Air ticket Price Prediction

Priyesh Singh, Sagar rai
Department of Computer Applications
Lovely Professional University

Abstract **- Airfare rates are highly volatile and sometimes erratic with respect to many volatile factors such as travel date, airline, stops, and flight duration. This work explores machine learning algorithms to forecast air ticket prices using structured historical flight data. Various models such as Random Forest, Decision Tree, Logistic Regression, and K-Nearest Neighbors (KNN) were experimented with. Among these, Random Forest showed the highest accuracy, suggesting that ensemble methods better identify the nonlinear relationships in airfare rates.**

I. INTRODUCTION

A. Background

Air ticket prices change frequently, influenced by time of booking, demand, airline, and route. These fluctuations confuse travelers and hinder aggregators in providing accurate recommendations. Traditional tools rely on general trends, ignoring specific flight attributes. Machine learning, when applied to structured data, can reveal hidden pricing patterns and make dynamic, data-driven predictions possible [1].

B. Problem Statement

While there are many different travel sites with predictive recommendations, they generally have no flight-specific knowledge. This study questions if machine learning models, having been trained on past flight information, can successfully predict airfares and be helpful to consumers and businesses when planning strategically.

Objectives

Collect and preprocess past flight ticket data, dealing with missing values and encoding categories.

- Engineer appropriate new features such as journey length, stops, and time bins to improve model performance.
- Train and compare the following machine learning models: Random Forest, Decision Tree, K-Nearest Neighbors, and Logistic Regression.
- Test model performance using RMSE, MAE, R^2, and Accuracy measures.
- Examine feature importance to determine major drivers of ticket price fluctuations.
- Suggest future enhancements, such as real-time dynamic data integration and more advanced models such as GRU/LSTM networks.

Prior studies have employed clustering [3] and reinforcement learning [4] for fitness recommendations. This work advances the field by combining diverse data sources and ensemble methods to achieve higher accuracy.

II. METHODOLOGY

2.1 Dataset & Preprocessing

Features in a public domain data include airline, source, destination, stops, duration, and price. Preprocessing involved:

- Null value handling
- Date-time parsing
- Duration conversion to minutes
- One-hot encoding for categorical features

2.2 Feature Engineering

Most important features developed:

- Total travel time
- Departure/arrival time slots (morning, evening, etc.)
- Number of stops
- Class (Economy/Business)

2.3 Algorithms Used

- Random Forest: Ensemble model that creates many trees and combines results.
- Decision Tree: Rule-based, interpretable but vulnerable to overfitting.
- Logistic Regression: Applied in this example to predict price bins (Low, Medium, High).

III. RESULTS

3.1 Metrics Used

RMSE: Penalizes large errors
MAE: Measures average absolute error
R^2: Variance explained by the model
Accuracy: For classification-based comparison

3.2 Performance Summary

Model	RMSE (↓)	MAE (↓)	R^2 (↑)	Accuracy (↓/↑)
Random Forest	Lowest	Low	High	~87%
Decision Tree	Moderate	Medium	Medium	~85%
KNN	High	High	Low	~80%
Logistic Reg.	-	-	-	~84%

IV. Discussion

The confusion matrix in Fig. 1 demonstrates the Random Forest model's ability to correctly classify workouts across all categories, with minimal misclassifications. The diagonal dominance confirms its robustness. Fig. 2 further validates the domain intuition that biometric and effort-based metrics drive personalized recommendations. Limitations include dataset imbalance and reliance on self-reported survey data.

Conclusion and Future Work

This project illustrates that traditional ML models, particularly Random Forest, can successfully forecast air ticket prices. More, however, is needed for actual deployment in real life: incorporating real-time information, seasonality, and UI tools.

Enhancements for the future:
Transformation from classification to continuous regression
Add live flight APIs
Host using Flask/Django with a UI
Implement SHAP for explainability
Investigate deep learning (e.g., GRU, LSTM)

Acknowledgment

The authors thank Dr. Punam Rattan for mentorship and Lovely Professional University for infrastructural support.

References

Bhambri, R. (2020). *Flight Price Prediction Using ML*, Towards DataScience.
Brownlee, J. (2016). *Logistic Regression for ML*.
Scikit-learn Documentation. (2024).
Kaggle Dataset: *Flight Fare Prediction*