FlyFarePredict: Machine Learning–Based Flight Ticket Price Prediction System

Harshada Chaudhari
Department of Computer Engineering
MKSSS's Cummins College of Engineering for Women
Pune,India
harshada.h.chaudhari@cumminscollege.in

Astha Nitnaware
Department of Computer Engineering
MKSSS's Cummins College of Engineering for Women
Pune,India
astha.nitnaware@cumminscollege.in

Pranali Nikose
Department of Computer Engineering
MKSSS's Cummins College of Engineering for Women
Pune,India
pranali.nikose@cumminscollege.in

***Abstract*— Flight ticket prices are difficult to comprehend and predict because the airline industry is dynamic and dependent upon numerous factors, such as airline, route, class, duration, date of booking, and fluctuating demand. The estimation of fares manually is complex and often goes wrong, thus making value-for-money decisions difficult for travelers. This work proposes FlyFarePredict, a machine learning-based flight ticket price prediction system, which can automatically estimate fares. The proposed framework includes a Random Forest regression model that is combined with feature engineering comprising categorical encoding and clustering analysis. Using a comprehensive flight dataset, this model predicts the prices of tickets by considering airline details, journey timings, class, duration, and historical fares. It also identifies pricing patterns across segments-economy and business. The model has secured an R^2 score of 0.97 and a mean absolute error of ₹2,137, thus demonstrating high reliability in price prediction. FlyFarePredict will enable informed booking decisions by travelers, dynamic pricing strategies by airlines, and provide actionable insights into fare trends presented in an interpretable way**

Keywords—machine learning, flight price prediction, airfare forecasting, random forest regression, feature engineering, travel analytics, route-based pricing, airline fare estimation, predictive modeling, dynamic pricing, booking date impact, flight segmentation, regression analysis, clustering analysis, ticket cost prediction, machine-assisted fare evaluation, travel decision support, price trend analysis

I. INTRODUCTION (*HEADING 1*)

Flight ticket prices are very difficult to predict by ordinary travelers due to the interplay of various factors involved, including airline, travel dates, departure and arrival cities, flight duration, class, and seasonal demand. This uncertainty can result in poor booking decisions, increased travel cost, and frustration. Manually checking numerous websites or using their own experience to estimate the fares themselves is a very time-consuming and impractical undertaking and is error-prone in daily life. Such limitations emphasize the requirement for automated tools that help users in quicker travel decisions.

While Skyscanner and Kayak are among the most popular travel platforms online, their scope in presenting flight search and historical price trends still falls short in accurately predicting the price for any individual itinerary. Skyscanner, for example, only offers approximate price trends and alerts without directly making any predictions of expected fares for a certain date, airline, or route [1]. Evaluations of other fare comparison tools have also concluded similarly-that while they help aggregate data, travelers themselves have to interpret trends and make manual booking decisions [2]. Since these latter platforms are more data presentation-oriented rather than predictive insight-oriented, they do not adequately satisfy the need for personalized, real-time price forecasting.

The following research addresses these challenges by proposing a machine-learning-based flight ticket price prediction system. It adopts historical data of flights and various features related to travel and implements Random Forest regression models for accurate ticket price prediction. This work also uses clustering techniques to divide flights into categories, such as economy, business, and premium fares, enabling users to make sense of price patterns in various market segments. The model has a very high R^2 score of 0.97, thus proving the potential of machine learning in providing speedy, reliable, and actionable insights for travelers.

The rest of this paper is organized as follows: Section II presents related work in airfare prediction and travel analytics. Section III elaborates on the methodology, including dataset structure, preprocessing, and model development. Section IV presents the model evaluation and results. Section V concludes the study and outlines the directions of future improvement**s.**

II. RELATED WORK.

Commercial flight search-and-price-tracking services also provide a range of functionalities but have key shortcomings. Skyscanner and Kayak, among others, allow consumers to view flights, price history, and receive alerts, but they essentially provide the raw data rather than accurate, predictive estimates of fares. Most third-party systems provide some degree of automation, where fare alerts or recommendations are automatically generated based on historical trends, but still require significant user interpretation and monitoring. These gaps have fueled interest in applying machine learning for automated airfare prediction [3].

In the last few years, several ML-based systems have attempted to enhance flight price analysis by incorporating historical price data, time-series forecasting, and route-specific modeling. A notable example is FlightPriceAI, which uses regression models to predict ticket prices and achieves a prediction accuracy of nearly 85% on selected routes [4]. Broader surveys in travel analytics also show that AI and machine learning are being used with increasing frequency for demand forecasting, dynamic pricing, and personalized travel recommendations. However, these studies emphasize prediction, trend estimation, or recommendation engines rather than providing simple, interpretable fare insights for everyday users [5].

Despite progress in both commercial tools and research-based ML systems, there is a clear gap between them. Existing platforms have focused more on data presentation rather than actionable fare prediction, while research prototypes require complex inputs, like multi-day historical trends or external market data. In this regard, the current study proffers a Random Forest–based flight ticket price prediction system that uses structured travel features to predict flight fares with great accuracy. The full model configuration is documented in the preprocessing and training procedures in the implementation script of this project, known as train_model.py [6].

III. METHODOLOGY

A. Dataset Description

The dataset used in this study includes 300,153 records of flights and 12 features associated with domestic Indian air travel .

The key attributes include:

- airline – airline operating the flight
- source_city – city of departure
- destination_city – city of arrival
- departure_time – time slot of departure
- arrival_time – time slot of arrival
- stops – number of layovers
- class - Economy or Business
- duration - total travel duration
- days_left - days between the booking date and departure
- price - ticket fare target variable

In addition, an extra column, Unnamed: 0, was removed as it represented an index and carried no learning value. The flight column was also dropped because it served as an identifier and has no bearing on prediction**.**

B. Data Preprocessing

1. Removal of Irrelevant Attributes:

Non-informative columns such as Unnamed: 0 (index column) and flight (unique identifier) were removed as they did not contribute to the prediction task.

2. Handling Missing Values:

Records containing missing values in the target variable price were eliminated to ensure that the model was trained only on valid and complete samples.

3. Categorical Feature Encoding

The dataset contained several categorical variables such as:airline,source_city,destination_city,departure_time,arrival_time,stops,class .As machine-learning algorithms operate on numerical representations, all categorical attributes were converted into integer-based encoded values using Label Encoding.Each unique category in a column was mapped to a corresponding integer label.After encoding, the dataset contained 10 numerical features, enabling compatibility with all machine-learning models used in the study.

4. Exploratory Data Analysis (EDA)

To understand the underlying patterns in the data, several exploratory analyses were conducted:Distribution of ticket prices revealed a right-skewed distribution, indicating that most fares fall within a lower range with fewer high-priced outliers.Box plots comparing airlines, travel class, number of stops, and cities with respect to price showed significant variation across categories.A correlation heatmap was generated to study relationships among numerical variables.Correlation analysis indicated that:class exhibited the strongest negative correlation with ticket price (−0.93),airline and duration showed moderate positive correlation,while cities exhibited very low correlation values.These insights guided the selection of predictive models.

5. Feature Selection:

The final set of independent variables included airline, source_city, destination_city, departure_time, arrival_time, stops, class, duration, and days_left. The target variable was price.

6. Train–Test Split:

The dataset was divided into training and testing sets using an 80:20 ratio. A total of 240,122 samples were used for training and 60,031 samples for testing. A fixed random seed (42) ensured the reproducibility of results.

C. Model Training and Evaluation

1. Model Selection:

Two supervised machine-learning algorithms were selected for training and performance comparison:

- Linear Regression – a baseline linear model used to understand linear relationships between features and ticket prices
- Random Forest Regressor – a powerful ensemble-based non-linear algorithm capable of capturing complex feature interactions.

These two models were chosen for their complementary nature and computational efficiency over extremely large datasets

2. Model Training Procedure

Both models were trained using the 80% training split (240,122 samples). The independent variables included nine numerical features obtained after preprocessing and encoding, while price was used as the dependent variable.

The Random Forest model was trained with the following hyperparameters:

- Number of trees (n_estimators) = 80
- Maximum depth (max_depth) = 12
- Random state = 42

The Linear Regression model used default scikit-learn parameters

3. Evaluation Metrics

Model performance was evaluated on the 20% test set (60,031 samples) using common regression metrics:

- MAE (Mean Absolute Error)
- RMSE (Root Mean Squared Error)
- R^2 Score (Coefficient of Determination)
- Accuracy (converted $R^2 \times 100$)

These metrics provide a comprehensive assessment of prediction error and model reliability.

4. Experimental Results

The results produced by each model are summarized below:

- Linear Regression
  - R^2 Score: 0.9046 (90.46%)
  - MAE: 4624.99
  - RMSE: 7014.31

Linear Regression served as a useful baseline model; however, it struggled to model the non-linear relationships present in flight pricing.

- Random Forest Regressor
  - R^2 Score: 0.9695 (96.95%)
  - MAE: 2137.69
  - RMSE: 3967.47

The Random Forest model significantly outperformed Linear Regression, yielding lower errors and a higher explanatory power — demonstrating its ability to capture complex price-determining factors such as airline, number of stops, class, and flight duration.

5. Model Comparison

Model	**Evaluation Metrics**			
	R^2 Score	***Accuracy (%)***	***MAE***	***RMSE***
Linear Regression	0.9046	90.46%	4624.99	7014.31
Random Forest	0.9695	96.95%	2137.69	3967.47

The comparison clearly shows that the Random Forest Regressor provides superior performance across all evaluation metrics.

6. Final Model Selection

Based on the R^2 score and overall error reduction, the Random Forest model was selected as the final predictive model.

The trained model was saved for deployment using Html,CSS,JAVAScript .Additionally, the list of input feature columns was saved for consistent future inference.

D. Frontend and Deployment

1. FRONTEND DESIGN (HTML, CSS, JAVASCRIPT)

- The user interface of the flight price prediction system was developed using standard web technologies—HTML, CSS, and JavaScript.
- HTML was used to create the structural layout of the input form where users can select:
- Airline,Source and destination cities,Departure and arrival time,Number of stops,Travel class,Days left before departure
- CSS was used to enhance the visual appearance, maintain responsiveness, and provide a clean user-friendly design.JavaScript handled client-side validation and ensured smooth interaction between the input form and the backend API.The interface allows users to enter flight-related details and receive the predicted ticket price .

2. Backend Server (Flask Framework)

- The backend of the system was implemented using Flask, a lightweight Python web framework.
- Flask handled:
- Routing between the frontend and model
- Receiving user input from the HTML form
- Loading the trained Random Forest model (.pkl file)
- Preprocessing the input data
- Generating the predicted ticket fare
- Sending the prediction back to the frontend
- Flask's simplicity, low overhead, and seamless integration with machine-learning models made it an ideal choice for deployment.

E.Model Integration

The trained Random Forest model and corresponding encoded feature columns were stored using joblib.

- During user interaction:

Flask receives input data from the HTML form.The values are preprocessed and encoded using the same method as training.The model predicts the fare.Flask returns the predicted result to the frontend interface.The result is displayed in the browser using HTML and JavaScript.This architecture makes the system lightweight and suitable for deployment on low-cost cloud servers.

F.End-to-End Workflow

1. User selects inputs on the HTML page
2. Data is sent via POST request to Flask

3,Flask loads the Random Forest model

4.Model predicts the price

5.Result appears on the webpage in real-time

1V. RESULTS AND DISCUSSION

A. Quantitative Results

The Linear Regression model achieved an accuracy of 90.46%, indicating that a majority of the variance in ticket prices could be explained by the linear relationships between features. However, the error values (MAE $\approx$ 4625 and RMSE $\approx$ 7014) show that the model struggled with capturing non-linear patterns present in flight pricing, especially for higher fares.

In contrast, the Random Forest Regressor achieved a significantly higher accuracy of 96.95% with substantially lower error values (MAE $\approx$ 2138 and RMSE $\approx$ 3967). This demonstrates the ability of ensemble-based non-linear models to better capture compatibility.

B. Discussion

- Capture non-linear trends in airline pricing
- Handle high-dimensional and categorical data effectively
- Reduce overfitting through the averaging of multiple decision trees
- Provide robustness to noisy and unbalanced data
- The model's R^2 score of 0.9695 indicates that it explains nearly 97% of the variability in flight ticket prices, making it reliable for practical fare prediction applications.

V.CONCLUSION

This research successfully developed a machine-learning-based flight ticket price prediction system using a dataset of 300,153 domestic flight records. The data underwent extensive preprocessing, including removal of irrelevant attributes, encoding of categorical variables, and exploratory analysis. Two regression algorithms were implemented and compared.Among the evaluated models, the Random Forest Regressor emerged as the most accurate, achieving an R^2 score of 0.9695 and outperforming the Linear Regression model across all metrics. This confirms that non-linear ensemble techniques are highly suitable for modeling the complex pricing structure of airline fares.The final model was integrated into a web-based interface built using HTML, CSS, JavaScript, and Flask, enabling users to predict flight prices in real time based on airline, route, class, timing, stops, and days left for departure.Overall, the system provides a reliable decision-support tool that can assist travelers in understanding price patterns and booking flights more efficiently.

REFERENCES

[1] [1] MakeMyTrip, "Flight Search & Price Comparison Features," Accessed: Nov. 2025. [Online]. Available: https://www.makemytrip.com/flights/

[2] [2] Skyscanner, "Flight Search and Fare Insights," Accessed: Nov. 2025. [Online]. Available: https://www.skyscanner.co.in/

[3] [3] S. Pawar, R. Deshmukh, and M. Patil, "Machine Learning Approaches for Flight Fare Prediction," International Journal of Computer Applications, vol. 182, no. 45, pp. 12–19, 2022.

[4] [4] Y. Zhou, L. Wang, and H. Li, "Airfare Prediction Using Ensemble Machine Learning Techniques," Journal of Travel Analytics, vol. 4, no. 2, pp. 33–46, 2023.

[5] [5] A. Sharma and P. Kaur, "Feature Engineering for Predictive Flight Pricing Models," Proceedings of the International Conference on Data Science, 2023, pp. 101–110.

[6] [6] FlightPricePredictor Project, "Model training and preprocessing pipeline," train.py, 2025.

[7] [7] FlightPricePredictor Dataset, "Preprocessor configuration," preprocessor.pkl, 2025.