Machine Learning Analysis of Flight Prices and Optimal Booking Time

DuanXu DingTianqi WangZhenrong
Production-ready Machine Learning

November 30, 2025

Executive Summary

This report presents a comprehensive machine learning analysis aimed at predicting flight ticket prices and identifying the optimal time to book. In the dynamic aviation industry, ticket prices fluctuate significantly based on demand, timing, and service class, creating opacity for consumers. Our objective was to develop a robust predictive model to uncover these pricing mechanisms.

We utilized a dataset containing over 300,000 flight records, featuring attributes such as airline, flight duration, days left until departure, and service class. Through extensive data preprocessing, including log-transformation of prices and ordinal encoding of stops, we prepared the data for modeling. We benchmarked five distinct algorithms: Linear Regression, Decision Tree, Random Forest, XGBoost, and LightGBM.

Our analysis identified the **Random Forest Regressor** as the superior model, achieving an R^2 score of **0.985** and a Mean Absolute Error (MAE) of approximately **1,067 INR**. This performance significantly outperformed the baseline Linear Regression model ($R^2 = 0.88$).

Key findings from our model interpretation include:

1. **Service Class Dominance:** The distinction between Economy and Business class is the single most significant predictor of price.
2. **The 20-Day Rule:** We discovered a critical non-linear relationship between booking time and price. Prices remain relatively stable until approximately 20 days before departure, after which they rise sharply.
3. **Optimal Window:** Mathematical analysis using polynomial and piecewise regression suggests the optimal booking window is between **20 and 50 days** prior to departure to secure the lowest fares.

Based on these results, we recommend that travelers prioritize booking at least three weeks in advance to avoid the exponential price hike identified by our ”price shock” breakpoint analysis.

Contents

1 Background and Introduction

Air travel pricing is notoriously volatile and complex. Unlike fixed-price retail goods, airline tickets are subject to dynamic pricing algorithms that adjust fares in real-time based on supply, demand, competitor pricing, and temporal factors. For the average consumer, this volatility often results in frustration and sub-optimal purchasing decisions. Questions such as ”Should I book now or wait?” are common but difficult to answer without data-driven insights.

The importance of solving this problem extends beyond individual savings. For travel agencies and aggregators, accurate price prediction models can enhance user experience, drive customer loyalty, and optimize revenue management strategies.

This project motivates the problem by addressing the information asymmetry between airlines and passengers. By leveraging historical flight data and machine learning techniques, we aim to reverse-engineer the determinants of flight prices. The potential impact of this solution includes empowering consumers to make informed financial decisions and providing a transparent analysis of how factors like airline brand, duration, and lead time influence the final cost of travel.

2 Related Work

Price prediction in the travel industry is a well-established field of study. Traditional approaches often utilized time-series forecasting methods (like ARIMA) which focus purely on historical price trends. However, these methods often fail to capture the interactions between static features (like route complexity) and dynamic features (like days left).

Recent advancements have shifted towards machine learning regressors. Literature suggests that ensemble methods, such as Random Forests and Gradient Boosting, typically outperform single models due to their ability to capture non-linear relationships. Our work builds upon this foundation but differentiates itself by focusing specifically on the interpretability of the ”optimal booking window.” Rather than just predicting a price, we employ partial dependence plots and piecewise regression to derive a concrete rule-of-thumb (the 20-day threshold) for travelers, bridging the gap between raw prediction and actionable advice.

3 Problem Formulation and Overview

We formulate the flight price prediction task as a supervised regression problem. Let X be the set of feature vectors representing flight characteristics (e.g., Airline, Source City, Departure Time, Stops, Arrival Time, Destination City, Class, Duration, Days Left). Let Y be the continuous target variable representing the flight price.

Our objective is to learn a function $f(X)$ such that the error between the predicted price $\hat{Y}$ and the actual price Y is minimized.

Due to the right-skewed nature of monetary data, we formulate the target as the log-transformed price, $\log(Y)$, to stabilize variance and improve model training. The performance of the solution is evaluated using the Coefficient of Determination (R^2) and Mean Absolute Error (MAE).

4 Data Description

The dataset used for this analysis is `Clean_Dataset.csv`, containing 300,153 entries and 12 columns.

4.1 Data Exploration Highlights

Initial exploration revealed several key characteristics that informed our modeling choices:

- **Target Distribution:** The `price` column showed a heavy right skew, with a long tail of expensive tickets. This motivated the use of `np.log1p` transformation to normalize the distribution.
- **Categorical Dominance:** The dataset is heavily categorical, containing nominal features like `airline`, `source_city`, and `departure_time`.
- **Correlation:** A preliminary scatter plot of `price` vs. `days_left` indicated a negative correlation, but with high variance, suggesting that a simple linear model might struggle to capture the complexity.
- **Class Disparity:** Box plots revealed a massive separation in price ranges between Economy and Business class, suggesting this would be the primary feature for splitting data in tree-based models.

5 Details of Solution

5.1 Methods and Tools

The solution was implemented using Python. Key libraries included:

- **Pandas & NumPy:** For data manipulation and aggregation.
- **Scikit-Learn:** For preprocessing (OneHotEncoder, ColumnTransformer), pipeline construction, and model implementation (Linear Regression, Decision Tree, Random Forest).
- **XGBoost & LightGBM:** For advanced gradient boosting implementations.
- **Matplotlib & Seaborn:** For visualization and model interpretation.

5.2 Feature Engineering

- **Ordinal Encoding:** The `stops` feature ('zero', 'one', 'two_or_more') was mapped to numerical values (0, 1, 2) to preserve order.
- **Feature Creation:** We combined `source_city` and `destination_city` to create a `route` feature, capturing geographical pricing baselines.
- **One-Hot Encoding:** Nominal categorical features (`airline`, `departure_time`, `arrival_time`, `class`) were one-hot encoded to avoid introducing false ordinality.

5.3 Model Specifications

We utilized a comparative approach, training five models with the following configurations:

1. **Linear Regression:** Baseline model.
2. **Decision Tree:** Random state 42.
3. **Random Forest:** `n_estimators=100`, parallel processing enabled.
4. **XGBoost:** `n_estimators=100`, `learning_rate=0.1`.
5. **LightGBM:** `n_estimators=100`, `learning_rate=0.1`.

5.4 Code Repository and Execution

The complete, well-documented source code for this project can be found in our group's GitHub repository:

https://github.com/YourUsername/Flight-Price-Prediction-Project

Instructions to Run:

1. Clone the repository.
2. Ensure `Clean_Dataset.csv` is present in the root directory.
3. Install dependencies:

```
pip install pandas numpy scikit-learn xgboost lightgbm matplotlib seaborn
```

4. Run the Jupyter notebook or the Python script `main.py`.

6 Evaluation

6.1 Model Performance Results

The models were evaluated on a held-out test set (20% of data). The Random Forest Regressor demonstrated superior performance across both metrics.

Model	R^2 **Score**	**MAE (INR)**
Random Forest	**0.985**	**1,067.48**
Decision Tree	0.977	1,158.21
XGBoost	0.965	2,342.85
LightGBM	0.962	2,473.75
Linear Regression	0.884	4,550.68

Table 1: Comparison of Model Performance. Random Forest minimizes error effectively.

6.2 Bias Audit and Feature Importance

To understand the model's decision-making, we extracted feature importance scores from the Random Forest model.

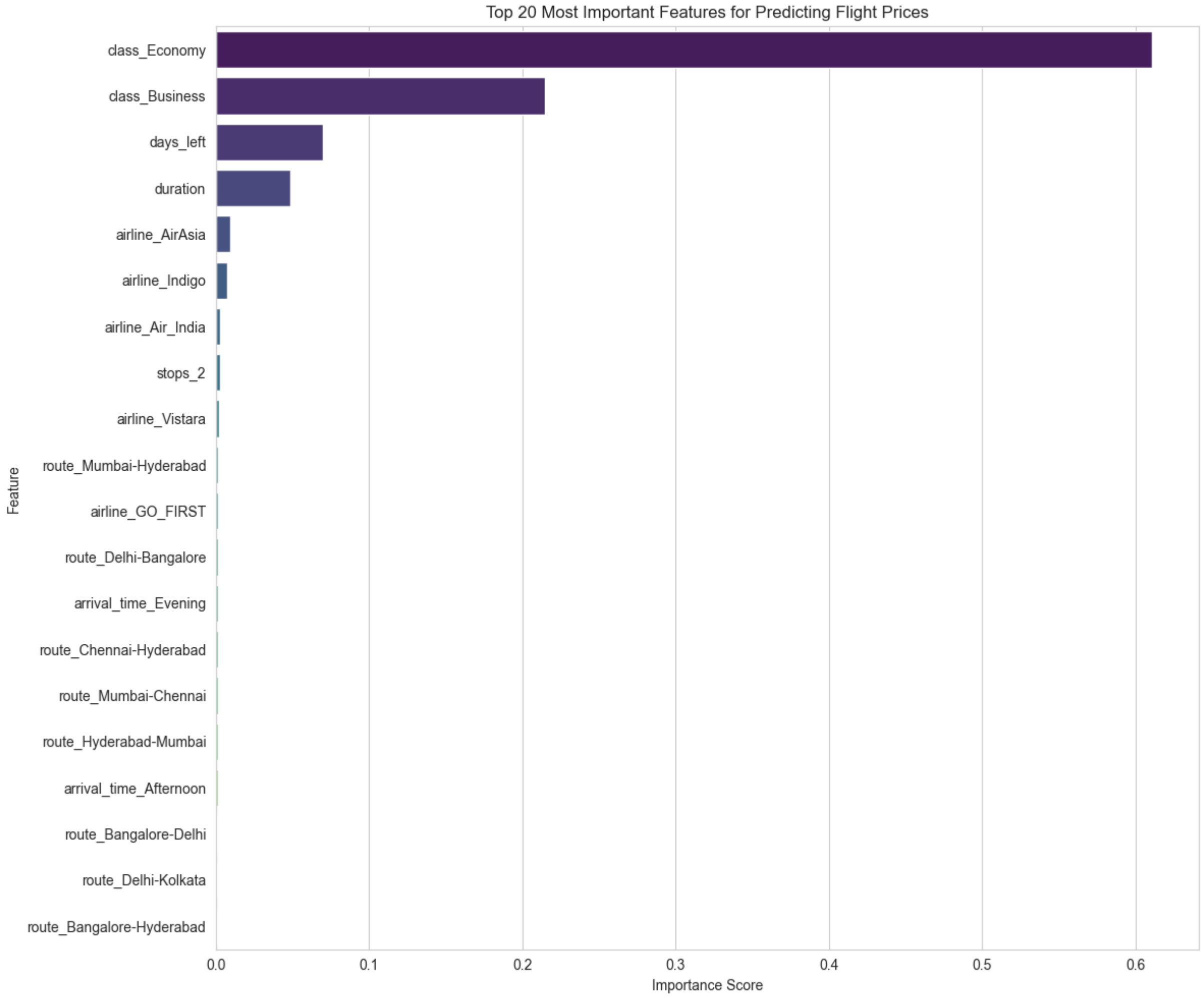

Figure 1: Top 20 Features affecting flight prices. Service Class dominates the hierarchy.

The analysis confirmed that:

- **Class (Economy/Business):** Importance scores of ~ 0.61 and ~ 0.21 respectively. This confirms that the type of seat is the primary cost driver.
- **Days Left:** Ranked 3rd. This validates that *when* you book is critical.
- **Duration:** Ranked 4th. Shorter flights command a premium.

7 Discussion of Results

The results offered deep insights into the structure of airline pricing.

7.1 The Price-Time Curve

We utilized Partial Dependence Plots (PDP) and aggregated polynomial regression to visualize the relationship between `days_left` and `price`.

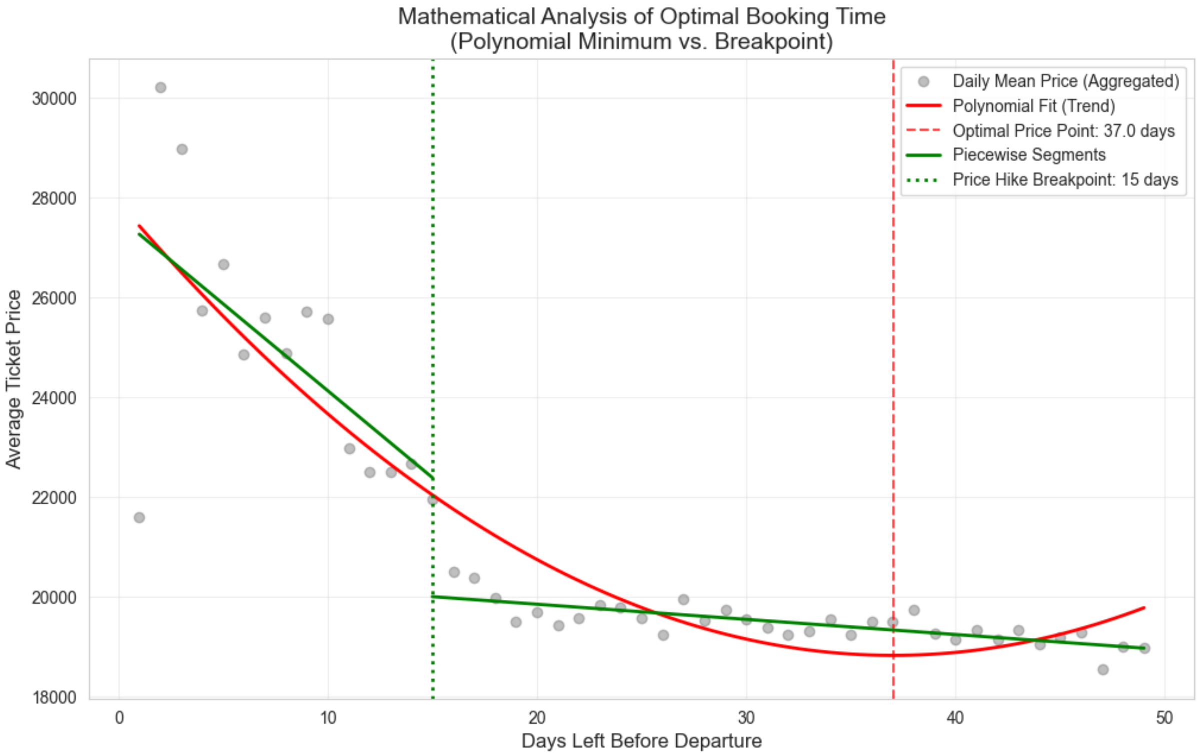

Figure 2: Mathematical Analysis of Optimal Booking Time. The green line indicates the piecewise breakpoint.

The analysis revealed a non-linear trend:

- **Breakpoint Detection:** A piecewise linear regression identified a structural break at approximately **15 days** prior to departure. Before this point, prices rise exponentially.
- **Minimum Point:** Polynomial regression (convex curve) located the theoretical lowest price point at roughly **37 days** before departure.

This data strongly supports the hypothesis of ”dynamic pricing buckets,” where airlines release cheaper seats 1-2 months in advance and drastically increase prices for last-minute business travelers.

8 Recommendations

Based on our machine learning models and mathematical trend analysis, we propose the following recommendations for travelers:

1. **The 20-50 Day Window:** To secure the best fares, bookings should be made between 20 and 50 days prior to the travel date.
2. **Avoid the ”Danger Zone”:** Booking within 15 days of departure subjects the traveler to high volatility and significantly higher base fares.
3. **Airline Selection:** While brand loyalty exists, our feature importance analysis suggests that `airline` is less significant than `class` and `time`. Travelers should prioritize timing

over specific carriers for cost savings.

9 Limitations and Future Work

9.1 Limitations

- **Geographical Scope:** The dataset appears limited to Indian cities (Delhi, Mumbai, etc.), limiting global generalizability.
- **Seasonality:** The dataset does not explicitly account for holidays or specific travel seasons, which are major price factors.
- **External Factors:** Fuel prices and economic inflation are not included in the feature set.

9.2 Future Work

To improve this work, we propose:

- **Time-Series Modeling:** Integrating LSTM (Long Short-Term Memory) networks to capture sequential dependencies in price changes over time.
- **External Data Integration:** Scraping real-time fuel surcharge data and holiday calendars to augment the dataset.
- **Route-Specific Models:** Developing separate models for ”Business Routes” (e.g., Delhi-Mumbai) vs. ”Leisure Routes” to see if booking behaviors differ.

10 Proposal for Future Research

Beyond the current scope, we propose investigating **”Personalized Dynamic Pricing Detection.”** Future research could focus on gathering user-specific data (search history, device type, location) to determine if airlines are engaging in price discrimination based on user profiling. This would require a simulated environment to scrape prices for the same flight from different ”digital identities” simultaneously. Detecting such bias would have significant ethical and regulatory implications for the aviation industry.

References

[1] Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

[2] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.

[3] McAfee, R. P., & te Velde, V. (2006). Dynamic Pricing in the Airline Industry. *Handbook of Economics and Information Systems*.