ISSN:2731-538X | E-ISSN:2731-5398
Vol. 19 No. 01 (2025)

# PREDICTION OF AIRLINE PRICES USING MACHINE LEARNING MODELS

**Ashok Kumar[1], Ritu Kumari[2] and Pratibha Srivastava[1*]**
[1]Department of Statistics, University of Lucknow, Lucknow, Uttar Pradesh, India.
[2]Department of Mathematics and Statistics, Integral University, Lucknow, Uttar Pradesh, India.
*Corresponding author, email: pratibha.srivastav07@gmail.com

## ABSTRACT

The aviation industry has experienced significant growth and transformation in recent years, driven by factors such as increased global connectivity, rising disposable incomes, and advancements in technology. Nowadays, airline ticket prices can vary dynamically for the same flight. From the passenger's perspective, if they want to save money, the appropriate model is required that predicts the ticket prices.

This study explores advanced machine learning techniques such as random forest, gradient boosting, and XGBoost algorithms in predicting the prices of flight tickets based on the features such as departure and arrival times, number of stops, and route specifics. The XGBoost with missing values outperforms the other models with an $R^2$ value of 0.864, indicating a high accuracy in capturing the variability in prices for Indian flights. The findings highlight the potential of machine learning models to enhance pricing strategies in the aviation industry, offering significant benefits for both airlines and passengers.

**Keywords** - Prediction, machine learning, random forest, XGBoost regressor, mean absolute error.

## INTRODUCTION

The demand for air travel has increased due to global connectivity, leading to intense competition among airlines to attract passengers and optimize their revenue. Passengers, on the other hand, are interested in low-cost flight tickets. It is well known that when tickets are booked months in advance, prices of flight tickets are often reasonable but when tickets are booked in a hurry, they are often higher prices. The point to be noted here is that the number of days/hours until departure isn't the only factor that decides flight fare, but several other factors such as departure and arrival times, number of stops, and route specifics. To overcome this challenge, machine learning and deep learning algorithms are used to predict flight prices.

Several studies have been conducted to develop accurate predictive models and gain insights into the factors influencing airline ticket prices. Ren at el. (2015) used linear regression, naive Bayes, softmax regression, and support vector machines to build a prediction model and classify ticket prices into five bins (60% to 80%, 80% to 100%, 100% to 120%, and etc.) to compare the relative values with the overall average price. They reported that the training error rate was close to 22.9% using the linear regression model. Janssen et al. [2014] developed an assumption model for the route from San Francisco to New York with already available data on flight fares for each day. The two important features were the day count from departure and which day of the week it was, whether it was weekday or weekend. Papadakis (2014) anticipated whether there would be a fall in airfare later on by addressing the issue as a classification task using Logistic Regression, Linear SVM and Ripple Down Rule Learner models. Groves and Gini (2011) used a partial least square (PLS) regression model to optimise the purchase of airline tickets. Kotsiantis (2013) applied artificial neural networks (ANN) and genetic algorithms (GA) to predict air ticket sales revenue from the travel agency.

Tziridis (2017) compared the performance of various machine learning algorithms, including artificial neural networks, linear regression, decision trees, and random forests, for predicting airfare prices. Their results showed that machine learning techniques can handle this prediction problem

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 19 No. 01 (2025)

with an accuracy of almost 88%. Subramanian (2022) applied feature selection algorithms along with hyperparameter methods to find the optimal model parameters and set of features for flight description in order to predict airfare price prediction. Panigarhi et al. (2022) have implemented the Artificial Neural Network (ANN), Linear Regression (LR), Decision Tree (DT), and Random Forest (RF) models. They have reported the values of root mean square (RMSE) and mean absolute percentage error (MAPE) for all the employed models. In this, decision tree models got the minimum RMSE and MAPE values while ANN got the high values compared to other models. Vaishnavi (2023) presented a machine learning-based flight fare prediction system that employs random forest regression to predict airline ticket prices. They also studied the various features that influence prices and conducted an experimental analysis of the proposed system.

This paper aims to use the random forest, gradient booster regressor (GBR), and extreme gradient boosting (XGBoost) regressor algorithms to predict airline prices in the Indian context. The rest of the paper is as follows: Section 2 describes the dataset and preprocessing of data to apply to machine learning models; Section 3 discusses the machine learning models used in the paper. Model evaluation matrices are also described in this section; Section 4 delves into the results and discussion, and Section 5 provides the conclusions.

## 1. Data Collection and Preprocessing

This section describes the proposed holistic approach focusing on the dataset used and the models selected to predict ticket prices. The dataset, feature descriptions and visualisation material are presented to highlight the level of competition and globalisation affection for ticket prices between destinations from different airline companies. The steps are as follows:

### 1.1 Data Collection

The assortment of the data that will be used to develop, test, and validate a machine learning model plays a crucial role in the implementation of the model. For this study, data is collected from the flight fare dataset imported from the Kaggle website (https://www.kaggle.com/). The collected dataset comprises information about different airlines in India. The dataset contains 10683 observations on 11 features or attributes. The features present in the dataset are the name of companies, date of travelling, origin, terminus, path of travelling, time of departure, time of arrival, travelling hours, total stoppage, additional info, and price.

### 1.2 Data Pre-processing

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. Real-world data generally contains noises, missing values, and may be in an unusable format that cannot be directly used for machine learning models. Data preprocessing involves cleaning, transforming, and preparing the data for applying the machine learning model. The sub-steps involved in the data preprocessing are:

- **Data Cleaning:** In this step, we removed the null and missing values from the dataset. If any duplicate value is there, then it will also be removed.
- **Formatting the Data:** Formatting the features is done, as well as labelling and encoding to convert category data to unique integer values. The data has also been normalized using *Standard Scaler* to ensure uniformity in scale across features.
- **Feature selection:** In this phase, the most informative features of a flight that determine the ticket prices are decided and extracted from the dataset to apply the machine learning models. As per the dataset, the features identified during feature extraction are listed in Table 1.

**Table.1** The List of Features Generated During Feature Extraction.

| lo. | ature Name | scription |
|---|---|---|
| 1 | rline | e name of the airline company |
| 2 | te of Journey | te of the journey |

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 19 No. 01 (2025)

| | | |
|---|---|---|
| 3 | urce city | ty from which the flight takes off |
| 4 | stination | e city where the flight will land |
| 5 | ute | ute of the flight |
| 6 | parture time | ores the information about the departure time |
| 7 | rival time | ores the information about the arrival time |
| 8 | ration | e overall amount of time it takes to travel between cities |
| 9 | tal stops | e number of stops between source and destination cities |

### 1.3 Splitting of Data

In this phase, we split the dataset into two parts training and testing datasets using an 80:20 ratio. From this, 80% of the data is used to train the model and 20% of the data is used to test the employed model.

## 2. Machine Learning Models

A model of machine learning is a set of programs that can be used to find the pattern and make a decision from a dataset. In this study, we have used random forest, gradient boosting regressor (GBR), and extreme gradient boosting (XGBoost) algorithms of machine learning to predict the ticket price of the flights. These models are defined below:

**3.1 Random Forest:** A random forest regressor algorithm is a supervised ensemble learning method that can be used for both classification and regression tasks. It combines the predictions from multiple decision trees, each of which produces its predictions. The basic idea behind this is to combine multiple decision trees to determine the final output rather than relying on individual decision trees. In the case of a regression problem, the final output is the mean of all the outputs whereas, for classification, the final output is based on the majority votes classifier. The random forest model is a type of additive model that makes predictions by combining decisions from a sequence of base models. More formally we can write this class of models as:

$$g(x) = f_0(x) + f_1(x) + f_2(x) + \cdots$$

where the final model $g(.)$ is the sum of simple base models $f_i(.)$.

Here, each base classifier is a simple decision tree. This broad technique of using multiple models to obtain better predictive performance is called model *ensembling*. In random forests, all the base models are constructed independently using a different subsample of the data. Random Forest Regressor can be used in a wide range of applications where the goal is to predict continuous numerical values. Its ability to handle high-dimensional data, capture complex relationships, and handle missing data and outliers make it a popular choice for many real-world problems.

**3.2 Gradient Booster:** The Gradient Booster (GB) regressor algorithm is an ensemble learning method for both classification and regression tasks. Gradient Boosting is a powerful machine learning algorithm that has many real-world applications in fields such as finance, healthcare, and e-commerce. Gradient boosting trains the model sequentially and each new model tries to minimize the loss function such as the mean squared error of the previous model using gradient descent. The predictions of the new model are then added to the ensemble, and the process is repeated until a stopping criterion is met.

Mathematically, let's assume we have a training set $(x_i, y_i)$ for $i = 1, 2, \ldots, n$, where $x_i$ are the input features and $y_i$ is the target variable. The goal is to learn a function $F(x)$ that maps the input features to the target variable. We start with an initial model $F_0(x)$, which is typically the mean of the target variable, then, for each iteration $m = 1, 2, \ldots, M$,

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 19 No. 01 (2025)

we compute the negative gradient.

$$g_i = -\left[\frac{\partial L\big(y_i, F(x_i)\big)}{\partial F(x_i)}\right]_{F=F_{m-1}(x)}$$

where $L(y_i, F(x_i))$ is the loss function and $F_{m-1}(x)$ is the model from the previous iteration. We then fit a weak learner $h_m(x)$ to the negative gradients $g_i$ using the input features $x_i$. The updated model is obtained by adding a scaled version of the new learner to the previous model:

$$F_m(x) = F_{m-1}(x) + \nu h_m(x)$$

where $\nu$ is the learning rate, which is a hyperparameter that controls the contribution of each new learner.

The process is repeated until a stopping criterion is met, such as a maximum number of iterations or a target value for the loss function.

**3.3 Extreme Gradient Boosting:** The Extreme Gradient Boosting (XGBoost) Regressor algorithm is a powerful machine learning algorithm in the ensemble learning category in regression. It started from a combination of gradient descent and boosting, called *Gradient Boosting Machine (GBM).* It is an optimized implementation of the GBR which offers several enhancements, making it a popular choice for predictive modelling**.** It uses decision trees as base learners and employs regularization techniques such as Lasso (L1) and Ridge (L2) regularization to prevent overfitting and enhance model generalization. It is known for its computational efficiency, feature importance analysis, and handling of missing values; hence it is applicable in various domains.

Regularization parameters of XGBoost are as follows:

- **Gamma ($\gamma$) :** minimum reduction of loss allowed for a split to occur. The higher the gamma, the fewer the splits.
- **Alpha ($\alpha$):** $L_1$ regularization on leaf weights, the larger the value, the more will the regularization, which causes many leaf weights in the base learner to go to 0.
- **Lambda ($\lambda$):** $L_2$ regularization on leaf weights, this is smoother than $L_1$ and causes leaf weights to smoothly decrease, unlike $L_1$, which enforces strong constraints on leaf weights.

The objective function of the XGBoost regressor is given by

$$\mathcal{L}(\phi) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(\mathrm{f_k})$$

where $\Omega(f) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2$,

If $\gamma$ is increased, the number of leaf nodes $(T)$ decreases, therefore, $\gamma$ penalizes $T$ and helps prevent the tree from becoming too complex.

**3.4 Model Evaluation**

The performance of the employed machine learning model is a crucial aspect of the study. A number of score matrices are available that can be used to evaluate the performance of machine learning models. Some of the performance evaluation metrics that have been used in this paper are given below. If $y_i$ and $\hat{y}_i$ are the i[th] observed and predicted values, respectively, and $n$ is the number of observations.

- **Root Mean Squared Error (RMSE):** It gives the root of the average squared difference between the actual values and the predicted values for a regression problem. Usually, an RMSE score of less than 1 is considered the best. The formula is given by

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n}}$$

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 19 No. 01 (2025)

- **Mean Absolute Error (MAE):** It gives the absolute difference between observed and predicted values. The higher negative mean values indicate the better performance of the model. The MAE is obtained as

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- **R-Squared ($R^2$):** This metric measures how well the regression model fits the data by comparing it to a baseline model that always predicts the mean value. It shows how much variation in the data is explained by the model. The formula is given by

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

## 3. Results and Discussions

Machine learning algorithms are employed to predict flight ticket prices. The Python software is used to obtain the results of the analysis of the flight ticket price dataset. The data reveals that the majority of travellers originate from Delhi and Cochin is the most popular destination. The distribution is evident in Figure 1, which indicates a significant proportion of flights departing from Delhi, and the distribution in Figure 2 shows that a significant proportion of flights arrive at Cochin.

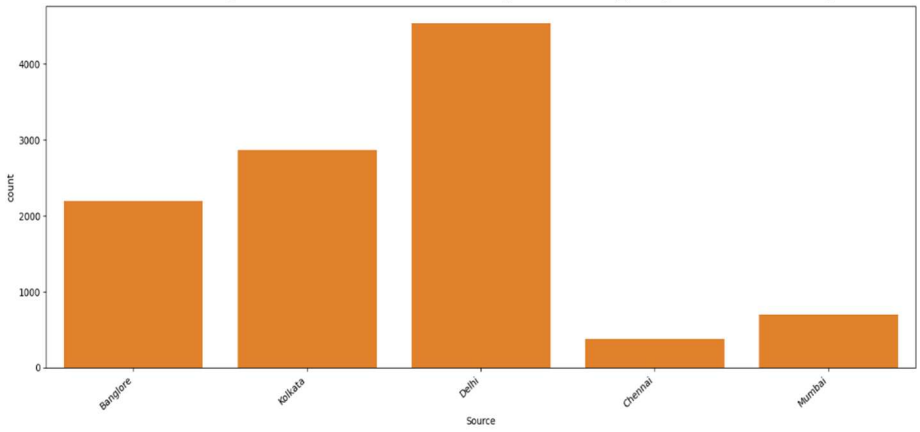


**Figure 1:** Distribution of Source for Flights Departure.

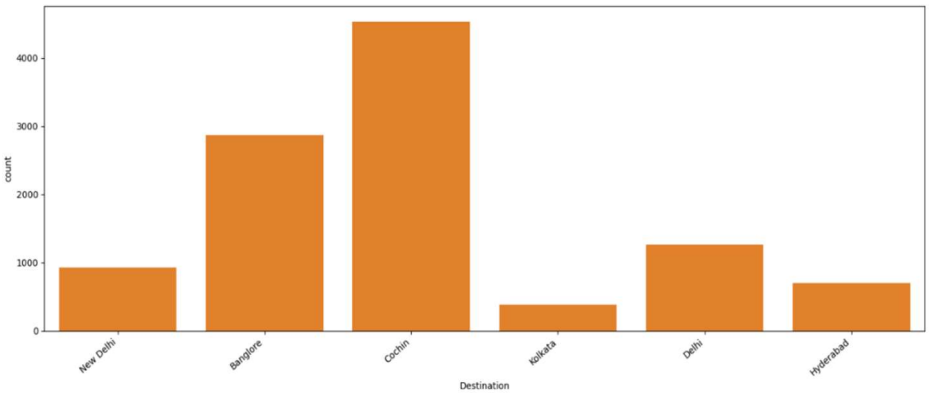


**Figure 2:** Distribution of Destination for Flights Arrival.

***Random Forest Regressor***

The training dataset contains 8545 rows of data and for testing 2137. *A random forest regressor model* with 500 estimators and a minimum sample split of 3 was implemented using the *scikit-learn library in Python.*

The random forest regressor's decision tree reveals a complex pattern of conditional statements that

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 19 No. 01 (2025)

govern the prediction of flight ticket prices. Therefore, we need to see the decision tree in part to understand the random forest regressor output. The zoomed image of the right-hand side for the random forest regressor is presented in Figure 3 and it shows that if the duration is less than or equal to 0.314, the model checks for route 1 and arrival time in minutes. If route 1 is less than or equal to 0.325, the model further checks for departure time, and if arrival time in minutes is less than or equal to 1.357, the model checks for its leave.

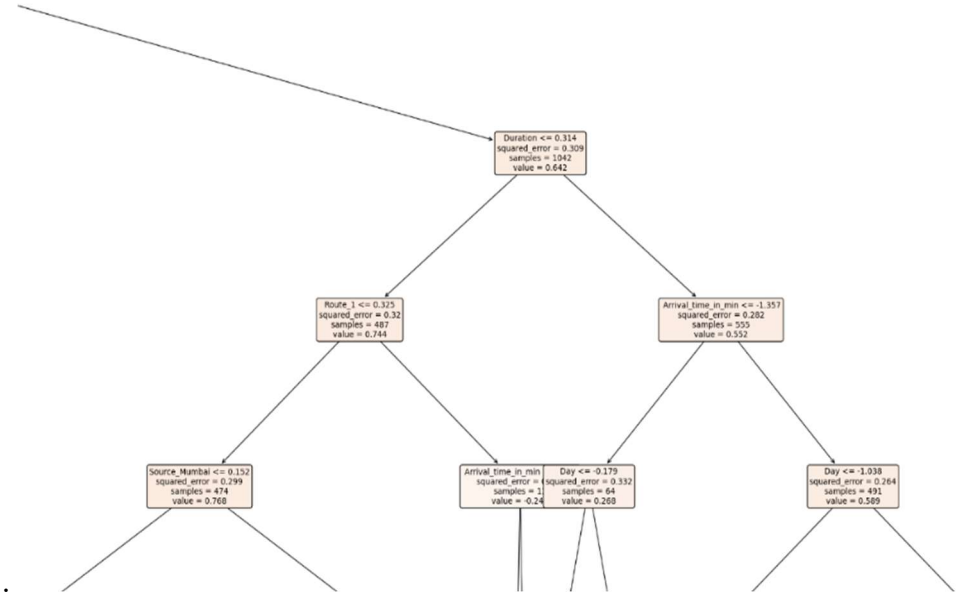


**Figure 3:** Left-hand side of random forest regressor decision tree.

Similarly, the zoomed image of the left-hand side for random forest regressor decision tree is presented in Figure 4 and it shows that if the duration is less than or equal to 1.104, the model checks for departure time in hours and arrival time in minutes. If departure time in hours is less than or equal to 0.431, the model further checks for flight Indigo, and if arrival time in minutes is less than or equal to 1.357, the model checks for its leave. Hence, carry the process to reach the conclusion. This nested structure of conditional statements enables the model to capture the intricate relationships between various features influencing flight ticket prices.

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 19 No. 01 (2025)

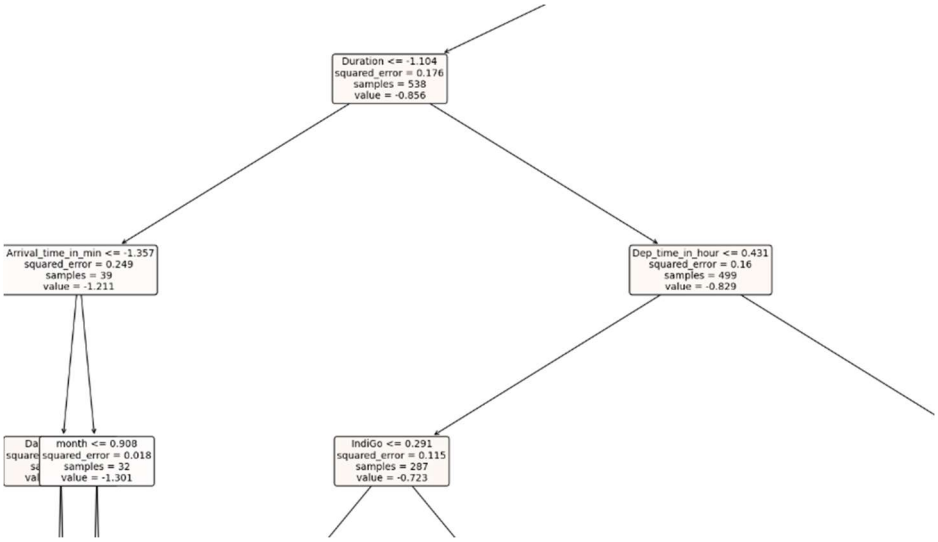


**Figure 4:** Right-hand side of random forest regressor decision tree.

***Gradient boosting regressor***

The Gradient Boosting Regressor (GBR) algorithm is used to predict the flight ticket price. To evaluate the performance of the model, we set a random state *(SEED = 23)* for reproducibility, ensuring that the results can be replicated across different runs and $Learning\ rate\ = 0.1$

The model's performance in predicting flight ticket prices accurately will be assessed by comparing the predicted values with the actual ticket prices. A plot between the number of boosting iterations and the deviance for both the training and the test datasets is given in Figure 5. It helps in visualising the difference between the actual and predicted price of the flight ticket. Figure 5 shows a noticeable difference between the values, the slight y-axis scale variation of 0.05 indicates a close approximation between the results, suggesting a reliable prediction outcome. It also shows that as the number of boosting iterations increases there is a higher difference between the actual and predicted price of the flight ticket.

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 19 No. 01 (2025)

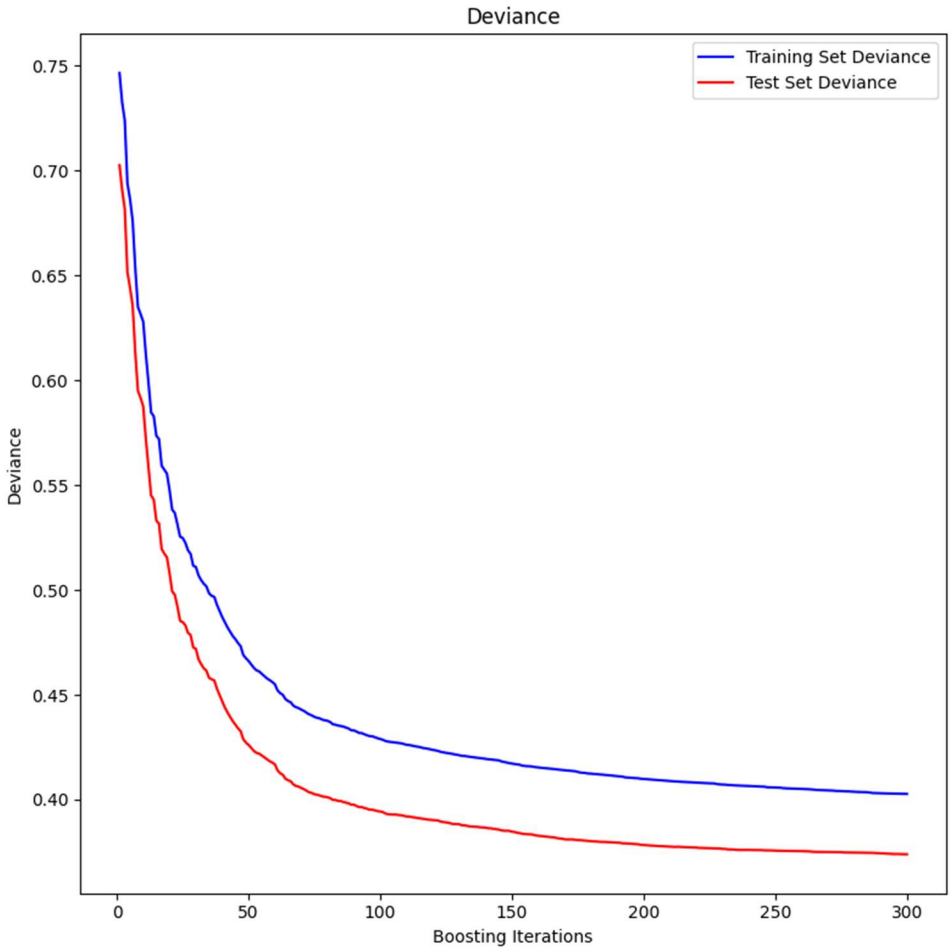


**Figure 5:** The Deviance of training and testing datasets for boosting iteration.

This gives RMSE of 0.41 which implies that the model's predictions are reasonably accurate, with deviations from the actual prices averaging around 0.41 units. It can be stated that the model's performance in predicting flight ticket prices accurately will be assessed by comparing the predicted values with the actual ticket prices.

***XGBoot Regressor***

We have performed XGBoost regressor for the prediction of flight prices on the considered dataset. without missing values and with missing values. The structure of the decision trees built by the XGBoost model is allowing to gain insights into the model's decision-making process and the importance of different features in predicting flight ticket prices. The resulting tree visualizations given in Figure 6 and provide a clear and intuitive understanding of how the XGBoost model makes predictions, enabling us to interpret the model's behaviour and identify the key factors influencing flight ticket prices.

The decision tree formed by the XGBoost model shows that if f0 <-0.61 and the value is not missing, the model checks if $f0 < -0.92$ . If $f0 < -0.61$ and the value is missing, the model checks if $f_{11} < 1.33$. The analysis continues down the tree based on these split conditions to make the final prediction.

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 19 No. 01 (2025)

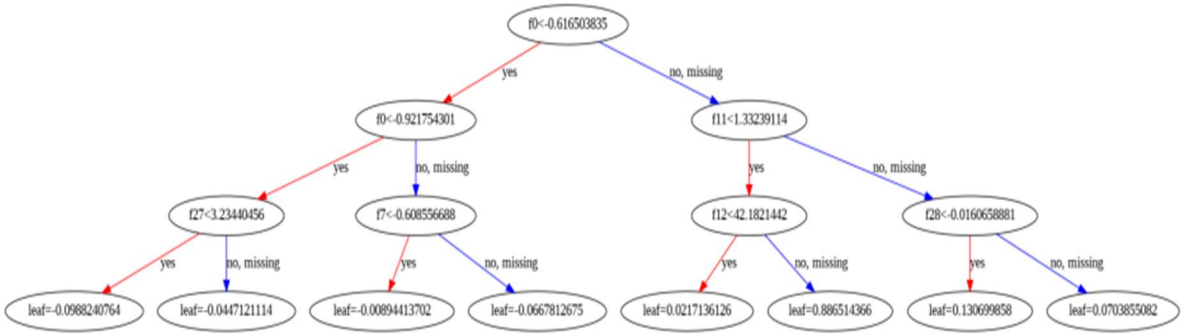


**Figure 6:** Decision tree for XGBoost regressor.

The comparative analysis underscores the strengths and limitations of each algorithm. Random Forest is noted for its robustness and ease of use, Gradient Boosting for its accuracy and ability to model complex interactions, and XGBoost for its efficiency and scalability. The research concludes that machine learning models, particularly XGBoost, can significantly enhance the accuracy of flight ticket price predictions, aiding airlines in optimizing pricing strategies and passengers in making informed purchasing decisions.

***Model evaluation***

Various performance measures of goodness of fit viz., R-squared ($R^2$), root mean square error (RMSE), means absolute error (MAE) for all the employed machine learning models are also calculated and are listed in Table 2.

**Table 2:** The performance matrices for the models.

| **Model** | $R^2$ | **RMSE** | **MAE** | **Time taken (in minutes)** |
|---|---|---|---|---|
| Random Forest | 0.824 | 0.423 | 0.244 | 3.24 |
| Gradient Boosting | 0.802 | 0.410 | 0.406 | 1.19 |
| XGBoost (without missing values) | 0.859 | 0.452 | 0.406 | 1.19 |
| XGBoost (with missing values) | 0.864 | 0.450 | 0.585 | 1.19 |

The random forest regressor model achieved a high accuracy of 86.78% in predicting flight ticket prices, with a low RMSE value indicating small errors in the predictions. The $R^2$ value of 0.824 suggests that the model explains a significant proportion of the variance in the data, meaning it captures a substantial amount of the variability in ticket prices.

The results of the analysis highlight the effectiveness of the GBR model in accurately predicting flight ticket prices, with a low RMSE of 0.410, indicating that the model's predictions deviate from the actual prices by an average of 0.410 units. This model takes very minimum time compared to the random forest model.

The XGBoost model without missing, the $R^2$ value 0.859 suggests that the model explains a significant portion (85.9%) of the variance in the target variable, which is the flight ticket price. The values of RMSE and MAE are 0.452 and 0.406 for the XGBoost regressor without missing values respectively, these low values of both indicators show that the model's predictions are good. Although, the XGBoost model with missing values has $R^2$ value 0.863, which suggests that the model explains a significant portion (86.3%) of the variance in the target variable. The MAE value is 0.585. A high $R^2$ value indicates a good overall fit, while a large MAE suggests that there are

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 19 No. 01 (2025)

significant errors in the predictions, particularly for data points where the model deviates substantially from the actual values.

## 4. CONCLUSION

This study highlights the potential of machine learning in transforming airline revenue management and passenger decision-making processes. The comparative analysis underscores the strengths and limitations of each algorithm. Random Forest is noted for its robustness and ease of use, Gradient Boosting for its accuracy and ability to model complex interactions, and XGBoost for its efficiency and scalability. The research concludes that machine learning models, particularly XGBoost, can significantly enhance the accuracy of flight ticket price predictions, aiding airlines in optimizing pricing strategies and passengers in making informed purchasing decisions. The key features for flight ticket price prediction based on random forest regressor are demand, date of journey, route, source, destination and route while GBR and XGBoost also have similar key features along with the airline. A detailed comparative analysis of the advanced machine learning algorithms, this study contributes valuable insights to the field of airline ticket price prediction, paving the way for more sophisticated and accurate prediction models.

## References

1. Abdella, J.A., Zaki, N.M. Shuaib, K. and Khan, F. (2021). Airline ticket price and demand prediction: A survey. *Journal of King Saud University-Computer and Information Sciences* 33, 375-391.
2. Babu, M.C., Rahul, E., Mohan, D. and Kumar, N.R. (2020). Flight ticket price prediction using machine learning. *ZKG International*, 8, 682-691.
3. Groves, W. and Gini, M. (2011). A regression model for predicting optimal purchase timing for airline tickets. *Technical Report 11-025, University of Minnesota, Minneapolis, USA*.
4. Groves, W. and Gini, M. (2013). An agent for optimizing airline ticket purchasing. *12th International Conference on Autonomous Agents and Multiagent Systems*, *Saint Paul, Minnesota, Minneapolis, USA,* 2, 1341-1342.
5. Janssen, T., Dijkstra, T.M.H., Abbas, S. and Riel, A.C.R.V. (2014). A linear quantile mixed regression model for prediction of airline ticket prices. *Radboud University, 1-32.*
6. Konstantinos, T., Kalampokas, T., Papakostas, G.A. and Diamantaras, K. (2017). Airfare prices prediction using machine learning techniques. *European Signal Processing Conference, Kos, Greece*. DOI: 10.23919/EUSIPCO.2017.8081365.
7. Kotsiantis, S. B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review,* 39 (4), 261-283.
8. Li, Y. and Li, Z. (2018). Design and implementation of ticket price forecasting system. AIP Conference Proceedings, 1967(1), 040009. DOI: 10.1063/1.5039083
9. Panigrahi, A., Sharma, R., Chakravarty, S., Paikaray, B.K. and Bhoyar, H. (2022). Flight price prediction using machine learning. *CEUR Workshop Proceedings*, 3283, 172-178.
10. Papadakis, M. (2014). Predicting Airfare Prices.
11. Ren, R., Yang, Y. and Yuan, S. (2105). Prediction of airline ticket price. *Technical Report, Stanford University.*

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 19 No. 01 (2025)

12. Sowjanya, Ikram, S., Khan, R., Jawad, S.H., and Arifeen, S.U. (2022). Prediction of Airfare Prices Using Machine Learning. *International Journal of Mechanical Engineering*, 7(6), 819-825.

13. Subramanian, R.R., Murali,M.S., Deepak, B., Deepak, P., Reddy, H.N. and Sudharsan, R.R. (2022). Airline fare prediction using machine learning algorithms. *4th International Conference on Smart Systems and Inventive Technology,* 877–884.

14. Vaishnavi, K.D.V.N., Bindu, L.H., Satwika, M., Lakshmi, K.U. Harini, M. and Ashok, N. (2023). Flight fare prediction using machine learning. *EPRA International Journal of Research and Development*, 8, 245-250.