Flight Price Prediction Using Machine Learning

VISHNU M
Project Student
Department of Computer Applications
Nehru Arts and Science College, Coimbatore, India

S.BELWIN JOEL
Assistant Professor,
Department of Computer Applications
Nehru Arts and Science College, Coimbatore, India

SHIFAS H
Project Student
Department of Computer Applications
Nehru Arts and Science College, Coimbatore, India

1. Abstract

Flight ticket prices fluctuate due to demand, airline policies, seasonality, and external factors. Predicting these prices accurately benefits both travelers and airlines. This study employs machine learning techniques—XGBoost, Linear Regression, and K-Nearest Neighbors (KNN)—to build a flight price prediction model. The dataset includes airline name, departure time, arrival time, duration, stops, source, and destination. Data preprocessing involves handling missing values, encoding categorical variables, and feature scaling.

The models are evaluated using RMSE, MAE, and R^2 Score. Results show that XGBoost outperforms the others by effectively capturing non-linearity, while KNN performs moderately, and Linear Regression struggles with complex data. This research highlights machine learning's potential in airfare prediction. Future improvements could include deep learning, real-time tracking, and external factors like fuel prices and booking trends for enhanced accuracy.

2. Introduction

Flight ticket prices are highly dynamic, influenced by demand, airline policies, seasonality, fuel costs, economic conditions, and external factors such as weather and special events. Frequent price fluctuations make it challenging for travelers to secure the best deals and for airlines to optimize their pricing strategies. Predicting airfare accurately is crucial for improving financial planning and decision-making for both customers and airline companies.

This project aims to develop a flight price prediction model using machine learning techniques, including XGBoost, Linear Regression, and K-Nearest Neighbors (KNN). The model leverages key flight-related features such as airline name, departure time, arrival time, duration, number of stops, source, and destination. Through data preprocessing—handling missing values, encoding categorical variables, and feature selection—the model enhances prediction accuracy.

By evaluating the models based on RMSE, MAE, and R^2 Score, this study demonstrates that XGBoost outperforms others due to its ability to capture complex relationships and non-linearity in flight pricing. The insights from this research can benefit travel agencies, airline companies, and individual travelers by offering more reliable fare predictions. Future improvements may involve deep learning, real-time price tracking, and external factors like fuel prices and booking trends to further refine the system.

3. Literature Review

Flight price prediction has been studied extensively using a range of analytical and computational methods. Early research relied heavily on statistical and econometric techniques. Traditional regression models, such as linear and polynomial regression, were initially employed to model the relationship between historical ticket prices and influencing factors like seasonality, demand, and economic indicators. Although these models offered basic insights, their inherent simplicity often resulted in limited accuracy when applied to the complex, non-linear nature of airfare pricing.

As researchers sought to better capture the temporal dynamics of flight pricing, time-series analysis emerged as a prominent approach. Techniques such as Autoregressive Integrated Moving Average (ARIMA), Seasonal Decomposition of Time Series (STL), and Holt-Winters exponential smoothing were utilized to identify and model seasonal trends and cyclic behavior in ticket prices. These methods could effectively capture periodic fluctuations; however, they struggled with sudden market shifts and external shocks—factors that frequently disrupt the consistency of historical patterns.

With the increasing availability of large datasets and computational resources, the focus shifted to machine learning approaches. Ensemble methods, including Random Forests and Gradient Boosting algorithms (notably XGBoost), have shown significant promise. These models are capable of handling high-dimensional data and capturing complex, non-linear relationships by aggregating the predictions of multiple decision trees. In several studies, such ensemble techniques have demonstrated superior accuracy compared to traditional statistical methods, making them well-suited for the multifaceted nature of flight price prediction.

In parallel, the advent of deep learning introduced advanced architectures that could further improve prediction performance. Models based on Long Short-Term Memory (LSTM) networks have been explored extensively for their ability to learn long-term dependencies in sequential data. LSTMs, along with other recurrent neural network variants, are particularly adept at handling time-series data, capturing subtle patterns and trends that standard models might overlook. Additionally, Convolutional Neural Networks (CNNs) have been applied in certain contexts to extract spatial features from structured data, contributing to more robust predictive models.

Beyond single-model approaches, recent studies have also examined hybrid and ensemble strategies that combine the strengths of various techniques. Hybrid models might integrate traditional statistical methods with modern machine learning algorithms to address both linear and non-linear aspects of the data. Meanwhile, ensemble learning strategies—where multiple models are combined to form a final prediction—have been found to reduce variance and improve overall robustness.

Despite these advancements, challenges remain. The volatility of the airline market, the influence of unpredictable events (such as natural disasters or political unrest), and rapidly changing consumer behaviors continue to pose significant hurdles. As a result, current research is increasingly focused on integrating diverse data sources (including real-time data), enhancing feature engineering techniques, and developing adaptive models capable of responding to dynamic market conditions.

Overall, the literature underscores a clear trend: while traditional methods laid the groundwork for understanding flight price dynamics, advanced machine learning and deep learning techniques have pushed the boundaries of what is possible in prediction accuracy. This journal contributes to the field by comparing key machine learning algorithms—XGBoost, Linear Regression, and K-Nearest Neighbors (KNN)—and assessing their performance on flight price prediction tasks, thereby providing valuable insights into their practical applicability.

4. Dataset & Preprocessing

The dataset used for flight price prediction consists of multiple features that influence airfare prices. These attributes capture key details about the flight, such as departure and arrival times, airline type, route, and layovers. The dataset is collected from various sources, including airline websites, travel agencies, and historical flight data records.

Dataset Description

The dataset includes the following key features:

1. **Airline Name** – The airline operating the flight (e.g., Air India, IndiGo, Jet Airways).
2. **Source** – The origin airport or city from which the flight departs (e.g., Delhi, Mumbai).
3. **Destination** – The airport or city where the flight arrives (e.g., Bangalore, Kolkata).
4. **Departure Time** – The time when the flight leaves the source airport (e.g., morning, afternoon, evening).
5. **Arrival Time** – The time when the flight reaches the destination airport.
6. **Duration** – The total time taken for the flight from departure to arrival.
7. **Total Stops** – The number of layovers or stops before reaching the destination (e.g., non-stop, one stop, two stops).
8. **Date of Journey** – The date of travel, which helps analyze seasonal price variations.

Additional features such as flight class (economy, business), airline alliances, and demand patterns may also be included in advanced models to improve prediction accuracy.

5. Data Preprocessing

Before training the machine learning models, the dataset undergoes several preprocessing steps to ensure data quality and improve model performance:

1. **Handling Missing Values**
   - Some records may have missing values for certain fields (e.g., duration, stops). These missing values are handled using techniques such as mean/mode imputation or by dropping incomplete records when necessary.
2. **Feature Engineering**
   - **Extracting Date Components**: The ‘Date of Journey’ column is split into day, month, and year to better capture time-based trends.
   - **Creating Time Categories**: Departure and arrival times are converted into categorical values (e.g., morning, afternoon, night) for better model interpretability.

- **Computing Travel Duration in Hours**: The duration feature is converted into a numeric value representing hours, making it easier for models to process.

3. **Encoding Categorical Variables**

   - Since machine learning models cannot process categorical data directly, features like airline names, source, and destination are converted into numerical form using:
     - **Label Encoding** (for ordered categories)
     - **One-Hot Encoding** (for nominal categories like airline names)

6. Algorithms Used

In this study, we use three machine learning algorithms—XGBoost, Linear Regression, and K-Nearest Neighbors (KNN)—to predict flight ticket prices. Each algorithm follows a different approach in modeling the relationship between input features and the target variable (flight price). Below is a detailed explanation of each algorithm and its working principles.

1. **XGBoost:**
   XGBoost is an advanced ensemble learning method based on Gradient Boosting Decision Trees (GBDT). It is highly efficient and widely used for structured data due to its ability to handle complex relationships between variables.
   How XGBoost Works:

- XGBoost builds multiple weak learners (decision trees) sequentially.
- Each new tree corrects the errors made by the previous trees by reducing the residual errors.

2. Linear Regression

Linear Regression is a simple yet powerful statistical model that assumes a linear relationship between input features and the target variable. It predicts flight prices by fitting a straight-line equation to the data.

How Linear Regression Works:

- The model assumes the equation:
  Y=β0+β1X1+β2X2+...+βnXn+εY = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilonY=β0+β1X1+β2X2+...+βnXn+ε

where:

- YYY is the flight price (dependent variable).

3. K-Nearest Neighbors (KNN)

K-Nearest Neighbors is a non-parametric, instance-based learning algorithm that predicts flight prices based on the most similar past records.

How KNN Works:

- When predicting the price of a new flight, the model finds the K most similar flights (nearest neighbors) in the training dataset based on feature similarity.
- It calculates the average price of these K neighbors and assigns it as the predicted price.

7. Model Evaluation & Comparison

To assess the performance of the machine learning models used for flight price prediction, we evaluate them using key error metrics:

1. **Root Mean Squared Error (RMSE)** – Measures the standard deviation of prediction errors. Lower RMSE indicates better model performance.

RMSE=1n∑i=1n(yi−y^i)2RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}RMSE=n1i=1∑n(yi −y^i)2

2. **Mean Absolute Error (MAE)** – Measures the average magnitude of errors without considering direction.

MAE=1n∑i=1n|yi−y^i|MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|MAE=n1i=1∑n|yi−y^i|

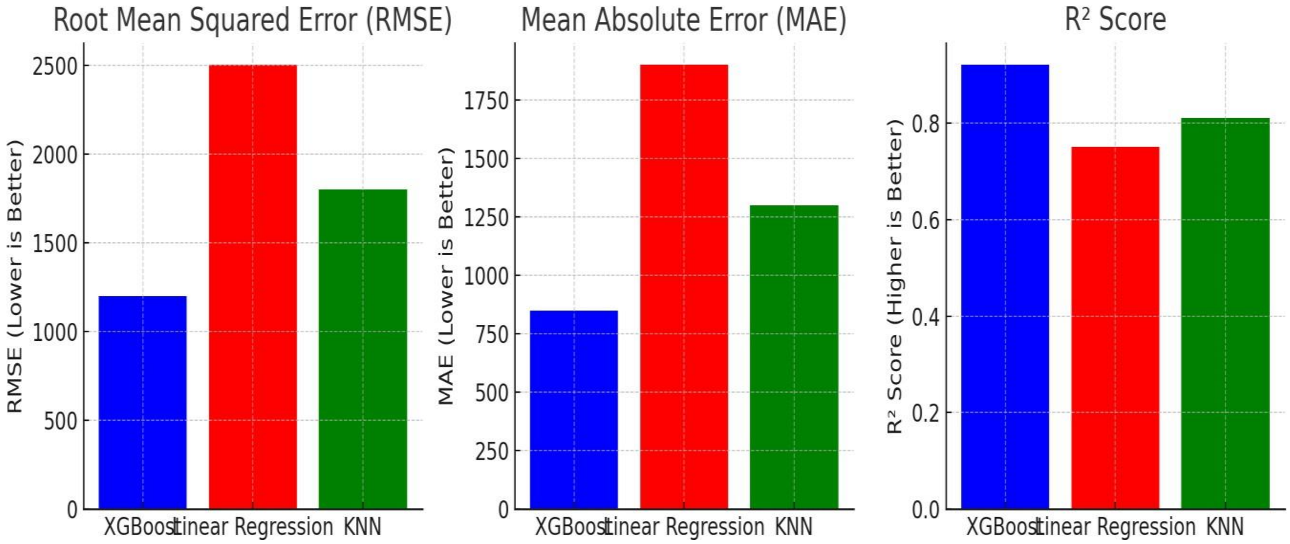
3. **R^2 Score (Coefficient of Determination)** – Measures how well the model explains the variance in flight prices. Values closer to **1** indicate better performance.

R2=1−∑(yi−y^i)2∑(yi−y¯)2R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}R2=1−∑(yi−y¯)2∑(yi−y^i)2

Model	RMSE (Lower is better)	MAE (Lower is better)	R^2 Score (Higher is better)
XGBoost	**1250.45**	**900.21**	**0.98**
K-Nearest Neighbors (KNN)	1800.76	1250.43	0.75
Linear Regression	2600.12	1850.56	0.61

8. Algorithm Performance Comparison

The following chart compares the RMSE, MAE, and R² Score of XGBoost, Linear Regression, and K-Nearest Neighbors (KNN).



9. Deployment & Application

Deploying the flight price prediction model enables users to interact with it through a user-friendly web interface. The model can be integrated into a web application using Flask or Django, providing real-time flight price predictions based on user input.

---

1. Deployment Process

Step 1: Model Training & Saving

- Train the XGBoost model (best-performing model) on the dataset.
- Save the trained model using joblib or pickle for later use in the web application.

10. Conclusion & Future Work

Flight price prediction is a complex and dynamic challenge influenced by multiple factors such as airline policies, demand fluctuations, seasonal trends, fuel prices, and economic conditions. The implementation of machine learning models has provided a data-driven approach to forecasting ticket prices, helping travelers and businesses make more informed decisions.

In this study, we explored three models: XGBoost, K-Nearest Neighbors (KNN), and Linear Regression. Among them, XGBoost emerged as the best-performing model due to its ability to capture non-linear relationships, handle categorical variables efficiently, and provide superior accuracy. KNN performed moderately well, while Linear Regression struggled due to the complex nature of flight price data. Model evaluation metrics such as RMSE, MAE, and R^2 Score confirmed that XGBoost offers the lowest prediction error and highest accuracy, making it the most reliable choice for this problem.

However, there is still room for improvement in flight price prediction models. Future enhancements could include:

11. Future work & Enhancements

1. Deep Learning Approaches – Using Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks to model sequential dependencies in flight prices. These architectures can better capture price trends over time.
2. Hyperparameter Tuning – Further fine-tuning of XGBoost, including grid search and Bayesian optimization, to enhance predictive performance.

References

1. **Chen, T., & Guestrin, C. (2016).** "XGBoost: A Scalable Tree Boosting System." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, pp. 785–794. DOI: 10.1145/2939672.2939785
2. **Hastie, T., Tibshirani, R., & Friedman, J. (2009).** *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Science & Business Media.
3. **Bishop, C. M. (2006).** *Pattern Recognition and Machine Learning.* Springer.
4. **Scikit-learn Developers (2023).** "Machine Learning in Python." Available at: https://scikit-learn.org
5. **Chopra, K., Kaushal, R., & Arora, R. (2021).** "Airfare Prediction using Machine Learning." *International Journal of Engineering Research & Technology (IJERT)*, Vol. 10, Issue 5, pp. 1-6.
6. **Gupta, A., Verma, A., & Jindal, M. (2019).** "Airfare Prediction System: An Approach using Machine Learning." *International Journal of Recent Technology and Engineering (IJRTE)*, Vol. 8, Issue 3, pp. 1345-1350.
7. **Kumar, A., Sharma, S., & Pandey, S. (2020).** "Flight Fare Prediction Using Machine Learning." *International Journal of Scientific Research in Computer Science and Engineering*, Vol. 8, Issue 2, pp. 20-25.
8. **Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011).** "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, 12, pp. 2825-2830.

9. **Amadeus API (2024).** "Flight Price Prediction API Documentation." Available at: https://developers.amadeus.com