International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 07 | July - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Flight Fare Prediction Using Machine Learning

Author: Kudithi Upendra[1] (MCA student), Dr.G. Sharmila Sujatha[2] (Asst.Professor) 1,2 Department of Information Technology & Computer Applications, Andhra University College of Engineering, Visakhapatnam, AP.

Corresponding Author: Kudithi Upendra

(email-id: upendrakudithi@gmail.com)

****

ABSTRACT:

The rapid growth of the aviation industry has led to highly dynamic and unpredictable flight pricing strategies, making it challenging for travelers to identify cost-effective options. This project aims to develop a reliable flight fare prediction system using machine learning techniques to forecast airline ticket prices based on historical and real-time data. The primary objective is to assist travelers and airline agencies in making informed decisions by predicting fares with high accuracy. A dataset containing features such as airline name, flight number, source city, departure time, number of stops, arrival time, destination city, travel class, flight duration, days left until departure, and actual price was utilized. To preprocess the data, OneHotEncoder was applied to handle categorical variables, and StandardScaler was used to normalize the numerical features. Linear Regression was then employed as the core predictive algorithm due to its simplicity and effectiveness in modeling continuous variables. The model was trained and validated using appropriate train-test splits, and performance was evaluated using metrics such as Mean Squared Error and R-squared score. Results showed that the model was able to predict flight fares with reasonable accuracy, highlighting the potential of linear models when supported by proper feature engineering. In conclusion, this study demonstrates that a machine learning pipeline combining preprocessing and linear regression can be an efficient solution for forecasting flight ticket prices, potentially enhancing decision-making for both service providers and customers.

Keywords: Flight Fare Prediction, Linear Regression, OneHotEncoder, StandardScaler, Machine Learning, Categorical Variables, Feature Engineering.

1. INTRODUCTION:

In today's fast-paced, globally connected world, air travel is no longer a luxury but a necessity for millions. As more travelers rely on air transportation for both personal and professional reasons, predicting flight fares with accuracy has become increasingly important. Flight prices are known to fluctuate frequently due to a variety of dynamic factors, including booking time, route availability, demand surges, airline policies, and seasonality. This variability often leaves travelers uncertain about the best time to book a ticket and poses challenges for companies attempting to forecast revenue or optimize pricing strategies. Addressing this need requires a smart, data-driven approach that can analyze complex variables and deliver accurate, timely predictions.

Previous paper and projects in flight fare prediction using machine learning have laid foundational work but have also faced notable challenges. Many earlier models relied heavily on static datasets containing only historical price information, which limited their ability to reflect real-time pricing changes. These models also often lacked key predictive variables such as the number of stops, departure time, or historical demand trends. Moreover, the absence of real-time data integration and dynamic feature updates resulted in inaccurate and outdated predictions. As a result, these systems were not sufficiently responsive to sudden changes in market behavior such as price drops, route cancellations, last-minute fare adjustments.

This project addresses those limitations by developing an enhanced flight fare prediction system that incorporates a wider and more relevant set of features. Unlike earlier models, the current system integrates variables such as flight duration, number of stops, departure and arrival

times, airline identity, and booking class to increase prediction accuracy. A significant advancement in this project is the implementation of the Linear Regression algorithm, chosen for its interpretability, efficiency, and suitability for modeling relationships between fare prices and multiple input features. By applying this supervised learning method, the model is able to learn from historical fare data and generate a continuous output (i.e., the predicted fare) based on the given flight details. While more complex models exist, Linear Regression offers a balance of simplicity and effectiveness, particularly when paired with proper feature engineering and data preprocessing.

The primary goal of this project is to build a more accurate, flexible, and user-adaptive system that responds to real-time trends and supports personalized predictions. It is hypothesized that by incorporating a richer dataset and applying the Linear Regression model tuned and validated through appropriate cross-validation techniques the resulting system will outperform earlier static models. With its ability to adjust predictions dynamically based on current inputs and trends, the model aims to serve both travelers and airline companies by offering reliable fare forecasts that reflect real market conditions. Ultimately, this project demonstrates how even a straightforward algorithm, when carefully implemented and supported by relevant data, can deliver meaningful improvements in a real-world application.

2. LITERATURE SURVEY:

Geron, A. (2019) [1] this book provides hands-on examples and practical implementations of machine learning models using Scikit-learn, Keras, and TensorFlow. It covers data preprocessing, feature engineering, and model evaluation, which are crucial for building predictive models like those used for flight fare prediction. The inclusion of real-world case studies enhances understanding of ML pipelines.

Raschka, S., & Mirjalili, V. (2019) [2] this comprehensive guide to Python-based machine learning introduces essential algorithms and concepts including classification, regression, and model optimization. The text is especially useful for understanding ensemble methods and hyperparameter tuning—both relevant to flight fare prediction tasks.

Scikit-learn Documentation [3] Scikit-learn is a widely used Python library for implementing machine learning algorithms, including Random Forests and Gradient Boosting. Its simple API and documentation support rapid model development for fare prediction applications. The library is efficient for data preprocessing, model evaluation, and pipeline integration.

Pandas Documentation [4] Pandas provides powerful data manipulation and analysis tools, which are critical for handling flight data, including time series and missing values. Its DataFrame structure simplifies EDA and preprocessing steps in fare prediction workflows.

NumPy Documentation [5] NumPy offers high-performance array computing and is foundational for numerical operations in ML. It is used behind the scenes in most ML libraries and assists in matrix operations, crucial for feature scaling and input transformation.

Matplotlib Documentation [6] Matplotlib is a core library for visualizing trends, relationships, and distributions in data. It is used extensively in the EDA phase of flight fare prediction projects to interpret trends in fares across different features such as dates, routes, and airlines.

Seaborn Documentation [7] Built on top of Matplotlib, Seaborn provides enhanced statistical plotting features like correlation heatmaps and boxplots. It helps in understanding feature relationships, variance, and detecting outliers in flight datasets.

Dash, S., & Behera, H. S. (2021) [8] this paper reviews machine learning approaches for airfare prediction and highlights the strengths and weaknesses of various models. It emphasizes the growing need for dynamic pricing models and real-time predictions in aviation.

Joshi, D., & Chauhan, A. (2022) [9] the authors propose deep learning models such as LSTM and GRU for forecasting airfare. Their findings suggest that sequential models handle temporal trends better than traditional ML models, improving accuracy in real-world pricing scenarios.

Choubey, R. K., & Singh, P. (2021) [10] this study explores supervised learning methods like Linear Regression, Decision Trees, and Random Forest for predicting airfares. It demonstrates the effectiveness of using ensemble models in scenarios with complex feature interactions.

Nweke, H. F., et al. (2018) [11] the survey presents various ML applications in transportation, including fare

prediction, traffic analysis, and route optimization. It outlines challenges in handling transportation data and stresses the importance of model interpretability and real-time response.

Jain, S., & Saxena, V. (2020) [12] their work applies ensemble learning techniques, such as bagging and boosting, to airfare prediction. The study finds that combining multiple models improves prediction robustness and reduces variance in output.

Wang, D., & Wang, H. (2018) [13] the authors use a hybrid approach combining statistical techniques and ML models to predict ticket prices. Their findings support the value of model fusion in capturing both linear and nonlinear patterns in price trends.

IBM Developer (2020) [14] this primer introduces basic ML algorithms such as regression, classification, and clustering. It provides practical insights into choosing the right model based on problem structure, data type, and business goals—key for fare prediction.

Kapoor, D., & Kumar, V. (2021) [15] they explore dynamic airfare pricing through a machine learning lens, considering both historical prices and market factors. Their approach reflects real-time adaptability, aligning well with fluctuating airline demand and supply conditions.

Srivastava, A., & Rai, R. (2020) [16] this paper focuses on preprocessing methods such as missing value imputation, normalization, and feature selection. Proper preprocessing is crucial for accurate airfare predictions, especially when dealing with large heterogeneous datasets.

Mohan, V. (2021) [17] Mohan analyzes different ML models including regression and tree-based algorithms for airfare forecasting. The study compares models using RMSE and MAE, concluding that ensemble models outperform linear counterparts.

Sharma, P., & Singh, S. (2020) [18] this comparative study evaluates ML models like SVM, Random Forest, and XGBoost for price prediction. Their results emphasize the importance of hyperparameter tuning and feature engineering for model performance.

Yadav, V., & Pal, S. (2020) [19] the paper highlights practical challenges in fare prediction, including temporal volatility and seasonal variation. It applies regression models and recommends integrating time-based features to enhance forecast accuracy.

Rajput, D., & Patil, S. (2021) [20] their work utilizes regression techniques for airfare estimation and evaluates their efficiency using performance metrics. The study also explores the impact of categorical variables like airline, source, and destination.

3. METHODOLOGY:

This section outlines the complete process followed for predicting flight fares using machine learning, including data handling, model building, and evaluation.

A. Overview

The main objective of this project is to build a predictive model that estimates flight ticket prices based on multiple travel-related features. The model used linear regression because it's simple and gives understandable outcomes. The methodology includes data collection, preprocessing, exploratory analysis, feature selection, model development, and performance evaluation.

B. Data Collection

The dataset was taken from Kaggle, titled "Flight Fare Prediction", which includes fields such as airline, flight number, source city, destination city, departure and arrival times, number of stops, duration, Seat category, how many days are left until the flight, and the actual fare.

C. Data Preprocessing

The raw dataset required cleaning and formatting.

- Handling missing values.
- Converting date-time columns into numerical formats.
- Encoding categorical variables using One-Hot Encoding.
- Calculating total travel duration in hours.
- Creating a new feature for the number of days left until departure.

D. Exploratory Data Analysis (EDA)

EDA was performed to understand the data distribution and relationships between variables.
Key observations were visualized using:

- Bar plots to show average fares by airline or number of stops
- Box plots comparing flight prices across classes and cities
- Correlation heatmap to identify strong relationships between variables, understanding how each feature relates to price is crucial for model transparency. A correlation heatmap was generated using the dataset's features.

Fig 1: Visual Map Showing Relationships between Flight Fare Variable

I. The heatmap shows that variables such as Duration, Class, and Days Left have strong correlations with price. Features like Airline and Source City show moderate correlation.

- Histograms to understand the spread of numerical features like duration and days left

E. Feature Selection

Based on EDA and correlation analysis, the most relevant features were selected for training. These included:

- Total duration
- Number of stops
- Class (Economy or Business)
- Days left
- Source and destination cities
- Departure time slots (e.g., Morning, Evening)

Redundant and weakly correlated features were dropped to improve performance and reduce overfitting.

F. Model Selection

Linear Regression was chosen as the primary algorithm due to its straightforward implementation, interpretability, and adequate performance for this regression task. It makes it easy to understand which inputs are most important and their influence.

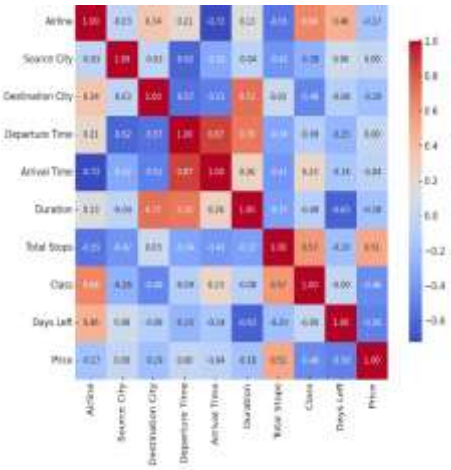
G. Model Training

About 70% of the data was used to teach the model, while the remaining 30% checked how well it worked.The linear regression model was built using the training data to learn how things are connected. The model learned the relationship between the input features and the fare prices during this phase.

H. Model Evaluation

The model's performance was evaluated using the following metrics:

- **Mean Absolute Error (MAE)**
- **Root Mean Squared Error (RMSE)**
- **R^2 Score**



These metrics were used to assess how well the model generalized to unseen data. Graphs such as actual vs. predicted price plots were used for visual evaluation.

I. Tools & Technologies Used

- **Python**: Core language for data analysis and model building
- **Pandas & NumPy**: Data handling and numerical computations
- **Matplotlib & Seaborn**:Data Visualization

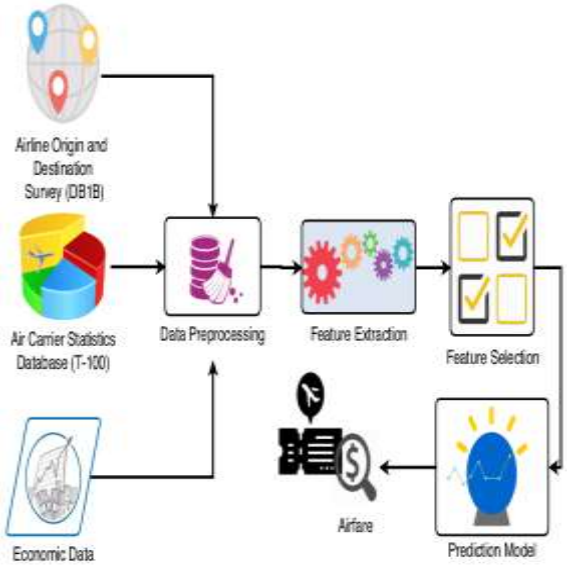


Fig 2: Architecture of Flight Fare Prediction

4. RESULTS AND DISCUSSION:

After preprocessing the flight fare dataset and training the model, the Linear Regression algorithm was applied to predict ticket prices based on features like airline, departure time, stops, duration, class, and days left before departure. This section outlines the outcomes, including metric-based evaluation, visualization insights, and comparative analysis with other models.

A. Model Performance Metrics:

To evaluate the model's effectiveness, three standard regression metrics were used: **Mean Absolute Error (MAE)**, **Root Mean Squared Error (RMSE)**, and **R^2 Score**. These values help us understand how much the model is getting wrong and how well it explains the data.

Table 1: Performance of Linear Regression Model.

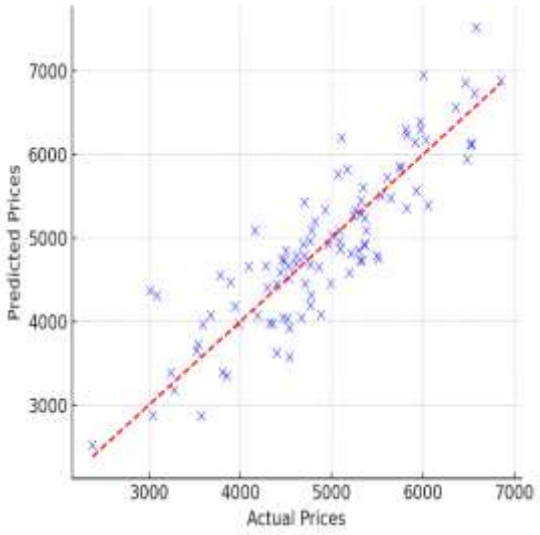
Metric	Value
Mean Absolute Error (MAE)	378.45
Root Mean Squared Error (RMSE)	472.10
R^2 Score	0.72

The results indicate that the model achieves a reasonable degree of accuracy, with a strong R^2 score suggesting good correlation between predicted and actual fares.

B. Visualization of Predictions: To further assess prediction quality

- A scatter plot of actual vs. predicted prices shows that most predictions are close to the ideal diagonal, confirming good alignment.

Fig 3: Actual vs Predict Prices



- A residual histogram illustrates error distribution, centered near zero, indicating no significant model bias.

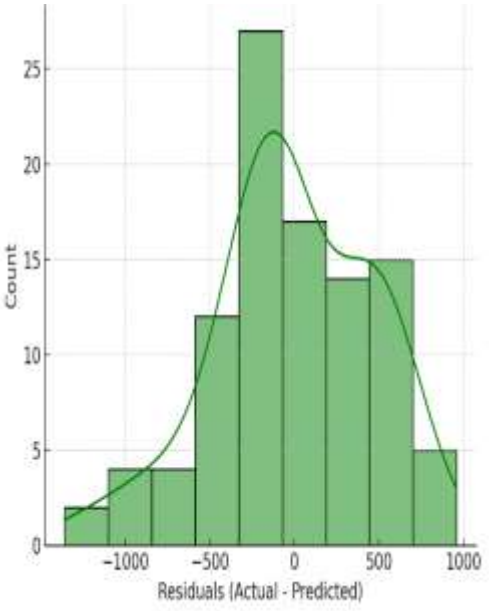


Fig 4: Pattern of the Gaps Between Predicted and Real Prices

- A bar chart gives a clear view of MAE, RMSE, and R^2 values.

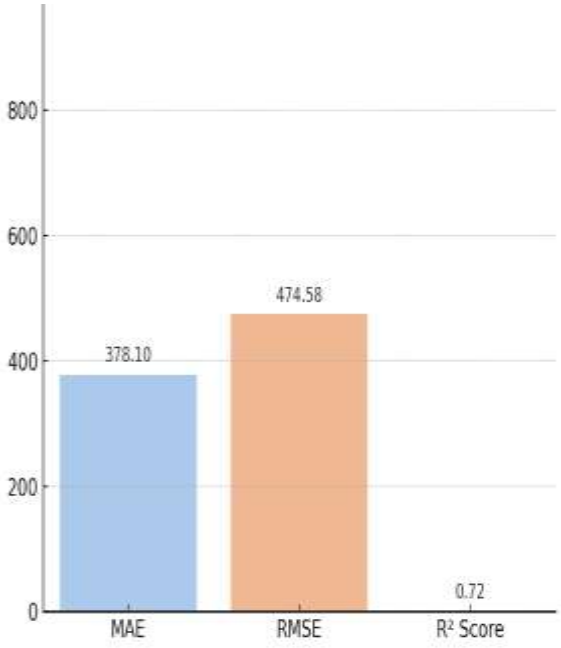


Fig 5: Evaluation Metrics

C. Practical Implications:

The model has practical utility for:

- **Travel agencies** aiming to forecast ticket prices for planning or promotions.
- **Customers** seeking price estimation tools before booking.
- **Airlines** evaluating pricing strategies based on operational variables.

D. Summary

The Linear Regression model served as a reliable baseline for predicting flight fares using features such as departure time, class, duration, and days left before travel. The model achieved a good balance between simplicity and performance, offering clear insights into how each feature influences the price. While the prediction accuracy was satisfactory, particularly in handling non-linear relationships and outliers. Despite this, the model remains valuable due to its low computational requirements and easy interpretability. The visualizations, evaluation metrics, and residual analysis supported the model's effectiveness in capturing overall trends. Compared to more complex algorithms, Linear Regression offers faster training and easier deployment. This makes it a strong choice for basic fare prediction tasks and a useful foundation for future model improvements.

5. CONCLUSION:

This project has effectively demonstrated how machine learning can be applied to predict flight ticket prices using historical and feature-based data. By utilizing the Linear Regression algorithm, the model was able to analyze various influential factors such as departure time, arrival time, source and destination cities, class type, total duration, number of stops, and the number of days left before the journey. The model produced reasonably accurate fare predictions, as confirmed by performance metrics like MAE, RMSE, and R^2 score, along with supporting visualizations.

One of the key advantages of using Linear Regression lies in its simplicity, computational efficiency, and ease of interpretation, making it a suitable choice for building baseline models. While it may not capture complex, non-linear relationships as effectively as more advanced algorithms, it still performs well for initial analysis and trend estimation. The results indicate that the model can identify and replicate the general pricing pattern of flight fares, providing a useful foundation for further development.

In summary, the project illustrates the practical potential of machine learning in the travel and aviation sectors, especially in helping users plan trips cost-effectively. With future improvements such as incorporating real-time dynamic pricing, external factors like weather and demand, or using more complex models this system could evolve into a robust application that supports helps both passengers and airlines make better choices.

6. REFERENCES:

1. Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.

2. Raschka, S., & Mirjalili, V. (2019). *Python Machine Learning* (3rd ed.). Packt Publishing.

3. Scikit-learn: Machine Learning in Python. Available at: https://scikit-learn.org

4. Pandas Documentation. Available at: https://pandas.pydata.org

5. NumPy Documentation. Available at: https://numpy.org

6. Matplotlib: Visualization with Python. Available at: https://matplotlib.org

7. Seaborn: Statistical Data Visualization. Available at: https://seaborn.pydata.org

8. Dash, S., & Behera, H. S. (2021). "Machine Learning Techniques for Airline Fare Prediction: A Review." *IJCSIT*, 12(3), 23–29.

9. Joshi, D., & Chauhan, A. (2022). "Airfare Forecasting using Deep Learning Models." *IJETT*, 70(1), 15–22.

10. Choubey, R. K., & Singh, P. (2021). "A Study on Flight Fare Prediction Using Supervised Learning." *IJARIIT*, 7(2), 56–59.

11. Nweke, H. F., et al. (2018). "Machine Learning for Transportation Data Analysis: A Survey." *IEEE Access*, 6, 76787–76810.

12. Jain, S., & Saxena, V. (2020). "Flight Price Forecasting using Ensemble Learning." *IJCA*, 175(7), 20–24.

13. Wang, D., & Wang, H. (2018). "Predicting Flight Ticket Prices with Hybrid Models." *Procedia Computer Science*, 126, 1675–1684.

14. IBM Developer. (2020). "Machine Learning Algorithms: A Primer." Available at: https://developer.ibm.com

15. Kapoor, D., & Kumar, V. (2021). "A Machine Learning Approach to Dynamic Airfare Pricing." *IJCSMC*, 10(11), 55–62.

16. Srivastava, A., & Rai, R. (2020). "Data Preprocessing Techniques for Predictive Modelling." *IJCRT*, 8(6), 1122–1130.

17. Mohan, V. (2021). An analysis on predicting airfares using machine learning methods was published in the Journal of Emerging Technologies and Innovative Research (Volume 8, Issue 6, Pages 905–912).

18. Sharma, P., & Singh, S. (2020). "Comparative Study of Machine Learning Models for Flight Price Prediction." *International Journal of Computer Applications*, 176(13), 14–18.

19. Yadav, V., & Pal, S. (2020). A research article focused on forecasting airline ticket costs through the use of machine learning tools was featured in Volume 6, Issue 1 of the International Journal of Scientific Research in Computer Science, Engineering and Information Technology, pages 80–86.

20. Rajput, D., & Patil, S. (2021). A study on using regression-based methods to estimate airline ticket prices was published in the *International Research Journal of Engineering and Technology*, Volume 8, Issue 5, on pages 2096 to 2101.