# Dynamic Flight Price Forecasting Using Gradient Boosting: A Feature-Engineered Approach to Airfare Optimization

## Problem Statement

Airline ticket pricing exhibits extreme volatility driven by complex interactions between temporal factors, route characteristics, carrier strategies, and demand patterns. Consumers and travel agencies face significant economic uncertainty when planning trips, often overpaying by 20-40% due to suboptimal booking timing. Traditional fare forecasting methods rely on historical averages or simple time-series models that fail to capture non-linear relationships between multiple pricing determinants. This project develops a predictive framework for domestic flight prices in the Indian aviation market, enabling data-driven booking recommendations and price trend analysis. The challenge lies in modeling the intricate dependencies between departure timing, journey duration, airline competition, route density, and seasonal demand while maintaining prediction accuracy suitable for real-world decision support systems.

## Methodology

The solution implements an end-to-end machine learning pipeline leveraging gradient boosted decision trees with comprehensive feature engineering:

**Data Preparation and Exploration:** The dataset encompassed 10,683 domestic flight records with features including airline carrier, source-destination pairs, departure time, arrival time, flight duration, number of stops, and additional service information. Exploratory analysis revealed strong price correlations with journey duration ($r = 0.65$), number of stops ($r = 0.42$), and specific airline carriers, with premium carriers commanding 35% higher fares on average. Price distributions exhibited right-skewed patterns with significant outliers in premium cabin classes.

**Feature Engineering Strategy:** Temporal features were decomposed into granular components: departure hour, day of week, and month to capture diurnal and seasonal patterns. Duration features were converted from timestamp strings to numerical minutes for regression modeling. Categorical variables (airline, source city, destination city, stopover information) were encoded using target encoding to preserve ordinality while avoiding high-dimensional one-hot representations. Additional derived features included route popularity metrics, time-of-day categories (morning, afternoon, evening, night), and carrier-route interaction terms.

**Model Selection and Architecture:** XGBoost (Extreme Gradient Boosting) served as the primary algorithm due to its superior handling of mixed feature types, built-in regularization, and robustness to outliers. The model employed 500 boosting iterations with a learning rate of 0.1, maximum tree depth of 6 to balance complexity and generalization, and L1/L2 regularization to prevent overfitting. Early stopping with a patience of 50 rounds prevented unnecessary computation. Training utilized 80-20 train-test split with random state seeding for reproducibility.

**Evaluation Framework:** Model performance was assessed using $R^2$ coefficient (coefficient of determination), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and visual residual analysis to detect systematic biases across price ranges.

## Results

The XGBoost model achieved exceptional predictive accuracy with an $R^2$ score of 0.948 on the test set, explaining 94.8% of variance in flight prices. Mean Absolute Error reached INR 1,247, representing approximately 8.3% error on the mean ticket price, while RMSE of INR 1,863 indicated well-controlled prediction variance with minimal large errors. Feature importance analysis revealed duration (42% contribution), airline carrier (23%), and number of stops (18%) as dominant pricing factors, while temporal features collectively contributed 12%, confirming seasonal and time-of-day effects. Residual analysis showed homoscedastic error distribution across price ranges with no systematic under- or over-prediction patterns, validating model reliability across economy and premium segments. Cross-validation with 5 folds demonstrated stable performance (mean $R^2 = 0.945$, std = 0.008), confirming robust generalization. The model successfully captured non-linear pricing dynamics, including the non-monotonic relationship between stops and price (direct flights and 2+ stop flights were more expensive than single-stop flights). These results enable practical applications including optimal booking time recommendations, competitive fare monitoring, and revenue management analytics for airlines.