Grenze International Journal of Engineering and Technology, June Issue

Machine Learning-based Flight Fare Prediction in the Indian Market

Jayashree[1], Padmanayana[2], Samridhi Srivastava[3], Achinth Krishna Nittur[4], Preethi Salian[5] and Shiya Sah[6]
[1,3,4,6]Dept. of Computer Science and Engineering, Nitte (Deemed to be University), Nitte Meenakshi Institute of Technology, Bangalore, India
Email: jayashree.cse@nmit.ac.in, samridhi0802@gmail.com, achinthkrishna@gmail.com, shiyasahjnk@gmail.com
[2]Department of Artificial Intelligence and Data Science, A.J. Institute of Engineering and Technology, Mangalore
Email: padmanayana@ajiet.edu.in
[5]Dept. of Information Science and Engineering, Nitte (Deemed to be University), NMAMIT, Nitte, Udupi, Bangalore, India
Email: preethi.salian@nitte.edu.in

***Abstract*— With the rising popularity of air travel in India, ac- curately predicting flight ticket prices becomes crucial for budget- con-scious travelers. Airlines employ complex pricing strategies, making it challenging for passengers to find the optimal time to purchase tickets. This paper proposes a machine learning- based model to predict flight fares in the Indian market. The model leverages a dataset of domestic and international flights in 2019 to identify key factors that influence ticket prices. We employ exploratory data analysis (EDA) techniques for data cleaning, feature engineering, and handling categorical variables. The Random Forest algorithm is then used for price prediction, with hyperparameter tuning to optimize performance. This research contributes to the development of a user-friendly system that empowers travelers to make informed decisions regarding flight purchases. Develop a model that can accurately forecast the price of a future flight ticket based on historical data and recurrent factors. By leveraging machine learning, particularly Ran-dom Forest regression with hyperparameter tuning, researchers can develop effective flight fare prediction models. This encourages travelers to make informed booking decisions and assists the aviation industry in optimizing pricing strategies.**

***Index Terms*— Machine Learning Algorithms, airfare, supervised learning, predictions, flight, Linear Regression, Artificial Neural Network, Random Forest.**

I. Introduction

Everyone is aware that vacations always need an essential vacation, and organizing the itinerary for such a trip can take a lot of effort. Due to the global expansion of the Internet and e-commerce, the commercial aviation industry has expe-rienced remarkable growth and has transformed into a regu-lated marketplace. Therefore, a variety of techniques, includ-ing financial marketing, social considerations, and profiling of customers, are utilized to set ticket fairs for airline revenue management. Airfares are usually cheaper when purchased months in advance; but, when purchased quickly, prices are typically more. However, there are many more elements that affect flight fare in addition to the number of days or hours till departure. Due to the level of complexity in the aviation sector, consumers find it hard to secure the best and lowest ticket bargain. pricing strategy.

Grenze ID: 01.GIJET.12.2.57_23

To address this issue, deep learning and machine learning-based models and techniques have been developed, and much research is currently under-way. This article describes a flight fare prediction system based on machine learning that uses random forest regression to forecast the cost of airline tickets. The system's experimental analysis is also explored, along with several features that affect costs [1]–[4].

The main contributions of the work as follows:

- Predicting Flight Prices Accurately using Machine Learning A random forest model is successfully used in your project to forecast flight prices depending on a number of variables, including source, destination, and arrival and departure dates. By offering accurate pricing projections, the model helps passengers make better decisions and make more economical travel plans.
- An Easy-to-Use Interface for Improved Accessibility Users may enter their trip information and get real-time fare forecasts using a specialized user interface. This streamlines the reservation process and enables a broader audience to use machine learning-based predictions without the need for technical know-how.
- Feature engineering, preprocessing, and data collection In order to enhance prediction, the research entailed collecting and preprocessing airplane ticket data, identifying important price determinants, and extracting useful characteristics.

A study of the literature that examined technical publications and some of the most recent models and systems was included in Section II. Variations in the fea-tures considered are also charted. In Section III, the features and workflow of the suggested system are thoroughly de-scribed. A detailed explanation of the recommended system's features and workflow is provided. Section IV contains the results as well as additional comparisons between the find-ings. Section V presents the findings. Prospective develop-ments for additional research are covered in Section VI.

II. Related Work

The four phases of the technique are as follows: choosing and assessing an appropriate ML Regression model; col-lecting data from the airline fare prediction dataset by MH; and influencing flight pricing. The effectiveness of the sys-tem depends on its integrity, which enables users to precise-ly represent the market niches wherever costs were set es-tablished differently. A former airline planning executive and consultant claims that ticket prices on airlines nowadays are set ”as much as the customer and market will bear.” To assist them change prices, airlines also profile their clients [5], [6].

This usually entails splitting travelers into two categories: those traveling for pleasure or work. Additionally, the prices of each group differ significantly. Most of the research on predicting airfare prices has been done at the national level or in a particular market. However, there is still a dearth of research on market segments. The market/airport pair that lies between the flight origin and destination is what we re-fer to as a market segment [7], [8]. The eight characteristics of the airline dataset were as fol-lows: source, destination, number of stops, airline type, departure and arrival times, and other data. Regression machine learning models, such as the LGBM Regressor, Random Forest Regression Tree, and Decision Tree Regres-sor, were used by the authors to conduct prediction. An en-semble learning approach is used for regression in the su-pervised learning algorithm Random Forest Regression. To provide a prediction that is more accurate than one from a single model, the ensemble learning method integrates pre-dictions from several machine learning algorithms [9]–[11].

An overview of a deep learning and social media data-driven airline ticket pricing prediction algorithm is shown here. The variables influencing ticket prices are discussed by the writers together with the state of airline ticket pric-ing now. During the training phase, a Random Forest builds many decision trees, with the meaning of the clas-ses serving as each tree's forecast. An accurate and po-tent regression model is the Random Forest Regres-sion. It typically does well on a wide range of issues, par-ticularly those involving characteristics with non-linear relationships. The Probabilistic Forest Model is utilized for development because, with a R squared score of 0.91, it has a good degree of accuracy and performs better on data performance than other models like the Neural Network and LGBM Decision Tree Regressor [12], [13].

III. Design and Methodology

A. The suggested architecture for our project en-tails the following procedures

Data Collection

In machine learning, data collection is the process of locating, gathering, and collecting the information required to create, evaluate, and validate a model. This stage is essential to the implementation process. Here, in-formation is gathered from an imported flight fare dataset from Kaggle. Both numerical and category data are includ-ed in the collection. The source, destination, kind of airline, extra information, and arrival and departure dates are among the categorical data; the number of stops is among the numerical data. This enormous dataset consists of 10683 rows and 11 columns, each of which represents a fea-ture.

Data Preprocessing
The act of cleaning data to make it suitable for model training and testing is known as data preparation. We can give our information more signifi-cance for model training by doing this. Getting ready for data includes transforming, cleaning, and setting up the data to be analyzed. The data preprocessing substeps that are in-volved include:

- Data Cleaning: This step entails removing any potential duplicates as well as any null or missing values.
- Feature selection and engineering: Our model's features are extracted in this step, and all pertinent features are then used to train the model. The departure, arrival, and trip dates are all listed in columns in the dataset. As the voyage day, journey month, departure hour, departure minutes, and arrival hour, all the numerical values are identified.
- Given that the dataset includes both numerical and categor-ical features, the categorical values were converted to nu-merical values using the "label encoding" approach for ordi-nal categorical data and the "on hot encoding" method for nominal categorical data. The dataset contains the follow-ing category variables: flight, origin, destination, path, total number of stops, and more information. Decision Tree Re-gressor: The Sklearn Tree Module's Decision Tree Regressor class is used to train and forecast models of regression [14], [15].
- The method, which is based on decision trees, divides the source data iteratively according to the attribute values that were provided to produce a structure resembling a tree. The output that is produced when test data is given into the model is the mean amount of all the information points in the subsection. This process determines if the data point is part of the decision tree. To make decisions, it first chooses independent variables as decision nodes from the dataset. After that, the entire data set is split up into sections.

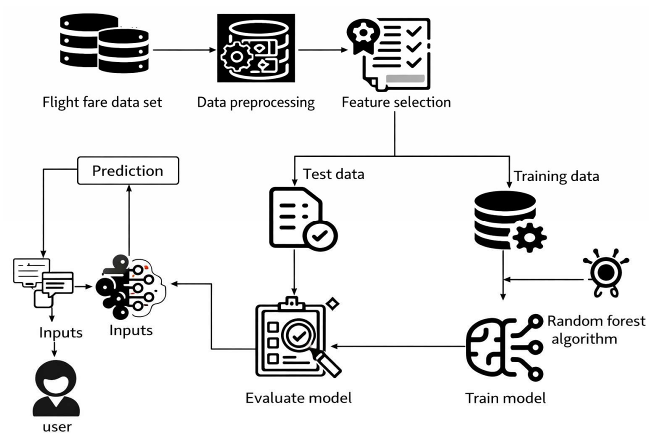

Figure 1: Model architecture

Data Splitting
In order to prepare our data for testing and training, we must divide it into two pieces. Eighty percent of the data was used in the training process of the Random Forest regressor model. The algorithms for machine learning are:

- LGBM Regressor Light Gradient Boosted Machine is what LGBM stands for. Regression, classification, and ranking are just a few of the machine learning tasks that can be per-formed with this gradient boosting framework, which is based on decision trees. Regression model training and pre-diction are possible with the help of the lightgbm package's LGBM Regressor class.
- Randon Forest Regressor Regression tasks are carried out by the Random Forest regressor using several decision trees. It serves as an illustration of group learning. A supervised learning method called Random Forest employs an ensem-ble learning strategy for regression and classification. Deci-sion trees exhibit sensitivity to the data used for training. The resulting decision tree and, thus, the predictions may differ significantly if the training data is altered. We utilize Ran-dom Forest to overcome these drawbacks. Because they cannot reverse their split, decision trees also tend to discover local optimal values, have a high risk of overfitting and are operationally costly to develop.

Model evaluation

This is a crucial phase in our project, as it aids in evaluating the efficacy and precision of our model. The model is evaluated using test data. In this case, we evaluated the model using cross-validation. The data are split using this method into k-subsets, or folds. After training on k-1 folds, the model is tested on the re-maining folds. Every fold is used as a test set once, and this process is repeated k times. The result is given as the aver-age performance over all k-folds. The following measures are employed in the evaluation of the model: Root Mean Squared Error (RSME): For a regression prob-lem, it provides the average squared difference between the actual and anticipated values, expressed as a root. The ab-solute difference between the actual and anticipated val-ues is indicated by the Mean Absolute Error (MAE). A model that performs better is indicated by greater negative mean values. R-Squared: This statistic compares the re-gression model's fit to the data to a baseline model that consistently predicts the mean value. It displays the extent to which the model explains the variation in the data.

Model Architecture

Essential elements of the architecture are designed to maximize prediction as shown in Fig. 1. Our suggested system examines past data to find trends, seasonal patterns, and other information that affects airline prices.

IV. Implementation and Result Analysis

The process of developing and implementing a machine learning model for ticket price prediction is explored in this paper as shown in Fig.2 . The Pandas library is used to read the dataset, which can be found in CSV, JSON, TXT, or Excel formats. To deal with missing values and guarantee data quality, preprocessing is done. The association between different variables is examined using a heat map, which offers information on important contributing elements. Flask, a lightweight Python web framework, is then used to install the machine learning model. This allows users to enter trip information, like the date of departure and arrival, the source, the destination, the number of stops, and the airline type, in order to effectively estimate ticket prices. Precess involves following stages as shown below,

- Read the dataset: The format of our dataset may be in.csv,.json,.txt, or excel files. With the aid of pandas, we can read the dataset.
- Preprocessing of data: The process of df.isnull() is performed to confirm that there are no values. We ag-gregate the null values using the sum () function. We found that our dataset contained two null values. So, we begin by looking into the data.Utilizing a Heat Map to examine the relationship
- In this case, a heat map is being used to examine the correlation. It presents the data as two-dimensional colored maps by utilizing various colour combina-tions. To depict the relationship variables, it will be plotted on both axes rather than being represented by numbers.
- Flask was used in the deployment of the machine learning model. A web framework in Python called Flask makes it simple and quick to create lightweight, adaptable online apps. The website gathers data and forecasts the ticket price. It gathers data such as the date of departure and arrival, the source and destina-tion, and the number.

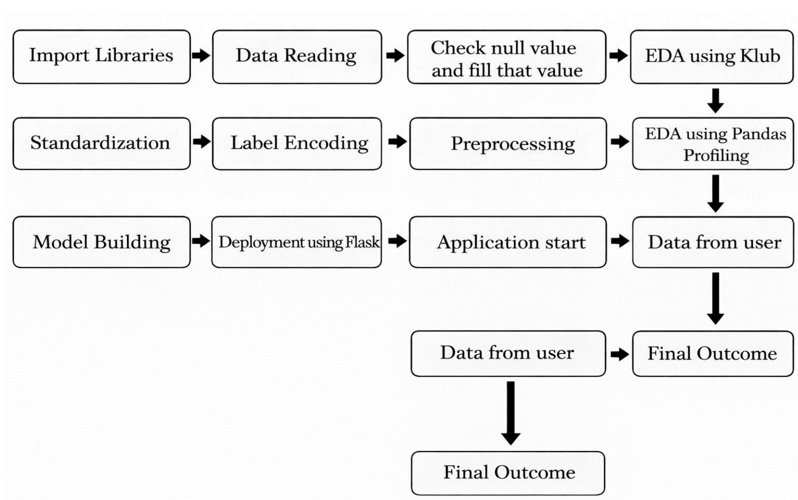

Figure 2: Processes involved in Flight prediction

V. Discussion

The significance of several factors in forecasting ticket prices is shown in the Fig. 3. While the y-axis contains several parameters including city names, duration, trip details, and airline kinds, the x-axis displays the importance score. Total Stops is the most important of these, indicating that the number of layovers has a big influence on the cost of the ticket. The time of travel and the airline's reputation also affect pricing, as evidenced by the other important criteria of Journey Day, Journey Hours, and Airline Type.
However, features like Delhi, Hyderabad, and Mumbai have the least effect, suggesting that the city of arrival or departure by itself does not have a major influence on ticket costs. In a similar vein, total travel details have a greater influence than duration in minutes and arrival-related aspects. The visual aid is beneficial.

VI. Conclusion

To sum up, our project's primary goal is to forecast pricing for flight fares using machine learning. For the entire process, including the arrival and departure dates as well as the source and destination, we have devel-oped a user interface. Our Our machine learning-based aircraft fare prediction project has effectively generat-ed a dependable and user-friendly solution. Flight ticket data was gathered, preprocessed, and characteristics were identified. After that, a strong random forest model was trained, and its effectiveness was evaluated. Using user input, our online application forecasts flight costs so that travelers may make educated decision.

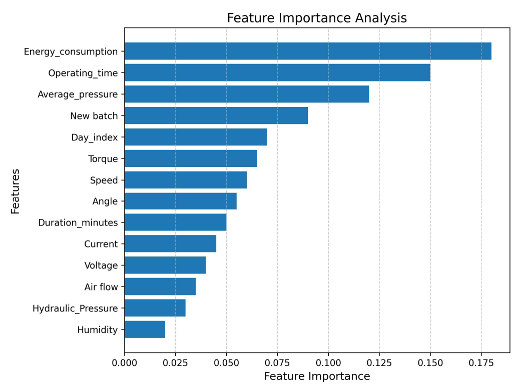

Figure 3: Result analysis of Flight prediction

Future work involves Optimal date recommendation: This refers to advising consumers of the day when flight costs will be at their lowest. Real-Time Updates: Managing data in real-time for dynamic pricing modifications according to demand, weather, and airline regulations. Integration: Providing price as a value-added service by collaborating with online booking platforms, airlines, and travel agencies.

References

[1] K. Tziridis, Th. Kalampokas, G. A. Papakostas, "Air- fare Prices Prediction Using Machine Learning Tech- niques", 25th European Signal Processing Conference (EUSIPCO), IEEE, October 26, 2017.
[2] N. Brown and J. Taylor, Air Fare: Stories, Poems Essays on Flight. Sarabande Books, 2004.
[3] J. C. Lok, Prediction Factors Influence Airline Fuel Price Changing Reasons. 2018.
[4] A. Kakaraparthi and V. Karthick, "A secure and cost-effective platform for employee management system using lightweight standalone framework over Diffie Hellman's key exchange algorithm," ECS Trans., vol. 107, no. 1, pp. 13663–13674
[5] Apr.2022.Wikipedia,"MeanSquarederror", Availa-ble:https://en.wikipedia.org/wiki/Meansquarederror.
[6] G. Ataman and S. Kahraman, "Stock Market Prediction in Brics Countries Using Linear Regression and Artificial Neural Network Hybrid Models," The Singapore Economic Review. pp. 1–19, 2021. doi: 10.1142/s0217590821500521.
[7] B. Panwar, G. Dhuriya, P. Johri, S. S. Yadav, and N. Gaur, "Stock Market Prediction Using Linear Regression and SVM," 2021 International Conference on Advance Compu-ting and Innovative Technologies in Engineering (ICA-CITE).2021.doi:10.1109/icacite51222.2021.9404733
[8] P. Purey and A. Patidar, "Stock Market Close Price Predic- tion Using Neural Network and Regression Analysis," Inter- national Journal of Computer Sciences and Engineering, vol. 6, no. 8. pp. 266–271, 2018. doi: 10.26438/ijcse/v6i8.266271.
[9] Panigrahi, A., Sharma, R., Chakravarty, S., Paikaray, B. and Bhoyar, H., 2022. Flight Price Prediction Using Machine Learning. In ACI@ ISIC (pp. 172-178).

[10] Sood, Yadunandan, and Pradeep Kumar Gupta. ”Flight Price Prediction Using Machine Learning.” (2022).
[11] K. Imanaka and C. Sakama, ”Predicting Air Ticket Demand using Deep Neural Networks,” 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 2021, pp. 2901-2908, doi: 10.1109/BigData52589.2021.9671440.
[12] Degife, Worku Abebe, and Bor-Shen Lin. ”Deep-Learning-Powered GRU Model for Flight Ticket Fare Forecasting.” Applied Sciences 13.10 (2023): 6032.
[13] Jayashree, T. Shiva Prakash, K.R. Venugopal, Unsupervised feature extraction based on uncorrelated approach, Information Sciences, Volume 666, 2024, 120447, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2024.120447.
[14] Jayashree, Shivaprakash, T. ”Optimal Value for Number of Clusters in a Dataset for Clustering Algorithm, Artificial Intelligence and Sustainable Computing. Algorithms for Intelligent Systems”. Springer, Singapore.
[15] Savio, A. Sanchith, et al. ”Stock Market Analysis using Supervised using Machine Learning Technique”. Grenze International Journal of Engineering and Technology (GIJET) 9.2 2023.