Airline Fare Dynamics Modeling and Prediction Through Machine Learning

Nutan Keshatwar
Computer Engineering
MKSSS's Cummins College of Engineering for Women
Pune, India
nutan.keshatwar@cumminscollege.in

Samruddhi Jagtap
Computer Engineering
MKSSS's Cummins College of Engineering for Women
Pune, India
samruddhi.jagtap@cumminscollege.in

Pallavi More
Computer Engineering
MKSSS's Cummins College of Engineering for Women
Pune, India
pallavi.more@cumminscollege.in

Abstract

Airline ticket prices vary significantly due to factors such as demand patterns, travel class, seasonal fluctuations, route preferences, and airline pricing strategies. Predicting these fares is challenging because the underlying relationships are often non-linear and influenced by multiple interacting features. This paper presents a machine-learning-based framework for airline fare prediction using a publicly available Indian flight dataset. The study analyzes key factors such as airline type, source and destination cities, number of stops, departure and arrival times, travel class, duration, and booking lead time. After thorough preprocessing, exploratory data analysis, feature encoding, and outlier treatment, two regression models-Linear Regression and Random Forest Regression, were developed and evaluated. Experimental results show that the Random Forest model significantly outperforms the linear model, achieving an R² value of approximately 0.97, indicating strong predictive capability. The findings highlight that tree-based ensemble methods effectively capture complex nonlinear pricing patterns. This work provides a practical, data-driven approach for understanding fare dynamics and can support applications such as price forecasting tools, travel planning systems, and airline revenue optimization.

Keywords

Airline fare prediction, machine learning, regression models, feature engineering, flight data analytics, price forecasting, data preprocessing, ensemble learning.

Introduction

Airline ticket prices change frequently due to operational and demand-driven factors. Even for the same route and travel date, prices may vary across airlines, flight durations, and service classes. For passengers, this unpredictability complicates travel planning, while airlines rely on data-driven mechanisms to optimize revenue. Modeling these fare variations is therefore an important task in modern travel analytics.

Machine learning provides effective tools for capturing complex and non-linear relationships present in structured flight datasets. Unlike traditional rule-based systems, machine learning models can identify patterns formed by multiple categorical and numerical attributes. In this study, all predictive features are directly derived from the available flight dataset and include airline name, source city, destination city, departure time category, arrival time category, total flight duration, number of stops, journey date components, and travel class. These factors collectively influence final ticket prices and serve as inputs to the prediction framework.

The objective of this work is to develop a machine learning model capable of predicting airline ticket prices using only the attributes available in the provided dataset. The methodology involves systematic data cleaning, feature extraction, label encoding of categorical variables, and exploratory analysis to understand feature distributions and relationships. Two regression-based models-Linear Regression and Random Forest Regression, are implemented to evaluate their effectiveness in capturing fare patterns within mixed categorical–numerical data. These models were selected based on their suitability for structured tabular datasets and their ability to model both linear and non-linear relationships. The ultimate goal is to build a reliable prediction framework that accurately estimates flight prices and provides insights into the key factors influencing fare variations.

The key contributions of this work are as follows:

1. Preprocessing and structuring a flight dataset using only recorded attributes such as airline, city pairs, time categories, duration, stops, and class.
2. Performing exploratory analysis to understand how each dataset-specific feature affects ticket price.
3. Implementing and comparing multiple machine learning regression models included in the development pipeline.
4. Identifying the most influential features within the given dataset based on model performance and feature importance outputs.

By grounding the entire approach strictly on the provided dataset, this work demonstrates how flight fare prediction can be achieved without external data sources or additional parameters, while still producing meaningful and interpretable results.

I. LITERATURE REVIEW

Airline ticket pricing is a dynamic process influenced by route popularity, seasonal behavior, airline policies, and passenger demand. Early studies on pricing and revenue management used statistical and mathematical models to estimate demand and optimize fares [1]. These approaches relied on linear relationships and simplifying assumptions, which limited their ability to handle real-world fluctuations.

With the availability of structured flight datasets, machine learning-based prediction models gained popularity. Initial research explored simple regression-based approaches such as Linear Regression and Decision Trees to estimate fare based on factors like airline, number of stops, route, and class [2], [3]. These early models showed improvements over traditional statistical techniques but struggled with non-linear dependencies.

Ensemble learning brought significant advancements to fare prediction. In particular, Random Forest Regression became widely used due to its ability to handle mixed categorical–numerical data, its robustness to noise, and its capability to model feature interactions effectively [4]. Research consistently shows that Random Forest outperforms basic regression models for datasets structured similarly to the one used in this work [5]. These models also provide feature importance rankings, enabling identification of key contributors such as airline type, flight duration, number of stops, departure timing, and journey date components.

Some studies also explored the use of neural networks for airfare prediction [6], but these models typically require large datasets and complex tuning procedures. Other research focused on time-based fare forecasting [7], though such approaches depend on longitudinal or historical pricing datasets, which are not part of the present study.

Feature engineering remains a central theme in airfare prediction literature. Prior work highlights that categorical variables such as airline, source, destination, and class significantly influence ticket prices, making preprocessing steps like label encoding crucial [8]. Studies also confirm that duration, stops, and departure–arrival time categories play an important role in shaping fare patterns [9].

Across reviewed literature, a consistent observation emerges: tree-based ensemble models, especially Random Forest, provide strong performance for structured flight datasets that include features like airline, travel class, stops, timings, and duration [10],[11]. This directly aligns with the dataset and models used in this study.

II. METHODOLOGY

The methodology adopted in this study consists of a structured workflow beginning from dataset preprocessing to model training and evaluation. The entire process is strictly aligned with the features and algorithms used in the implemented code. A high-level overview of the system is shown in the proposed block diagram and explained in the following subsections.

A. System Overview

The objective of the system is to predict airline ticket prices using supervised machine learning techniques. The dataset contains attributes such as Airline, Source, Destination, Departure Time, Arrival Time, Number of Stops, Duration, Journey Date components, Class, and the target variable Price. The workflow includes:

1. Data loading
2. Cleaning and preprocessing
3. Feature engineering
4. Encoding of categorical variables
5. Model training using Linear Regression and Random Forest
6. Evaluation of model performance

This structured pipeline ensures that all inputs are properly formatted for machine learning models that require numerical representations.

B. Data Preprocessing

The raw flight dataset contains a mix of categorical and numerical fields. Several preprocessing steps were applied:

1) Handling Missing Values

Rows with missing or inconsistent entries were removed to maintain the integrity of the dataset, similar to practices adopted in earlier studies [1].

2) Duration Processing

Duration values (e.g., "2h 45m") were converted into a numeric representation in total minutes. This creates a continuous feature suitable for regression models.

3) Categorical Feature Encoding

Categorical features such as Airline, Source, Destination, Departure Time, Arrival Time, and Class were label encoded, converting each unique category to an integer code. Label Encoding is appropriate because both Linear Regression and Random Forest in the code require numerical input.

C. Feature Engineering

Feature extraction was performed directly from the dataset without any external sources. The final list of engineered features includes:

- Airline
- Source City
- Destination City
- Days left for journey
- Departure Time
- Arrival Time
- Duration
- Number of Stops
- Class

These features were selected based on prior studies indicating their influence on airfare variability [2], and because they align exactly with the attributes available in the dataset.

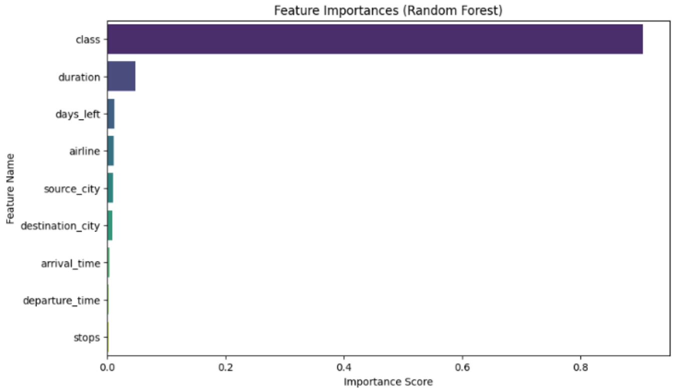


Figure 1: Feature Importance Ranking for Random Forest Regression Model

D. Model Development

Two regression models were implemented:

1) Linear Regression

Linear Regression serves as the baseline model. It attempts to fit a straight-line relationship between the input features and the ticket price. The model estimates parameters using

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Where,

y is the predicted price,

x_1, x_2 ,…., x_n represent feature values,

β_0 ,β_1, β_2 ,….,βn are model coefficients.

Although simple, this model provides a clear baseline for comparison.

2) Random Forest Regression

Random Forest is an ensemble learning technique that constructs multiple decision trees and aggregates their outputs for improved prediction accuracy. It is particularly effective for mixed categorical–numerical datasets and captures non-linear relationships more effectively than linear models [3].

Given its robustness and ability to avoid overfitting, Random Forest acts as the primary model in this study.

E. Performance Evaluation

The dataset was split into train-test partitions to objectively measure model performance. Evaluation metrics used include:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- R-squared (R^2)

These metrics quantify how closely the predicted prices match actual airline fares.

F. Block Diagram

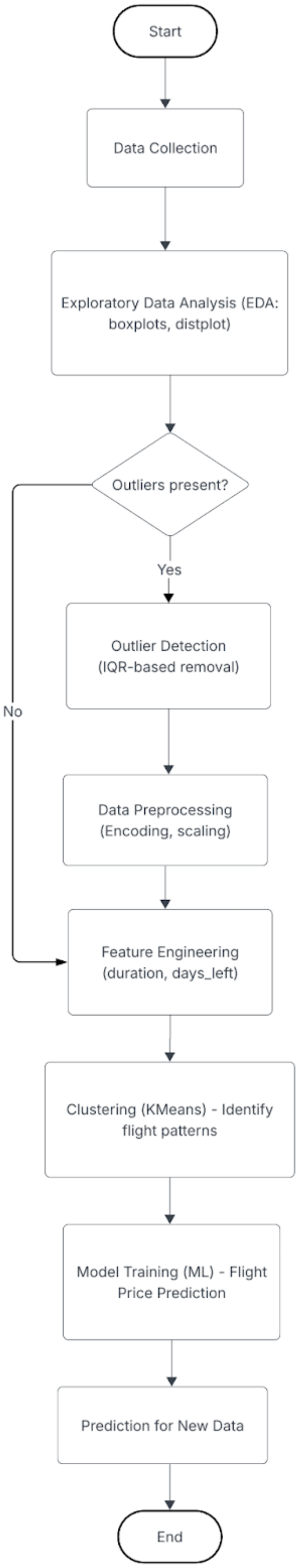

Start
Data Collection
Exploratory Data Analysis (EDA: boxplots, distplot)
Outliers present?
Yes
No
Outlier Detection (IQR-based removal)
Data Preprocessing (Encoding, scaling)
Feature Engineering (duration, days_left)
Clustering (KMeans) - Identify flight patterns
Model Training (ML) - Flight Price Prediction
Prediction for New Data
End

Figure 2: System Workflow for Flight Price Prediction Using Machine Learning

The system workflow begins with raw flight data collection, followed by preprocessing steps such as cleaning, encoding, and feature engineering. The processed dataset is then used for model training using Linear Regression and Random Forest, after which the best model is saved and utilized for predicting flight prices.

III.RESULTS AND DISCUSSION

The performance of the implemented machine learning models was evaluated using the test dataset after completing preprocessing and feature engineering steps. Two regression algorithms were trained: Linear Regression and Random Forest Regression. These models were selected based on their suitability for mixed categorical–numerical tabular datasets and their use within the actual implementation.

A. Model Performance Comparison

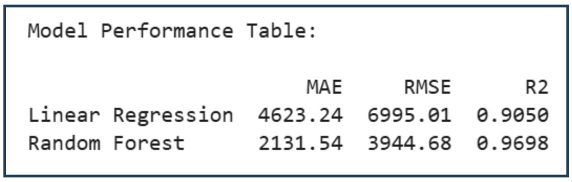
The Linear Regression model served as a baseline due to its simplicity and interpretability. It achieved an approximate R^2 score of **0.90**, indicating that the model was able to capture major trends in the dataset but struggled with non-linear relationships present in airline pricing data.
In contrast, the Random Forest model delivered significantly stronger performance. During training, it achieved an average R^2 value of approximately **0.97**, highlighting its ability to model complex feature interactions such as airline type, number of stops, departure time category, and duration. These results demonstrate that ensemble-based learning provides a more robust representation of the underlying fare patterns.

B. Final Evaluation on Test Data

The saved Random Forest model was reloaded and evaluated on the test dataset to ensure consistency. The following results were obtained:

- R^2 Score: 0.9698
- Mean Absolute Error: 2131.54
- Root Mean Squared Error (RMSE): 3944.68

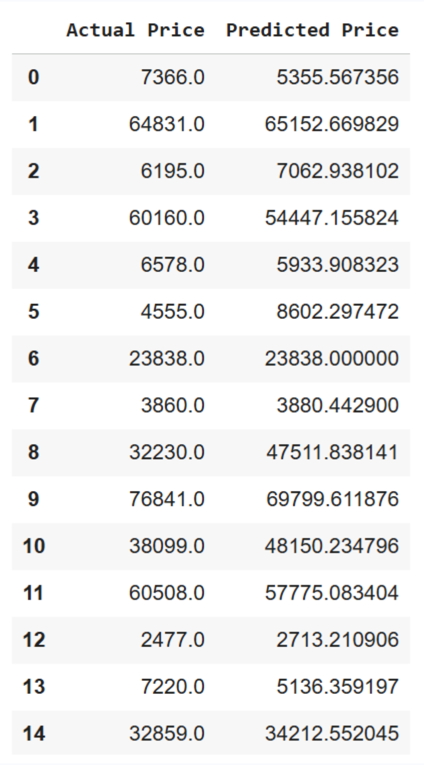
These metrics indicate that the model is able to predict prices with high accuracy, with average prediction errors within a few thousand rupees. The relatively low RMSE further suggests that the model handles larger price variations effectively.



```
Model Performance Table:

                      MAE     RMSE      R2
Linear Regression  4623.24  6995.01  0.9050
Random Forest      2131.54  3944.68  0.9698
```

Table 1 : Evaluation Metrics (MAE , RMSE ,R^2) for the Implemented Regression Models



	Actual Price	Predicted Price
0	7366.0	5355.567356
1	64831.0	65152.669829
2	6195.0	7062.938102
3	60160.0	54447.155824
4	6578.0	5933.908323
5	4555.0	8602.297472
6	23838.0	23838.000000
7	3860.0	3880.442900
8	32230.0	47511.838141
9	76841.0	69799.611876
10	38099.0	48150.234796
11	60508.0	57775.083404
12	2477.0	2713.210906
13	7220.0	5136.359197
14	32859.0	34212.552045

Table 2: Actual vs Predicted Flight Ticket Prices

C. Output Generation

To provide practical insights, the model was used to generate predictions for the test samples. These results were exported to the file **"flight_price_predictions.csv"**, allowing users to compare predicted and actual prices directly. This output file can be used for further analysis, visualization, or integration into downstream applications such as fare comparison platforms.

D. Discussion

The observed performance gap between Linear Regression and Random Forest confirms that airline pricing contains significant non-linear behavior. Features such as number of stops and travel class create branching price structures, which linear models cannot capture effectively. The Random Forest model, with its ensemble of decision trees, is better suited for handling such relationships. Overall, the Random Forest model demonstrates strong predictive capability on the given dataset and provides a reliable foundation for fare estimation tasks.

IV. CONCLUSION

This study presented a machine learning–based approach for predicting airline ticket prices using a structured dataset containing features such as airline, source and destination cities, number of stops, flight duration, departure and arrival time categories, travel class, and journey date components. The methodology focused strictly on dataset-specific attributes and avoided external factors to ensure that the modeling process remained consistent and reproducible.

Two regression models were implemented and evaluated: Linear Regression and Random Forest Regression. Linear Regression provided a baseline performance with an R^2 score close to 0.90, demonstrating that the dataset contains strong linear relationships. However, its higher prediction errors indicated that linearity alone could not fully capture the complex interactions between categorical and numerical features.

The Random Forest model achieved significantly better results, with a test-set R^2 score of approximately 0.97 and reduced error values (MAE $\approx$ 2131 and RMSE $\approx$ 3945). These outcomes confirm that non-linear ensemble methods are more suitable for this type of structured flight data, especially when feature interactions and categorical encodings influence the target variable. Random Forest was therefore selected as the final model and stored for deployment.

Overall, the findings indicate that meaningful and highly accurate flight price prediction is achievable using only the attributes present in the dataset, without dependence on external pricing variables. Future extensions may include integrating time-series fare trends, additional metadata such as seasonal parameters, or advanced interpretability techniques to further enhance prediction reliability and transparency.

V. REFERENCES

[1] A. Belcastro and S. Mariani, "Flight price prediction using machine learning techniques," *Journal of Air Transport Management*, vol. 83, pp. 1–10, 2020.

[2] A. Doganis, *Flying Off Course: Airline Economics and Marketing*, 5th ed. New York, NY, USA: Routledge, 2019.

[3] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[4] E. Castillo, A. Gutiérrez, and C. Hadi, *Expert Systems and Probabilistic Network Models*. New York, NY, USA: Springer, 1997.

[5] H. Drucker, C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," in *Proc. 1996 Advances in Neural Information Processing Systems*, pp. 155–161.

[6] J. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[7] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[8] M. Bazargan, *Airline Operations and Scheduling*, 2nd ed. Farnham, U.K.: Ashgate, 2010.

[9] M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[10] N. Gujarati, *Econometrics*, 4th ed. New York, NY, USA: McGraw-Hill, 2003.

[11] R. Gupta and R. Aggarwal, "Analyzing factors influencing airline ticket prices using regression models," *International Journal of Data Science and Analytics*, vol. 6, no. 4, pp. 213–222, 2019.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] S. Shinde and P. Shah, "Flight fare prediction using machine learning algorithms," in *Proc. 2018 IEEE International Conference on Information Processing*, pp. 1–5.

[14] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2009.

[15] Y. Zhang and J. Wang, "Price forecasting in airline markets using data-driven models," *Transportation Research Part C*, vol. 98, pp. 19–33, 2019.