Project Report on

Airline Ticket Price Prediction System

Prepared by

Adit Ghorpade	**Srushti Deshmukh**
612210054	642303007

Department of Computer Engineering,
COEP Technological University (COEP Tech)
(A Unitary Public University of Govt. of Maharashtra)
Shivajinagar, Pune-411005, Maharashtra, INDIA

April 2025

Contents

1 Introduction

The Airline Price Detection System is a machine learning project designed to predict airline ticket prices based on various flight parameters. Airline prices vary due to factors like flight duration, number of stops, airline, and destination. This project analyzes historical flight data and builds a predictive model to estimate ticket prices.

The process starts with data preprocessing, where missing values are handled, categorical variables are encoded, and outliers are detected. Important features influencing the price are extracted and analyzed. Exploratory Data Analysis (EDA) is performed using visualizations to identify trends and patterns.

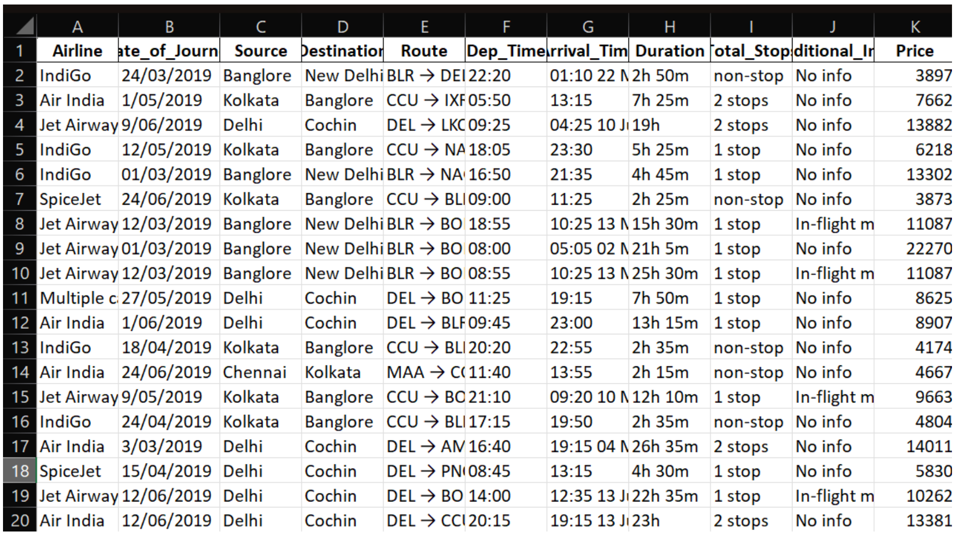
For prediction, the project uses the RandomForestRegressor model, which is trained on processed flight data. The model is then tested on unseen data, and its performance is evaluated using metrics like the R^2 score. The final output is an accurate fare prediction system that helps travelers plan their trips efficiently and allows airlines to optimize their pricing strategies.

2 Dataset Description & Features

2.1 Dataset Description

- **Total Records**: **10,683**
- **Total Columns**: **11**
- **Column Names**:
  1. **Airline** – Name of the airline
  2. **Date_of_Journey** – Date of travel
  3. **Source** – Departure city
  4. **Destination** – Arrival city
  5. **Route** – Flight route taken
  6. **Dep_Time** – Departure time
  7. **Arrival_Time** – Arrival time
  8. **Duration** – Total flight duration
  9. **Total_Stops** – Number of stops between source and destination
  10. **Additional_Info** – Extra flight-related information
  11. **Price – (Target Variable)** Ticket price

2.2 Dataset Preview



	A	B	C	D	E	F	G	H	I	J	K
1	Airline	ate_of_Journ	Source	estinatior	Route	Dep_Time	rrival_Tim	Duration	otal_Stop	ditional_Ir	Price
2	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEI	22:20	01:10 22 M	2h 50m	non-stop	No info	3897
3	Air India	1/05/2019	Kolkata	Banglore	CCU → IXF	05:50	13:15	7h 25m	2 stops	No info	7662
4	Jet Airway	9/06/2019	Delhi	Cochin	DEL → LKC	09:25	04:25 10 Ji	19h	2 stops	No info	13882
5	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NA	18:05	23:30	5h 25m	1 stop	No info	6218
6	IndiGo	01/03/2019	Banglore	New Delhi	BLR → NA	16:50	21:35	4h 45m	1 stop	No info	13302
7	SpiceJet	24/06/2019	Kolkata	Banglore	CCU → BLI	09:00	11:25	2h 25m	non-stop	No info	3873
8	Jet Airway	12/03/2019	Banglore	New Delhi	BLR → BO	18:55	10:25 13 M	15h 30m	1 stop	In-flight m	11087
9	Jet Airway	01/03/2019	Banglore	New Delhi	BLR → BO	08:00	05:05 02 M	21h 5m	1 stop	No info	22270
10	Jet Airway	12/03/2019	Banglore	New Delhi	BLR → BO	08:55	10:25 13 M	25h 30m	1 stop	In-flight m	11087
11	Multiple c	27/05/2019	Delhi	Cochin	DEL → BO	11:25	19:15	7h 50m	1 stop	No info	8625
12	Air India	1/06/2019	Delhi	Cochin	DEL → BLF	09:45	23:00	13h 15m	1 stop	No info	8907
13	IndiGo	18/04/2019	Kolkata	Banglore	CCU → BLI	20:20	22:55	2h 35m	non-stop	No info	4174
14	Air India	24/06/2019	Chennai	Kolkata	MAA → C(	11:40	13:55	2h 15m	non-stop	No info	4667
15	Jet Airway	9/05/2019	Kolkata	Banglore	CCU → BO	21:10	09:20 10 M	12h 10m	1 stop	In-flight m	9663
16	IndiGo	24/04/2019	Kolkata	Banglore	CCU → BLI	17:15	19:50	2h 35m	non-stop	No info	4804
17	Air India	3/03/2019	Delhi	Cochin	DEL → AM	16:40	19:15 04 M	26h 35m	2 stops	No info	14011
18	SpiceJet	15/04/2019	Delhi	Cochin	DEL → PN	08:45	13:15	4h 30m	1 stop	No info	5830
19	Jet Airway	12/06/2019	Delhi	Cochin	DEL → BO	14:00	12:35 13 Ji	22h 35m	1 stop	In-flight m	10262
20	Air India	12/06/2019	Delhi	Cochin	DEL → CCI	20:15	19:15 13 Ji	23h	2 stops	No info	13381

2.3 Dataset Features

1. Input Features (Independent Variables):

   1. Airline (Categorical) – Name of the airline operating the flight.
   2. Date_of_Journey (Date/Time) – Date on which the journey is scheduled.
   3. Source (Categorical) – Departure city of the flight.
   4. Destination (Categorical) – Arrival city of the flight.
   5. Route (Categorical) – The specific route taken by the flight (e.g., direct or via stops).
   6. Dep_Time (Date/Time) – Flight departure time.
   7. Arrival_Time (Date/Time) – Flight arrival time.
   8. Duration (Categorical/Numerical) – Total flight duration in hours and minutes.
   9. Total_Stops (Categorical/Numerical) – Number of stops between source and destination.
   10. Additional_Info (Categorical) – Extra information about the flight, such as meal availability or layovers.

2. Target Variable (Dependent Variable):

   11. Price (Numerical) – Target variable, representing the airline ticket price (in INR).

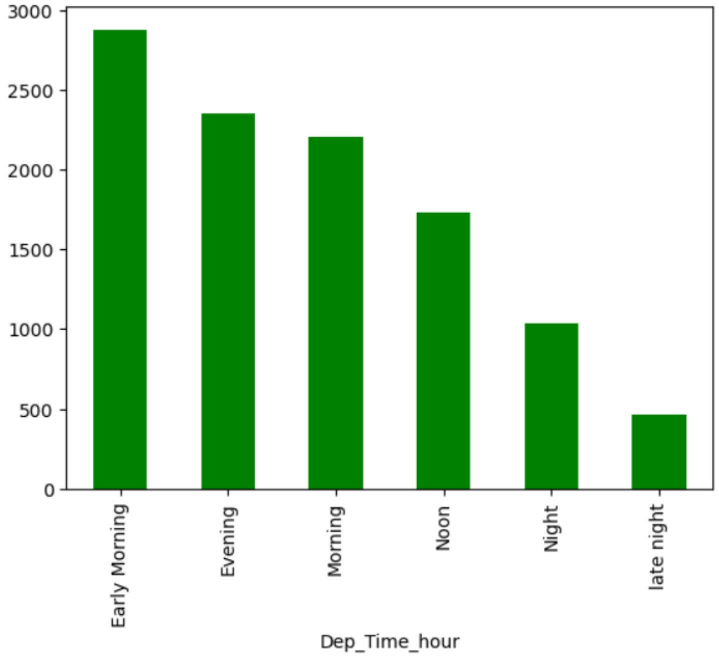
Feature Types Breakdown:

- Categorical Features:
  - Airline, Source, Destination, Route, Total_Stops, Additional_Info.
- Numerical Features:
  - Price (Target), Duration (converted to minutes), Dep_Time, Arrival_Time.
- Date/Time Features:
  - Date_of_Journey, Dep_Time, Arrival_Time

3 Plotting

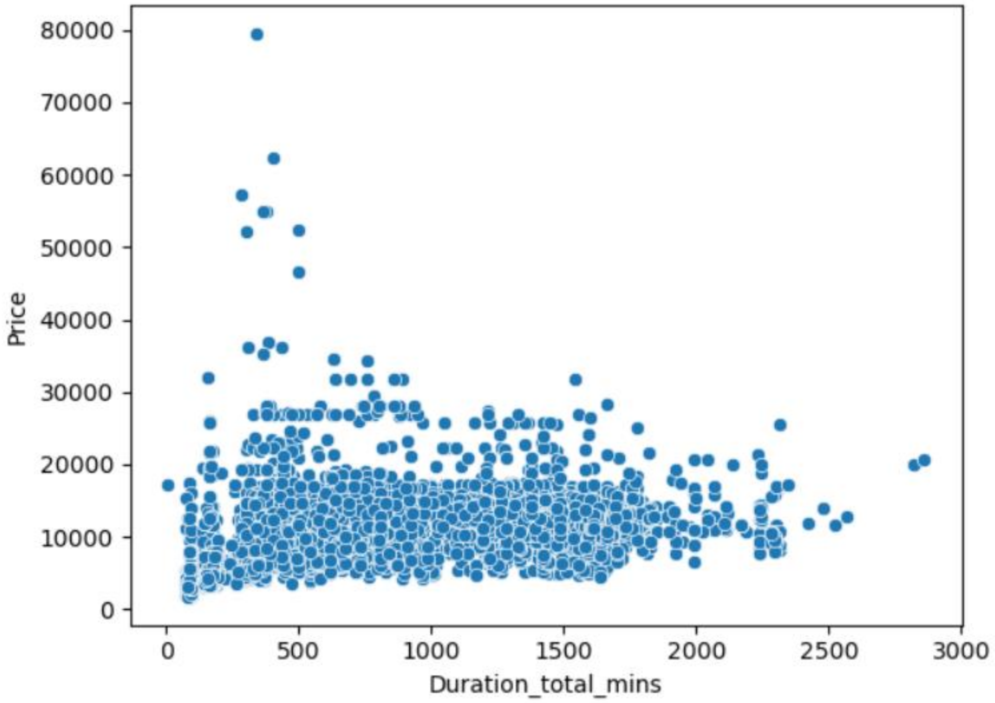
Price vs Dep_Time_hour

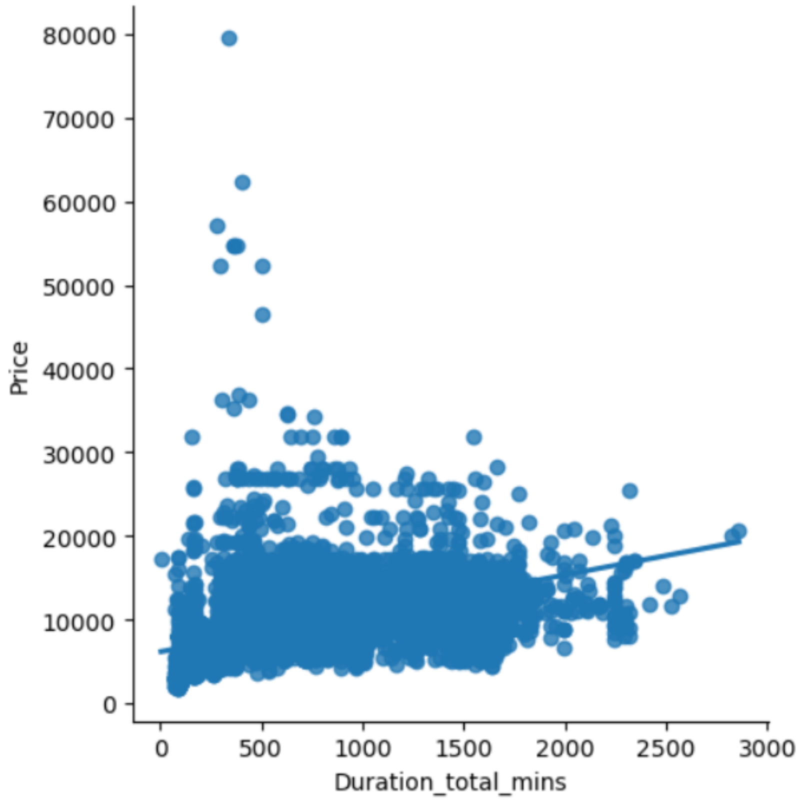
```
<Axes: xlabel='Dep_Time_hour'>
```



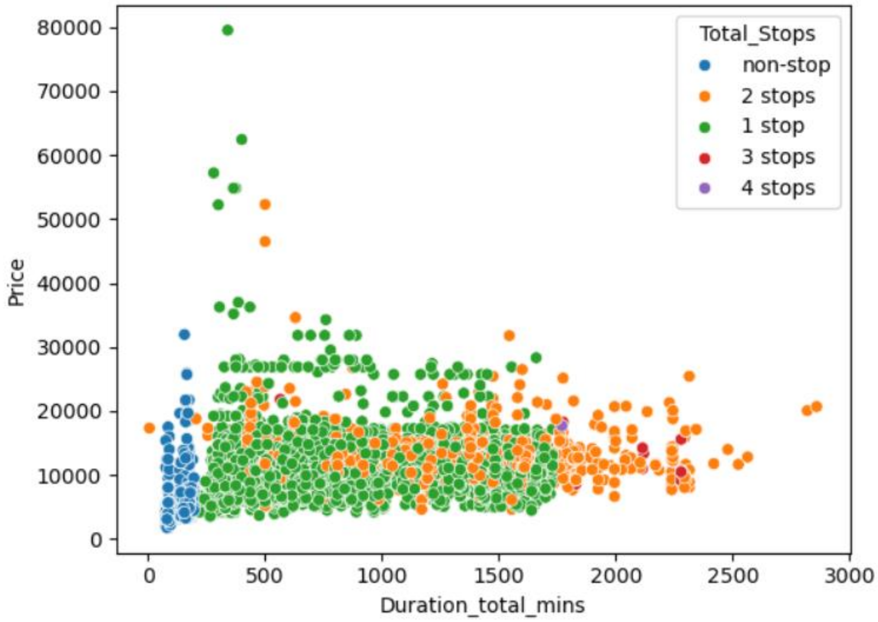
Duartion_total_mins vs Price Dependency

```
<Axes: xlabel='Duration_total_mins', ylabel='Price'>
```

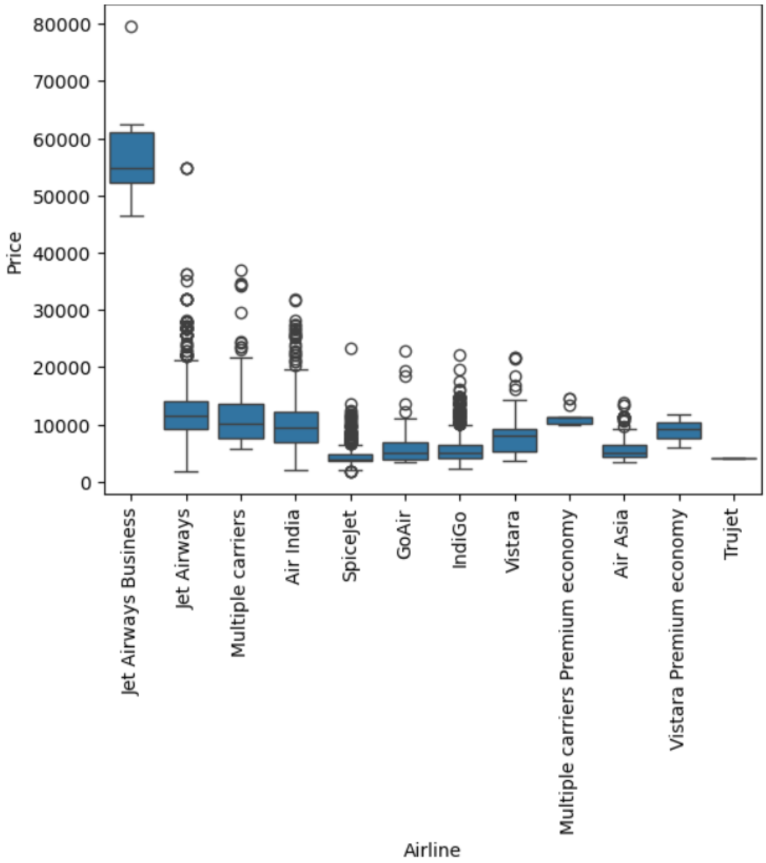




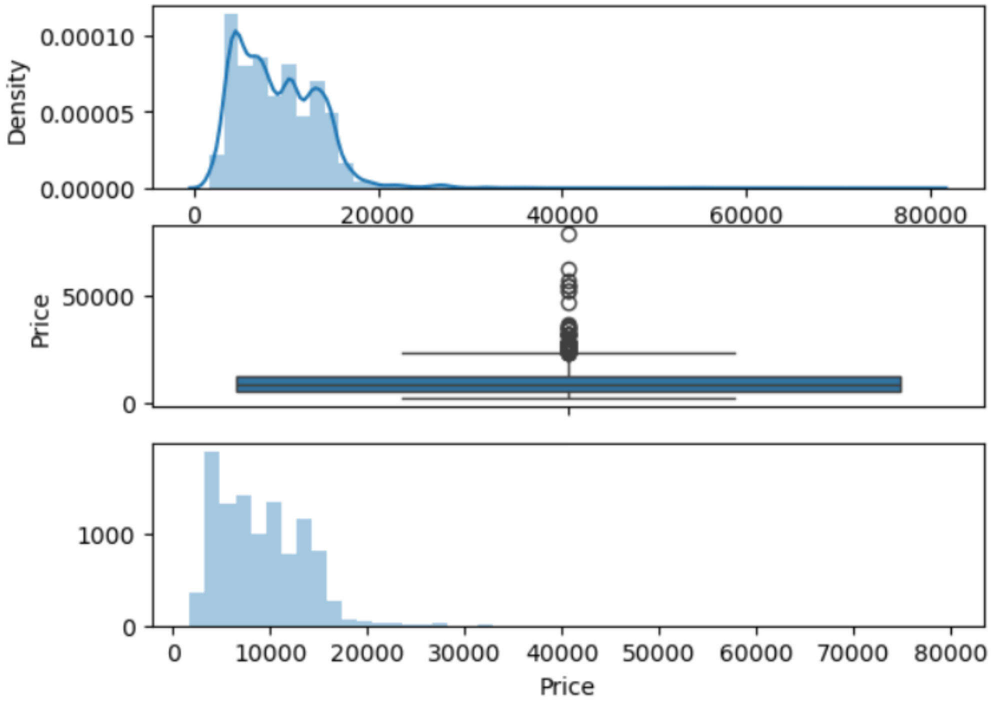
Duartion_total_mins vs Price



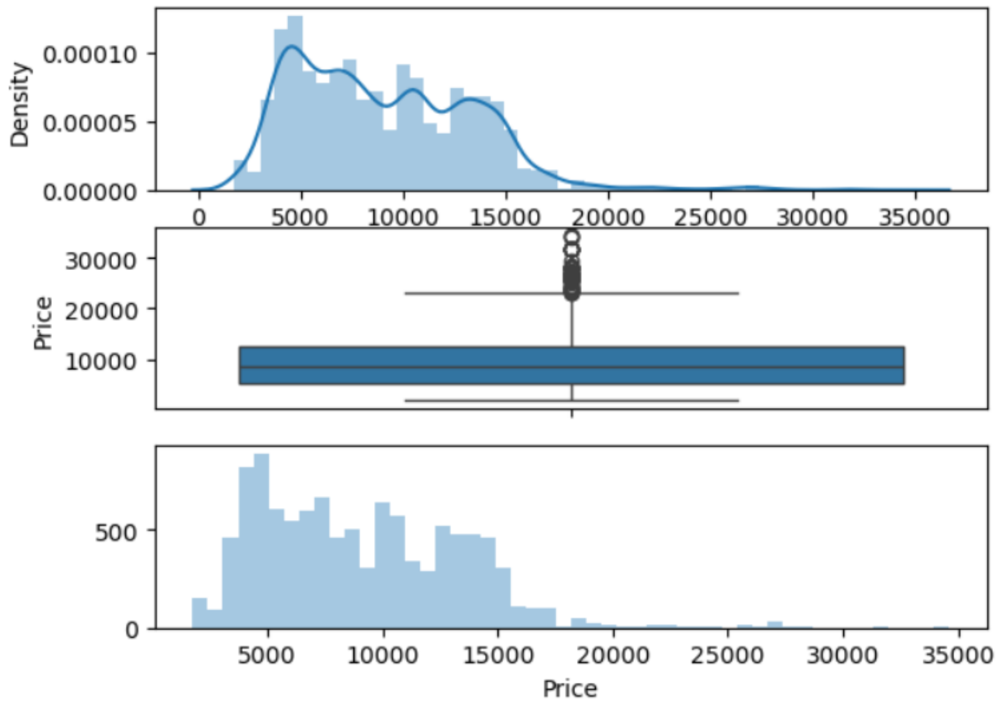
Airlines vs Price

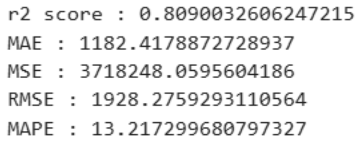


Outliers in Price

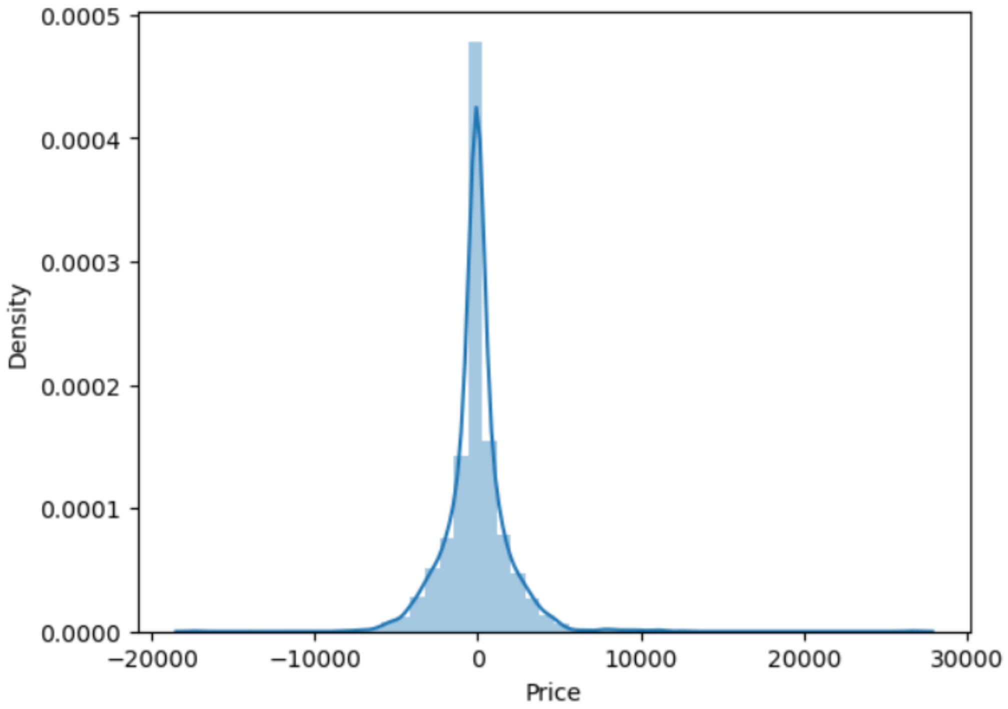


After Removing outliers





```
r2 score : 0.8090032606247215
MAE : 1182.4178872728937
MSE : 3718248.0595604186
RMSE : 1928.2759293110564
MAPE : 13.217299680797327
```



4 Preprocessing Techniques Used

Data preprocessing is a crucial step in building a machine learning model, ensuring that the dataset is clean, structured, and ready for analysis. The following techniques were applied to process and transform the dataset before model training:

4.1 Handling Missing Values

Missing values in the dataset can lead to inaccurate predictions and biases in the model. We handled missing values by analyzing their distribution and applying suitable imputation techniques to maintain data consistency. Any records with critical missing values were either filled using appropriate statistical measures (mean/median/mode) or dropped if they contributed to data inconsistencies.

1. **Dashboard Module:** Admin and Student views
2. **Feedback Module:** Feedback collection and reporting

4.2 Extratcing Useful Features

Feature engineering was performed to derive meaningful insights from existing columns:

- Departure Time (Dep_Time) and Arrival Time (Arrival_Time): Extracted hour and minute components separately to capture the effect of different flight times on pricing trends.
- Journey Date (Date_of_Journey): Extracted day and month from the journey date to analyze the impact of seasonal variations on ticket prices.
- Duration: Converted the duration column into numeric format to quantify its effect on pricing.

4.3 Encoding Methods

Categorical data needs to be converted into numerical format for machine learning models. The following encoding techniques were used:

- **One-Hot Encoding**: Applied to Source (departure city), where each category was transformed into binary variables. This method prevents ordinal relationships from being misinterpreted by the model.
- **Label Encoding**: Applied to Number of Stops, where stop values were assigned numerical labels (e.g., 0 for non-stop, 1 for one-stop, etc.). This helps the model recognize stop patterns without increasing dimensionality.
- **Target Mean Encoding**: Applied to Airline and Destination, where each category was replaced by the mean ticket price for that category. This technique helps capture the actual influence of categorical variables on the target price while avoiding excessive feature expansion.

4.4 Outlier Detection

Outliers can distort model training and affect prediction accuracy. We used box plots and scatter plots to visualize the distribution of ticket prices and identify outliers. Flights with unrealistically high or low prices were carefully examined and either transformed or removed to ensure a balanced dataset.

4.5 Data Visualization For Feature Analysis

To understand which factors influence airline ticket prices the most, we plotted:

- Box Plots & Scatter Plots to analyze relationships between flight features and ticket prices.
- Visual comparisons of flight duration, stops, airline types, and destinations to observe trends and variations in pricing.

.

5 Algorithms/Models Used

5.1 Random Forest Regressor

The RandomForestRegressor is a machine learning model that belongs to the ensemble learning category. It is a part of the **scikit-learn** (sklearn) library, which provides a wide range of tools for machine learning and data analysis. Random Forest is an extension of Decision Trees that builds multiple trees and combines their outputs to improve accuracy and reduce overfitting.

For our airline price prediction problem, we need a regression model that can handle both numerical and categorical data, capture complex relationships between features, and provide accurate predictions. RandomForestRegressor is chosen because:

- It handles non-linearity well, making it suitable for price prediction where relationships between features and prices are not strictly linear.
- It reduces overfitting by averaging multiple decision trees, ensuring that the model generalizes well to unseen data.
- It works effectively with categorical and numerical data without requiring extensive transformations.
- It provides feature importance scores, helping us understand which features contribute the most to price variations.

5.2 Working Of the Model

1. **Dataset Splitting:**
   The dataset is split into training (X_train, y_train) and testing (X_test, y_test) sets using train_test_split() from sklearn.
   Typically, 75% of the data is used for training and 25% for testing to evaluate model performance.
2. **Model Training**:
   Multiple decision trees are trained on different subsets of the dataset.
   Each tree learns patterns from random samples of data and predicts ticket prices.
3. **Prediction and Averaging**:
   Each decision tree gives an independent prediction.
   The final prediction is obtained by averaging the outputs of all trees, reducing variance and improving accuracy.
4. **Model Evaluation**:
   The trained model is tested on X_test, and predicted prices (y_pred) are compared with actual prices (y_test).
   Performance is measured using the R² score, which indicates how well the model explains the variance in ticket prices.

5.3 Random Forest Over Other Regression Models

Better than Linear Regression: Linear regression assumes a strict linear relationship between input features and the target variable, which does not hold in real-world pricing models.

Better than Decision Trees: A single decision tree may overfit the data, leading to poor generalization. Random Forest, by combining multiple trees, reduces variance and improves robustness.

More accurate than basic regression models: Traditional regression models may not capture complex interactions between features like flight duration, stops, and airline types, but Random Forest can efficiently learn these patterns

6 Comparative Measures used & Metrics Used for different Methods

To assess the accuracy and reliability of our Airline Price Prediction Model, we use various performance metrics. These metrics help us understand how well the model predicts airline prices and whether it generalizes well to unseen data.

6.1 R^2 Score

The R² score measures how well the model explains the variance in the target variable (ticket prices). It ranges from 0 to 1, where:

1 means perfect predictions (all variance is explained).

0 means the model does not explain any variance in the target.

Negative values indicate that the model performs worse than a simple average prediction.

Model's R² Score:

- Training R² Score: **0.9513** → The model explains 95.13% of the variance in training data.
- Test R² Score: **0.8090** → The model explains 80.90% of the variance in test data.

Interpretation:

- ➢ The model performs very well on training data, indicating it has learned patterns effectively.
- ➢ The drop in test R² (from 0.95 to 0.81) suggests slight overfitting, meaning the model fits training data better than unseen test data.
- ➢ However, 80.9% variance explanation is still a strong performance for a real-world pricing prediction model.

6.2 Mean Absolute Error(MAE)

MAE calculates the average absolute difference between actual and predicted values.

Model's MAE: **1182.41**

On average, our model's ticket price predictions are off by Rs.1182.41 from actual values.

Lower MAE values indicate better accuracy.

6.3 Mean Squared Error(MSE)

MSE measures the average squared difference between actual and predicted values.

Model's MSE: **3718248.06**

MSE penalizes larger errors more heavily than MAE.

A high MSE suggests some large prediction errors exist, but squaring amplifies them, making the interpretation tricky.

6.4 Root Mean Squared Error (RMSE)

RMSE is the square root of MSE and provides an error value in the same unit as the target variable.

Model’s RMSE: **1928.27**

This means the model’s predictions deviate from actual prices by approximately ₹1928.27 on average. RMSE is useful for understanding the model’s typical prediction error. A lower RMSE indicates a more reliable model

6.5 Mean Absolute Percentage Error (RMSE)

MAPE calculates the percentage error between actual and predicted values, making it easier to compare models across datasets.

Our Model’s MAPE: 13.21%

This means the model makes an average 13.21% error in price prediction.

A lower MAPE is preferred; anything below 10-15% is considered a good predictive model.

Train & Test Performance

Metric	Training Score	Test Score
R^2Score	0.9513	0.8090
MAE	-	1182.41
MSE	-	3718248.06
RMSE	-	1928.27
MAPE	-	13.21%

Final Interpretation

- Random Forest Regressor model performs well, explaining 80.9% of ticket price variance in test data.
- The low MAE, RMSE, and MAPE values indicate that predictions are reasonably accurate.
- There is some overfitting, as the training score is much higher than the test score, which could be improved using techniques like hyperparameter tuning or reducing tree depth in RandomForest.

7 Conclusion

The Airline Price Detection System successfully predicts airline ticket prices by analyzing key flight parameters using machine learning techniques. The project involved data preprocessing, feature extraction, exploratory data analysis, model training, and evaluation to develop an accurate price prediction model.

By utilizing the RandomForestRegressor, the system effectively learns from historical flight data and provides reliable price estimations. The model’s performance was assessed using the R^2 score, ensuring accurate predictions. Factors such as flight duration, number of stops, airline, and destination were identified as major contributors to price variations.