Pricing With Strategic Customers," Mathematics of Operations Research 43, pp. 1119-1142, 2019.
- [12] D. Tanouz, R. R. Subramanian, D. Eswar, G. V. P. Reddy, A. R. Kumar, and C. V. N. M. Praneeth, "Credit Card Fraud Detection Using Machine Learning," in 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 2021.
- [13] Saran Jaya Thilak et. al, "A Comparison Between Machine Learning Models for Airticket Price Prediction," 2022 3rd International Informatics and Software Engineering Conference (IISEC)