**Flight Fare Prediction Using Machine Learning**

By 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. The resulting system demonstrates the practical application of machine learning in addressing real-world problems in the travel industry.

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

**I. 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:

1. Identify the key factors that influence flight prices in the Indian domestic market.
2. Develop an accurate machine learning model for price prediction.
3. Create a user-friendly web application that makes predictions accessible to travelers.
4. Evaluate the model's performance against real-world pricing data.

By achieving these objectives, the project aims to empower travelers with data-driven insights that can potentially reduce travel costs and improve planning efficiency. Additionally, the methodology and findings may contribute to the broader understanding of price prediction techniques applicable to other markets and transportation sectors.

**II. METHODOLOGY**

**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. This domain-specific approach preserved the data distribution better than simple statistical imputation.
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). This feature helped normalize prices across different route distances.
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 to establish minimum performance expectations
2. **Support Vector Regression (SVR)**: With RBF kernel to capture non-linear relationships
3. **Decision Tree Regressor**: To model complex decision boundaries
4. **Random Forest Regressor**: Ensemble approach to improve generalization and handle feature interactions
5. **XGBoost Regressor**: Gradient boosting implementation known for performance in regression tasks

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]

**III. RESULTS AND DISCUSSION**

**A. Model Performance Comparison**

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 |
| Support Vector Regression | 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

**B. Feature Importance Analysis**

Analysis of feature importance from the Random Forest model revealed key factors influencing flight prices in the Indian domestic market. Fig. 1 illustrates the top 10 features by importance.

**Fig. 1. Top 10 Features by Importance in the Random Forest Model**

The most influential features were:

1. **Total flight duration** (17.8%): Longer flights generally cost more due to increased operational costs
2. **Days until departure** (12.3%): Prices typically increase as the departure date approaches
3. **Time of departure** (9.7%): Early morning and late evening flights often cost less
4. **Airline** (9.4%): Premium carriers charge higher prices than budget airlines
5. **Total stops** (8.6%): Non-stop flights command premium prices

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.

**C. 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 and interactive elements
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.

**IV. CHALLENGES AND LIMITATIONS**

Several challenges were encountered during the development of this flight price prediction system:

**A. Data Challenges**

1. **Temporal constraints**: The dataset covers a limited time period, potentially missing longer-term seasonal patterns and trends
2. **Feature coverage**: Some potentially relevant factors like seat availability, competitor pricing, and fuel costs were not available in the dataset
3. **Route imbalance**: Major routes (e.g., Delhi-Mumbai) were overrepresented compared to secondary routes

**B. Technical Challenges**

1. **Categorical encoding**: High cardinality in the Route feature required custom encoding strategies
2. **Hyperparameter tuning**: The extensive hyperparameter space required significant computational resources for optimization
3. **Memory constraints**: One-hot encoding of categorical variables with many levels created high-dimensional sparse matrices
4. **Deployment complexities**: Ensuring consistent preprocessing between training and inference environments

**C. Model Limitations**

1. **Prediction lag**: The model cannot account for real-time pricing changes due to sudden demand fluctuations
2. **Price outliers**: Extreme pricing events (festival seasons, emergencies) are challenging to predict accurately
3. **New routes**: The model may perform poorly on routes or airlines not represented in the training data

Despite these challenges, the implemented system provides valuable price estimates with reasonable accuracy for most common domestic routes in India.

**V. CONCLUSION AND FUTURE WORK**

This project successfully developed a machine learning system for predicting domestic flight prices in India with good accuracy (R² score of 0.81). The Random Forest regression model, deployed as a user-friendly web application, provides travelers with a practical tool for estimating flight costs based on various parameters.

Key contributions of this work include:

1. Identification of the most influential factors affecting flight prices in the Indian market
2. Development of an effective feature engineering pipeline for flight data
3. Comparative analysis of regression algorithms for price prediction
4. Creation of an accessible web interface for non-technical users

Future work could focus on several enhancements:

1. **Real-time data integration**: Incorporating live pricing data from airline APIs to improve prediction timeliness
2. **Additional features**: Including external factors like fuel prices, events, and holidays
3. **Temporal modeling**: Implementing time series techniques to better capture seasonal patterns
4. **Personalization**: Developing user profiles to customize predictions based on individual travel histories
5. **Expanded coverage**: Extending the model to international flights and additional markets

As airline pricing becomes increasingly complex, machine learning approaches like the one developed in this project offer valuable tools for consumers to navigate price volatility and make more informed travel decisions.

**REFERENCES**

[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. https://doi.org/10.1007/s41870-022-01040-6

[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. https://doi.org/10.1177/0047287520934880

[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.