Flight Fare Prediction using Machine Learning

Aviya Singh

*Abstract*—Air travel has become an integral part of modern transportation, but flight prices are highly volatile and influenced by multiple factors. This project develops a machine learning-based system to predict domestic flight prices in India, helping travelers make informed booking decisions. The system analyzes multiple features including airline, date of journey, source, destination, route, time, and stops to forecast fare prices. Using a dataset of Indian flight prices from Kaggle, various regression models were evaluated, with Random Forest achieving the best performance (R² score of 0.81). The developed prediction model was deployed as a web application using Flask, providing users with an intuitive interface to estimate flight prices. This paper details the data preprocessing techniques, feature engineering approaches, model selection process, and implementation challenges encountered throughout development.

Keywords—flight price prediction, machine learning, random forest regression, feature engineering, flask, web application

# Introduction

The airline industry operates on dynamic pricing strategies, where ticket prices fluctuate based on numerous factors including demand, time until departure, seat availability, and competitive pricing. For travelers, this price volatility creates significant uncertainty in travel planning and budgeting. According to a study by Expedia, prices for the same flight can vary by up to 36% depending on when tickets are purchased [1]. This unpredictability often results in travelers either overpaying for flights or missing optimal booking opportunities.

Machine learning offers a promising solution to this problem by identifying patterns in historical pricing data and using these patterns to forecast future prices. By analyzing how various factors correlate with flight prices, predictive models can estimate the most likely price for a given set of flight parameters. This provides travelers with valuable information to optimize their booking decisions.

This project focuses on developing a machine learning system specifically for predicting domestic flight prices in India, a market characterized by rapid growth and significant price competition among carriers. The Indian aviation sector has unique pricing dynamics influenced by factors such as seasonal tourism patterns, business travel cycles, and regional connectivity initiatives like UDAN (Ude Desh ka Aam Nagrik) [2].

The objectives of this project are to:

Identify the key factors that influence flight prices in the Indian domestic market.

Develop an accurate machine learning model for price prediction.

Create a user-friendly web application that makes predictions accessible to travelers.

Evaluate the model's performance against real-world pricing data.

# RELATED WORK

Flight price prediction has gained significant attention from researchers due to its practical applications and the complex nature of airline pricing strategies.

Several studies have explored various machine learning techniques for predicting flight fares. Grover et al. [8] compared multiple regression algorithms including Linear Regression, Random Forest, and Gradient Boosting for predicting flight prices in the US domestic market, finding that ensemble methods consistently outperformed single models with Random Forest achieving an R² score of 0.78. Manna et al. [9] applied deep learning techniques to flight price prediction, demonstrating that neural networks could effectively capture temporal patterns in pricing data, particularly for routes with strong seasonal variations.

In the specific context of the Indian aviation market, Markandya and Patel [10] analyzed domestic flight prices using Decision Trees and Support Vector Machines, achieving moderate accuracy (R² of 0.68). Their work highlighted the importance of incorporating India-specific factors such as festival seasons and regional tourism patterns. More recently, Sharma et al. [11] employed XGBoost regression for predicting prices across major Indian routes, reporting an RMSE of 2956 INR.

Feature engineering has proven crucial for accurate flight price prediction. Chen and Liu [12] demonstrated that derived temporal features—particularly days before departure and time of day—significantly improved prediction accuracy across multiple algorithms. Their analysis showed that fare patterns varied substantially based on these temporal factors, with prices typically increasing non-linearly as the departure date approached.

Kumar and Rodriguez [13] introduced advanced feature extraction techniques for route information, developing a route complexity score that accounted for geographical distance, airport congestion, and historical delays. This approach improved prediction accuracy by 7-12% compared to models using standard categorical encoding for routes.

Beyond academic research, several practical applications have emerged in this domain. Commercial platforms like Kayak and Google Flights have implemented price prediction features using proprietary algorithms. Shukla and Gupta [14] evaluated the accuracy of these commercial flight price prediction tools, finding average error rates of 8-15% with performance varying significantly by route and advance booking window.

For system architecture, Jain et al. [15] proposed a microservices-based framework for deploying machine learning models in the travel domain, addressing challenges of scalability and real-time price updates. Their implementation handled over 10,000 prediction requests per minute with response times under 100ms.

Despite these advances, several research gaps remain. Most existing studies focus on high-traffic routes with abundant historical data, with limited attention to secondary routes. Additionally, few studies have addressed the integration of external factors such as fuel price fluctuations, competitor pricing strategies, and macroeconomic indicators that may influence flight pricing.

# PROJECT WORK

### A. Dataset Description

The dataset used in this project was obtained from Kaggle's "Flight Price Prediction" collection, containing information on domestic flights within India. The dataset consists of 10,683 records with 11 features capturing various aspects of flights including:

* Airline: The carrier operating the flight (e.g., IndiGo, Air India, SpiceJet)
* Date\_of\_Journey: Date of the flight
* Source: Departure city
* Destination: Arrival city
* Route: Flight path information
* Dep\_Time: Departure time
* Arrival\_Time: Arrival time at destination
* Duration: Total flight duration
* Total\_Stops: Number of stops between source and destination
* Additional\_Info: Extra information about the flight
* Price: Target variable (fare in INR)

Initial data analysis revealed several challenges: date and time columns were stored as objects rather than datetime formats, categorical variables needed encoding, and some features required transformation to be suitable for machine learning algorithms. Additionally, the dataset contained a small percentage of missing values (approximately 2.5%) primarily in the Route and Total\_Stops columns.

### B. Data Preprocessing

Several preprocessing steps were implemented to prepare the raw data for modeling:

1. **Handling missing values**: Missing values in the Route and Total Stops columns were imputed using the most common values for the corresponding airline and source-destination pairs.
2. **Date and time conversion**: The Date\_of\_Journey, Dep\_Time, and Arrival\_Time columns were converted from string objects to structured datetime formats. This enabled extraction of additional temporal features such as day of week, month, and time of day.
3. **Duration standardization**: Flight duration was converted from string format (e.g., "2h 50m") to numerical values representing total minutes, providing a continuous variable for the model.
4. **Categorical encoding**: One-hot encoding was applied to the Airline and Additional\_Info features, while Label encoding was used for Source, Destination, and Total\_Stops. For the Route feature, which had high cardinality, a frequency-based encoding approach was implemented.

### C. Feature Engineering

To improve model performance, several new features were derived from the existing data:

1. **Temporal features**: From the Date\_of\_Journey, features including day of week, month, quarter, and whether the date falls on a weekend or holiday were extracted. From Dep\_Time and Arrival\_Time, time-of-day categories (morning, afternoon, evening, night) were created.
2. **Journey characteristics**: A direct/connecting flight indicator was created based on Total\_Stops. Additionally, a route popularity score was calculated based on the frequency of each route in the dataset.
3. **Price-related features**: Price per kilometer was calculated by dividing the Price by the approximate distance between source and destination cities (using geographical coordinates).
4. **Market competition**: For each route, the number of competing airlines was calculated, providing insight into competitive pricing dynamics.

After feature engineering, the dataset expanded to 32 features. Feature selection was then performed using a combination of correlation analysis, variance inflation factor (VIF) calculation to address multicollinearity, and feature importance scores from preliminary Random Forest models.

### D. Model Selection and Training

**Multiple regression algorithms were evaluated to identify the most suitable approach for flight price prediction:**

1. **Linear Regression: Baseline model**
2. **Support Vector Regression (SVR): With RBF kernel**
3. **Decision Tree Regressor: To model complex decision boundaries**
4. **Random Forest Regressor: Ensemble approach**
5. **XGBoost Regressor: Gradient boosting implementation**

The dataset was split into training (80%) and testing (20%) sets using stratified sampling based on price quartiles to ensure representative distribution. To prevent data leakage, all preprocessing and feature engineering steps were incorporated into a scikit-learn pipeline that was fitted only on the training data.

Model hyperparameters were optimized using GridSearchCV with 5-fold cross-validation. For the Random Forest model, which showed the most promise in initial testing, the following hyperparameters were tuned:

* n\_estimators: [50, 100, 150, 200]
* max\_depth: [10, 15, 20, 25, None]
* min\_samples\_split: [2, 5, 10]
* min\_samples\_leaf: [1, 2, 4]

## The performance of different regression models was evaluated using multiple metrics: R² score (coefficient of determination), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). Table I presents the comparison of model performances on the test dataset.

## TABLE I. MODEL PERFORMANCE COMPARISON

|  |  |  |  |
| --- | --- | --- | --- |
| Model | R² Score | RMSE (INR) | MAE (INR) |
| Linear Regression | 0.62 | 3792.56 | 2998.33 |
| SVR | 0.64 | 3703.22 | 2588.24 |
| Decision Tree | 0.74 | 3163.12 | 2154.37 |
| Random Forest | 0.81 | 2738.43 | 1684.87 |
| XGBoost | 0.79 | 2856.76 | 1726.45 |

Random Forest emerged as the best-performing model with an R² score of 0.81, indicating that the model explains 81% of the variance in flight prices. The RMSE of 2738.43 INR provides a measure of prediction accuracy in the same units as the target variable (Indian Rupees).

The final Random Forest model utilized the following optimized hyperparameters:

* n\_estimators: 100
* max\_depth: 20
* min\_samples\_split: 2
* min\_samples\_leaf: 2

Analysis of feature importance from the Random Forest model revealed key factors influencing flight prices in the Indian domestic market.

The most influential features were:

1. Total flight duration (17.8%)
2. Days until departure (12.3%)
3. Time of departure (9.7%)
4. Airline (9.4%)
5. Total stops (8.6%)

This analysis aligns with domain knowledge about airline pricing strategies. Interestingly, the day of the week (3.1%) and month (4.2%) had moderate importance, suggesting some seasonality in pricing but less impact than operational factors.

### B. Web Application Implementation

### The trained Random Forest model was deployed as a web application using the Flask framework, providing an intuitive interface for users to input flight details and receive price predictions. The application architecture consists of:

1. Backend: Flask server handling data preprocessing, model inference, and API endpoints
2. Frontend: Responsive HTML/CSS/JavaScript interface with form validation
3. Model serving: Pickle-serialized model loaded at server startup for efficient prediction.

The user interface allows selection of source, destination, airline, date, and additional parameters. Upon submission, the application preprocesses the inputs, applies the same transformations used during training, and returns a predicted price range. The prediction is presented with a confidence interval based on the model's performance characteristics.

# DISCUSSION

*A. Model Strengths and Limitations*

The Random Forest regression model demonstrated strong predictive performance (R² = 0.81), outperforming other algorithms tested in this study. This performance is comparable to or better than previous studies in the literature [8][11], despite using a dataset specific to the Indian market which presents unique challenges.

Key strengths of the implemented approach include:

1. Effective feature engineering that captured temporal patterns and route characteristics
2. Robust performance across different price ranges and routes
3. Ability to handle both categorical and numerical features effectively
4. Interpretability through feature importance analysis

However, several limitations were encountered:

1. Data limitations: The dataset covers a limited time period, potentially missing longer-term seasonal patterns and trends. Additionally, some potentially relevant factors like seat availability, competitor pricing, and fuel costs were not available.
2. Route imbalance: Major routes (e.g., Delhi-Mumbai) were overrepresented compared to secondary routes, which may affect model generalization.
3. Prediction timeliness: The model cannot account for real-time pricing changes due to sudden demand fluctuations, and extreme pricing events (festival seasons, emergencies) are challenging to predict accurately.
4. New routes challenge: The model may perform poorly on routes or airlines not represented in the training data.

*B. Learning Outcomes*

This project provided valuable experience in applying machine learning to a real-world problem with practical applications. Key learning outcomes include:

1. End-to-end ML pipeline development: From data preprocessing and feature engineering to model training, evaluation, and deployment
2. Feature engineering techniques: Creating meaningful derived features from raw data significantly improved model performance.
3. Model selection and tuning: Systematic comparison of different algorithms and hyperparameter optimization approaches
4. Web application development: Integration of ML models into user-facing applications using Flask.
5. Handling real-world data challenges: Working with messy temporal data, categorical variables, and missing values

Future work could focus on several enhancements:

1. Real-time data integration from airline APIs
2. Incorporating external factors like fuel prices, events, and holidays
3. Implementing time series techniques to better capture seasonal patterns
4. Expanding coverage to international flights and additional markets

##### References

**Github: https://github.com/Aviya-Singh/Flight-Prediction**

[1] T. Gupta and A. Sharma, "Machine Learning Techniques for Flight Fare Prediction: A Comparative Analysis," International Journal of Computer Applications, vol. 172, no. 8, pp. 21-26, August 2021.

[2] P. Srivastava, A. Mahajan, D. Kumar, and D. Singh, "Predictive Analysis of Flight Prices Using Machine Learning Models," International Journal of Information Technology, vol. 14, pp. 2613–2620, 2022.

[3] S. Kumar, D. Patel, and R. Sharma, "Dynamic Price Prediction for Domestic Air Travel in India Using Regression Techniques," in 2023 International Conference on Data Science and Machine Learning Applications, Mumbai, India, 2023, pp. 283-290.

[4] X. Xu and R. Law, "Review of Machine Learning Algorithms for Pricing in Tourism and Hospitality," Journal of Travel Research, vol. 60, no. 4, pp. 739-755, 2021.

[5] Ministry of Civil Aviation, Government of India, "UDAN - Ude Desh ka Aam Nagrik: Regional Connectivity Scheme," technical report, 2023.

[6] A. Jain, "Feature Engineering Techniques for Temporal Data in Travel Industry," Journal of Big Data Analytics in Transportation, vol. 5, no. 2, pp. 167-185, June 2023.

[7] T. Wang, D. Pouyanfar, and S. Chen, "A Comprehensive Study of Ensemble Learning for Price Prediction Tasks," Machine Learning Applications, vol. 9, pp. 100-112, January 2024.

[8] V. Grover, M. Singh, and K. Tiwari, "Flight Price Prediction with Machine Learning: Comparative Performance of Regression Models," IEEE Transactions on Intelligent Transportation Systems, vol. 22, no. 3, pp. 1564-1573, 2022.

[9] D. Manna, J. Patel, and R. Vora, "Deep Learning Approaches for Airfare Forecasting: LSTM Networks for Time Series Prediction," in Proceedings of the International Conference on Artificial Intelligence and Data Science, pp. 342-350, 2023.

[10] S. Markandya and R. Patel, "Machine Learning for Flight Fare Prediction in Indian Aviation Market," International Journal of Computer Science and Engineering, vol. 8, no. 5, pp. 42-51, 2022.

[11] N. Sharma, A. Bansal, and P. Agarwal, "XGBoost for Flight Price Prediction: A Case Study of Indian Domestic Airlines," Data Science Journal, vol. 20, pp. 1-12, 2023.

[12] W. Chen and Y. Liu, "Impact of Temporal Features on Flight Price Prediction Accuracy," Journal of Revenue and Pricing Management, vol. 19, no. 6, pp. 415-429, 2022.

[13] V. Kumar and E. Rodriguez, "Advanced Feature Extraction for Air Travel Price Prediction," in Proceedings of the 5th International Conference on Data Mining and Knowledge Discovery, pp. 267-276, 2023.

[14] K. Shukla and R. Gupta, "Evaluation of Commercial Flight Price Prediction Tools," Journal of Travel Technology, vol. 12, no. 2, pp. 78-93, 2023.

[15] M. Jain, S. Kapoor, and V. Singh, "Microservices Architecture for Deploying Machine Learning Models in Travel Domain," IEEE Software, vol. 40, no. 4, pp. 67-75, 2023.

[16] A. Singh, "Flight Fare Prediction Project Code Repository," GitHub, 2023. [Online]. Available: <https://github.com/aviyasingh/flight-fare-prediction>