FLIGHT FARE PREDICTION USING MACHINE LEARNING

1. Abstract

Flight ticket prices vary significantly based on multiple factors such as airline, travel duration, number of stops, route, and journey timings. Accurate fare prediction can help travelers plan better and assist airline and travel platforms in pricing strategies.
This project focuses on building a **machine learning regression model** to predict flight ticket prices using historical flight data. Multiple regression models were trained and evaluated, and the best-performing model was selected based on performance metrics.

2. Introduction

The airline industry follows dynamic pricing strategies where ticket prices change frequently depending on demand, routes, and operational factors. Predicting flight fares is a challenging problem due to the involvement of both categorical and numerical variables.

This project aims to:

- Analyze flight fare data
- Perform data cleaning and feature engineering
- Build regression models to predict flight prices
- Evaluate and compare different models
- Select the best-performing model

3. Problem Statement

To develop a machine learning model that can accurately predict flight ticket prices based on available flight details such as airline, source, destination, duration, stops, and travel timings.

4. Type of Problem

This is a **Regression Problem** because:

- The target variable **Price** is a **continuous numerical value**

- The goal is to predict an exact numeric fare amount

5. Dataset Description

The dataset contains the following information:

- Airline
- Source
- Destination
- Date of Journey
- Departure Time
- Arrival Time
- Duration
- Total Stops
- Route
- Additional Info
- Price (Target Variable)

6. Data Preprocessing

6.1 Handling Date and Time Features

- Date_of_Journey was converted into:
  - Journey_Day
  - Journey_Month
- Dep_Time and Arrival_Time were processed to extract:
  - Dep_Hour, Dep_Minute
  - Arrival_Hour, Arrival_Minute

6.2 Duration Conversion

- Flight duration strings (e.g., “2h 50m”) were converted into total minutes (Duration_mins).

6.3 Total Stops

- Converted categorical stop values into numeric form:
    - Non-stop → 0
    - 1 stop → 1
    - 2 stops → 2, etc.

6.4 Route Feature

- Route information was transformed into Route_Segments, representing the number of flight legs.

6.5 Handling Missing Values

- Numerical features: Filled using **median imputation**
- Categorical features: Filled using **most frequent value**

Irrelevant columns such as raw time columns and additional information were dropped.

7. Feature Engineering

Final Selected Features

Categorical Features

- Airline
- Source
- Destination

Numerical Features

- Duration_mins
- Total_Stops_num
- Route_Segments
- Journey_Day
- Journey_Month

- Dep_Hour
- Dep_Minute
- Arrival_Hour
- Arrival_Minute

Target Variable

- Price

8. Exploratory Data Analysis (EDA)

- A **correlation heatmap** was used to analyze relationships between numeric features.
- Duration, number of stops, and route segments showed strong correlation with price.
- A **prediction vs actual scatter plot** showed that predicted prices closely followed actual prices, indicating good model performance.

9. Model Building

A machine learning pipeline was created using:

- ColumnTransformer
- OneHotEncoder for categorical variables
- StandardScaler for numerical variables

The following regression models were trained:

1. Linear Regression
2. Decision Tree Regressor
3. Random Forest Regressor
4. XGBoost Regressor

10. Model Evaluation

Evaluation Metrics Used

- **R² Score**
- **MAE (Mean Absolute Error)**
- **RMSE (Root Mean Squared Error)**

Performance Before Hyperparameter Tuning

Model	**R² Score**
Linear Regression	0.62
Decision Tree	0.77
Random Forest	0.82
XGBoost	0.85

11. Hyperparameter Tuning

Hyperparameter tuning was performed using **RandomizedSearchCV**.

Best Parameters

Random Forest

- n_estimators = 200
- max_depth = 12
- min_samples_split = 2

XGBoost

- n_estimators = 200
- max_depth = 7
- learning_rate = 0.1

12. Final Model Performance

Model	**R² Score**
Random Forest (Tuned)	0.842

Model	R² Score
XGBoost (Tuned)	**0.857**

XGBoost was selected as the final model due to its superior predictive performance.

13. Conclusion

This project successfully demonstrates an end-to-end machine learning pipeline for flight fare prediction. Proper data preprocessing, feature engineering, and model tuning significantly improved prediction accuracy. The final XGBoost model achieved an R^2 score of approximately **85.7%**, making it suitable for real-world price estimation tasks.

14. Skills Demonstrated

- Data Cleaning and Feature Engineering
- Exploratory Data Analysis (EDA)
- Machine Learning Pipelines
- Regression Modeling
- Hyperparameter Tuning
- Model Evaluation and Comparison

15. Future Enhancements

- Deploy the model as a web application
- Include real-time flight data
- Perform classification-based price range prediction
- Improve accuracy using advanced ensemble techniques

Final Project Report :

Prepared by:
Arifa
Malathi
Neha