



# Indian Institute of Technology Jodhpur

Introduction To Data Science (MAL7011), October 2024

Report On

---

## Flight Fare Prediction System

---

*Group Member 1:* Vimal Kumar Verma (M24MAC015)  
*Group Member 2:* Amit Kumar Verma (M24MAC016)  
*Group Member 3:* Shashank Mishra (M24MAC011)

## Abstract

Flight ticket fare is one of the most fluctuating datasets, varying frequently based on numerous factors such as demand, seasonality, availability, and external conditions like fuel prices or economic events. These fluctuations make it challenging to predict flight ticket fares accurately, as prices can change daily, weekly, or even within hours. However, with the right approach, it is possible to predict flight ticket fares with near accuracy, helping travelers and airlines make informed decisions.

The primary objective of our project, *Flight Fare Prediction System*, is to provide accurate and reliable predictions of future flight ticket prices. Our system leverages supervised machine learning algorithms, enabling it to learn from historical fare data and identify patterns and trends. By considering variables such as the time of booking, airline, destination, and flight duration, we aim to forecast future ticket prices more effectively. The model's performance will be enhanced through feature engineering and model optimization, leading to more precise predictions.

This system has the potential to benefit not only travelers by allowing them to book at the optimal time but also airlines by enabling better pricing strategies. Through this predictive model, we aim to bring more transparency and confidence to the flight booking process, ultimately improving user experience and decision-making.

# Contents

<b>1</b>	<b>Introduction</b>	<b>5</b>
<b>2</b>	<b>Project Goal</b>	<b>6</b>
<b>3</b>	<b>Approach</b>	<b>6</b>
<b>4</b>	<b>Objective</b>	<b>6</b>
<b>5</b>	<b>About the Project</b>	<b>6</b>
<b>6</b>	<b>Problem Validation &amp; Market Research</b>	<b>7</b>
<b>7</b>	<b>Scope</b>	<b>7</b>
<b>8</b>	<b>Technical Aspects</b>	<b>7</b>
<b>9</b>	<b>Dataset</b>	<b>9</b>
9.1	Features . . . . .	9
<b>10</b>	<b>Data Preprocessing and Feature Engineering</b>	<b>9</b>
<b>11</b>	<b>Exploratory Data Analysis (EDA)</b>	<b>10</b>
11.1	Visualizations . . . . .	10
11.2	Airline vs. Price . . . . .	11
11.3	Total Stops vs. Price . . . . .	11
11.4	Source and Destination Analysis: . . . . .	12
11.5	Further Exploratory Analysis . . . . .	13
<b>12</b>	<b>Statistical Analysis</b>	<b>15</b>
12.1	Kurtosis . . . . .	15
12.2	Skewness . . . . .	16
12.3	Descriptive Statistics . . . . .	17
12.4	Correlation Matrix . . . . .	17
12.5	Outlier Detection . . . . .	17
<b>13</b>	<b>Data Transformation</b>	<b>18</b>
13.1	Box-Cox Transformation . . . . .	18
13.2	Outlier Handling . . . . .	18
<b>14</b>	<b>Model Building</b>	<b>21</b>
14.1	Train-Test Split . . . . .	21
14.2	Feature Engineering . . . . .	21
14.3	Model Selection . . . . .	22
14.4	Best Model Selection . . . . .	22
<b>15</b>	<b>Hyperparameter Optimization</b>	<b>23</b>
15.1	Optuna Overview . . . . .	23

<b>16 Model Evaluation</b>	<b>23</b>
16.1 Comparison . . . . .	24
<b>17 Prediction of Flight Prices for a New Data Point</b>	<b>25</b>
<b>18 Results</b>	<b>26</b>
<b>19 Conclusion</b>	<b>28</b>
<b>20 Google Colab Link</b>	<b>31</b>

# 1 Introduction

Nowadays, airline corporations employ highly complex and dynamic strategies for flight ticket fare calculations. These strategies are influenced by a wide range of variables, including market demand, seat availability, booking time, competition among airlines, fuel prices, seasonal factors, and even the day of the week. The dynamic pricing models used by airlines are designed to maximize revenue by adjusting fares in real-time, based on both predicted and observed trends.

This makes it increasingly difficult for customers to anticipate flight prices, as the fares may fluctuate multiple times within a day. Additionally, airlines use yield management techniques, where pricing is adapted based on how many seats are left or how close the flight is to departure. These complexities leave customers often unsure of the best time to book, which can lead to either overpaying or missing out on lower fares.

Understanding these price variations and the factors behind them has become essential for travelers who wish to optimize their flight bookings. As a result, there is a growing demand for systems that can predict flight ticket prices more accurately. By leveraging advanced machine learning algorithms and analyzing large datasets, it is possible to build predictive models that can help users navigate this complexity and book their flights at the best possible rates.

## 2 Project Goal

Our project, Flight Price Prediction System, addresses this issue by providing a solution that enables users to predict flight ticket prices before making a purchase. By analyzing historical data and market trends, our system offers accurate fare predictions, helping travelers make informed decisions and book at the best possible time and price.

## 3 Approach

Our proposed method utilizes machine learning algorithms, specifically focusing on supervised learning techniques. We are sourcing our dataset from a website that provides details of short-duration Indian flights. The project involves feature engineering, where we preprocess the raw data and transform it into a structured dataframe. Once processed, we proceed with normalizing the dataframe.

The regression model selected for prediction is Extreme Gradient Boosting (XGBoost). We train the model using the normalized dataframe, and after experimenting with and fine-tuning the hyperparameters, we achieve predicted results and assess the model's accuracy.

## 4 Objective

The main objective of this project is to apply machine learning techniques to model and predict flight ticket prices over time. By leveraging historical data, the goal is to understand how flight prices fluctuate and forecast future prices with accuracy. This project aims to analyze the patterns and trends in airfare changes, identify key factors influencing these variations (such as demand, seasonality, and competition), and uncover the underlying pricing models used by airlines. Ultimately, the system will help users make better booking decisions by predicting the optimal time to purchase tickets.

## 5 About the Project

The initiative revolves around the prediction of flight ticket prices, which is a difficult task owing to the multitude of factors that influence pricing. To overcome this challenge, we have devised a model that can predict the prices of future flight tickets. Our project employs machine learning methodologies, specifically utilizing the Extreme Gradient Boosting (XGBoost) algorithm to ensure precise predictions. We are also developing an accessible web interface where users can input their flight parameters—such as departure location, destination, date of travel, and the airline for which they seek a price forecast. Upon submission, the system will analyze the provided data and yield a predicted ticket price generated by our model. This forecast enables users to plan their bookings more strategically, empowering them to make well-informed decisions based on the results of our algorithm. The interface is crafted for user convenience, providing rapid and reliable predictions, thus streamlining the entire process for users.

## 6 Problem Validation & Market Research

India’s civil aviation industry is experiencing rapid growth. According to reports, India aims to become the third-largest aviation market by 2020 and the largest by 2030. By FY2017, Indian domestic air traffic was projected to surpass 100 million passengers, up from 81 million in 2015, as noted by the Centre for Asia Pacific Aviation (CAPA). This growth highlights the increasing demand for air travel in the country.

Google Trends data shows that the term “Cheap Air Tickets” is frequently searched in India, reflecting the heightened interest in finding affordable flight options. As India’s middle class expands and more people gain access to air travel, the demand for cost-effective solutions rises significantly. This growing interest in budget-friendly airfares presents a valuable opportunity for a predictive system that helps users book flights at optimal prices, making travel more accessible for the masses. By addressing this demand, our project aims to assist consumers in navigating fluctuating airfares and finding the best deals.

## 7 Scope

Traditionally, when purchasing an airplane ticket, the common strategy is to buy well in advance of the flight’s departure date to avoid the risk of rapidly increasing prices. However, this approach is not always foolproof, as airlines sometimes lower ticket prices closer to departure to boost sales when demand is low or seats remain unsold. This makes predicting the best time to buy challenging for consumers.

Airlines use a wide range of variables to determine flight ticket prices, such as whether the flight is during a peak holiday season, the number of available seats, and the month of travel. While some factors, like seasonality or seat availability, are observable, many other variables remain hidden, contributing to the complexity of price fluctuations.

In this dynamic environment, buyers are constantly searching for the optimal time to purchase a ticket, aiming to secure the lowest possible price. Conversely, airlines are focused on maximizing overall revenue by adjusting fares based on demand and market conditions. Since airlines have the flexibility to change prices at any moment, travelers can save a significant amount of money if they manage to book their tickets when prices are at their lowest.

This uncertainty creates an opportunity for predictive models like ours to help travelers make better purchasing decisions by forecasting future price trends and identifying when ticket prices are likely to drop. By understanding and anticipating these fluctuations, our project can aid both travelers in saving money and airlines in managing their pricing strategies more effectively.

## 8 Technical Aspects

This project is highly focused on machine learning and statistics. We utilized Python and scikit-learn for model implementation and automation. These tools allowed us to efficiently process data, build models, and predict flight prices.

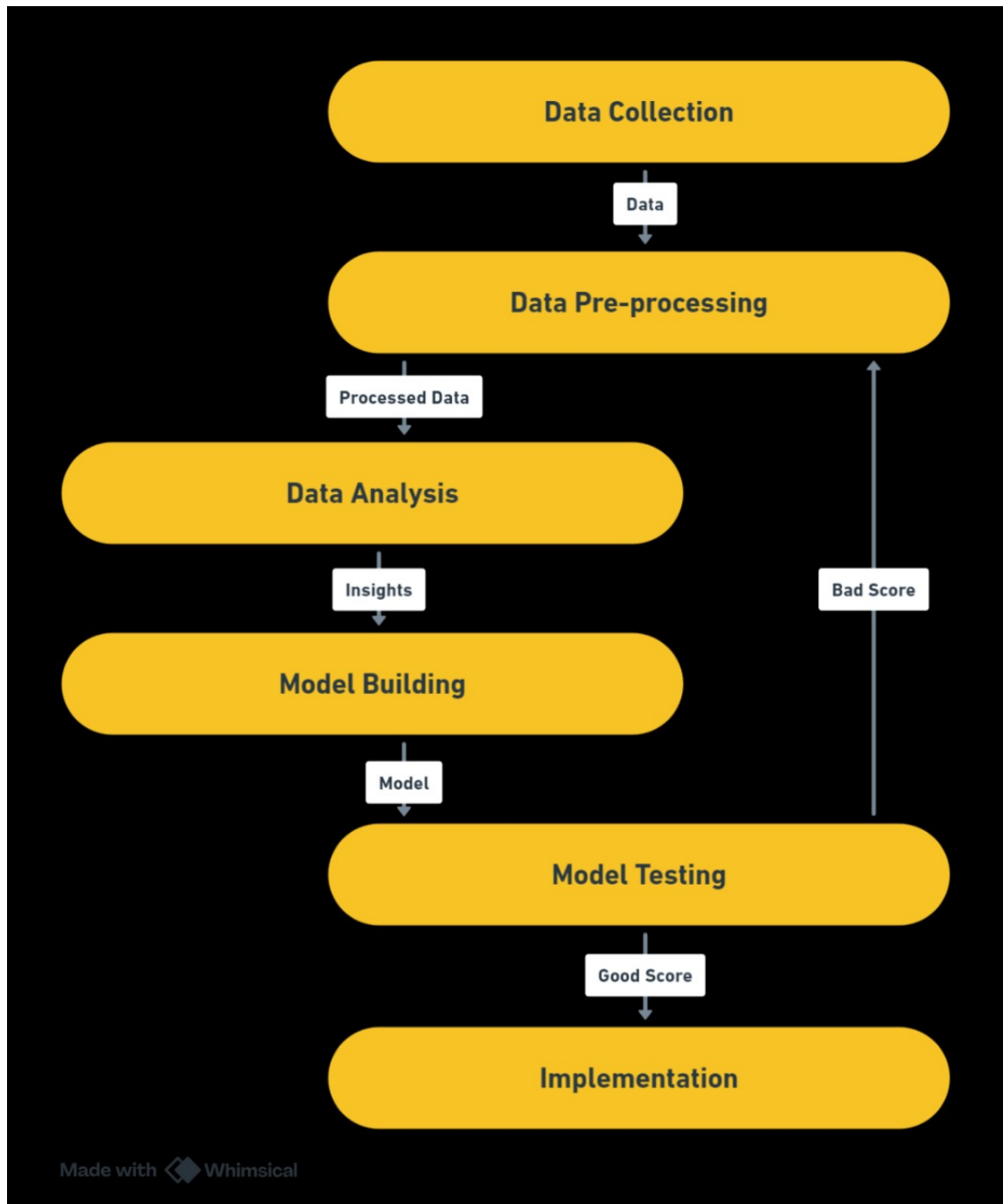


Figure 1: Flow diagram



## 9 Dataset

The dataset contains information about flight booking options from the website Ease-mytrip for flight travel between India's top 6 metro cities. There are 300,261 data points and 9 features in the cleaned dataset.

### 9.1 Features

The various features of the cleaned dataset are explained below:

1. **Airline:** The name of the airline company is stored in the airline column. It is a categorical feature having 6 different airlines.
2. **Flight:** Flight stores information regarding the plane's flight code. It is a categorical feature.
3. **Source City:** City from which the flight takes off. It is a categorical feature having 6 unique cities.
4. **Departure Time:** This is a derived categorical feature created by grouping time periods into bins. It stores information about the departure time and has 6 unique time labels.
5. **Stops:** A categorical feature with 3 distinct values that stores the number of stops between the source and destination cities.
6. **Arrival Time:** This is a derived categorical feature created by grouping time intervals into bins. It has six distinct time labels and keeps information about the arrival time.
7. **Destination City:** City where the flight will land. It is a categorical feature having 6 unique cities.
8. **Additional Info:** A categorical feature that contains information on seat class; it has many distinct values: Business, Economy, etc.
9. **Duration:** A continuous feature that displays the overall amount of time it takes to travel between cities in hours.
10. **Price:** The target variable that stores information about the ticket price.

## 10 Data Preprocessing and Feature Engineering

The following preprocessing steps were carried out:

- **Arrival Time Extraction:** Split "Arrival.Time" into "Arrival\_hr" and "Arrival\_minz".

```
dataset['day'] = pd.to_datetime(dataset['Date_of_Journey']).dt.day
dataset['month'] = pd.to_datetime(dataset['Date_of_Journey']).dt.month
dataset['year'] = pd.to_datetime(dataset['Date_of_Journey']).dt.year
```

- **Departure Time Preprocessing:** Extracted departure hour and minutes from "Dep\_Time".

```
dataset['Dep_hr'] = dataset['Dep_Time'].str.split(':', expand=True)[0].astype(float)
dataset['Dep_Minz'] = dataset['Dep_Time'].str.split(':', expand=True)[1].astype(float)
dataset.drop(columns={'Dep_Time'}, inplace=True)
```

- **Total Stops Handling:** Converted "Total\_Stops" from categorical to numeric format.

```
dataset['Total_Stops'] = dataset['Total_Stops'].str.replace('non-stop', '0')
dataset['Total_Stops'] = dataset['Total_Stops'].str.split(" ", expand=True)[0]
dataset['Total_Stops'] = dataset['Total_Stops'].astype(float)
```

## 11 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to derive meaningful insights from the dataset and uncover underlying patterns. The following key analyses were performed:

### 11.1 Visualizations

Various visualizations were created to explore the relationships between variables. We utilized:

- **Histograms and Box Plots:** These were employed to analyze the distribution of flight prices and other numerical features, helping to identify outliers and understand the central tendencies of the data.
- **Scatter Plots:** These visualizations were used to examine correlations between different features, such as departure times and flight duration, providing insights into how these factors interact with ticket prices.
- **Airline Analysis:** We explored the variation in average ticket prices across different airlines. This analysis highlighted pricing strategies employed by each carrier and helped identify the airlines that typically offer competitive rates.
- **Total Stops Impact:** The influence of the number of stops on average prices was analyzed. We found that non-stop flights tend to have higher average fares compared to flights with one or more stops, emphasizing the preference for direct travel among consumers.
- **Source and Destination Analysis:** We identified the most popular source and destination cities, providing a better understanding of travel patterns and demand within the Indian aviation market. This analysis can inform airlines about potential opportunities for route expansion or promotional offers.

- **Time Series Analysis:** We conducted a time series analysis to observe how flight prices change over different days and months. This allowed us to identify seasonal trends and patterns, helping travelers anticipate price fluctuations.
- **Price Distribution Analysis:** We analyzed the price distribution across various classes (Economy vs. Business) and times of travel. This revealed how class type influences pricing and the availability of discounted fares.
- **Impact of Day of the Week:** The analysis also considered how prices vary based on the day of the week, further refining our understanding of consumer behavior and pricing dynamics in the airline industry.

Exploratory analysis was conducted to understand the data:

## 11.2 Airline vs. Price

Observed how average flight price varies with different airlines.

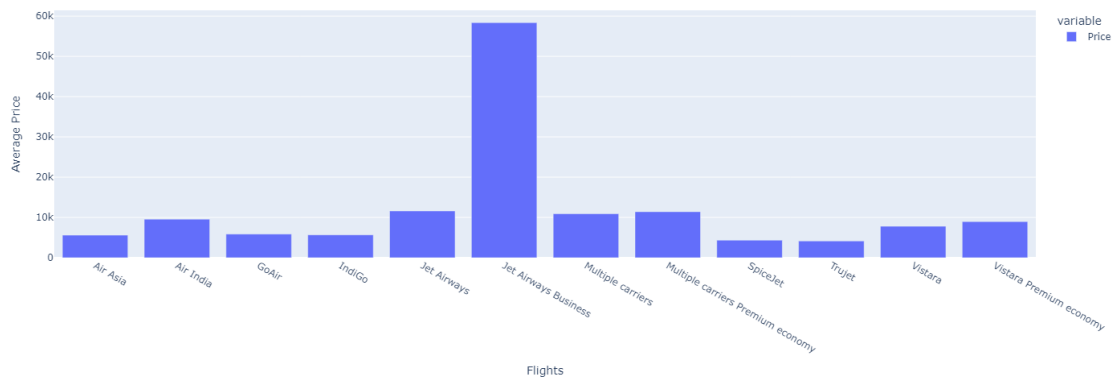


Figure 2: Your caption here

Regular flights usually have lower mean price while premium flights for business classes have high mean price.

## 11.3 Total Stops vs. Price

Analyzed the impact of the number of stops on average flight prices.

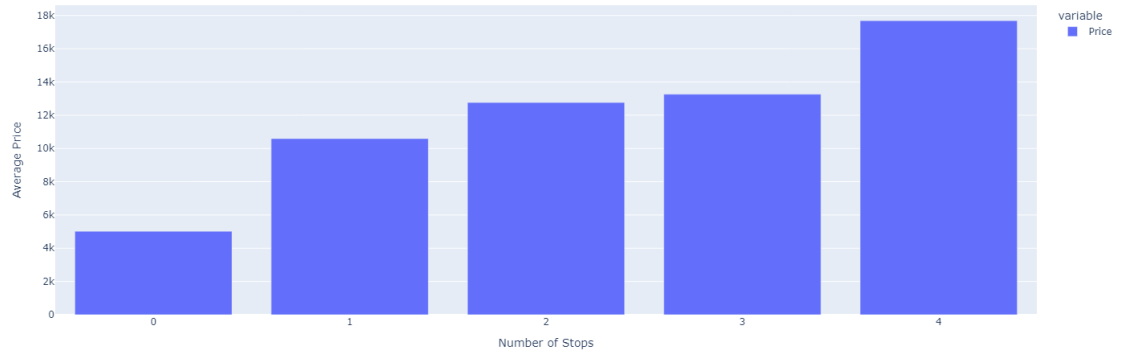


Figure 3: Your caption here

Higher stops have a longer flight duration and hence a higher price.

#### 11.4 Source and Destination Analysis:

Identified the most popular source and destination cities.:

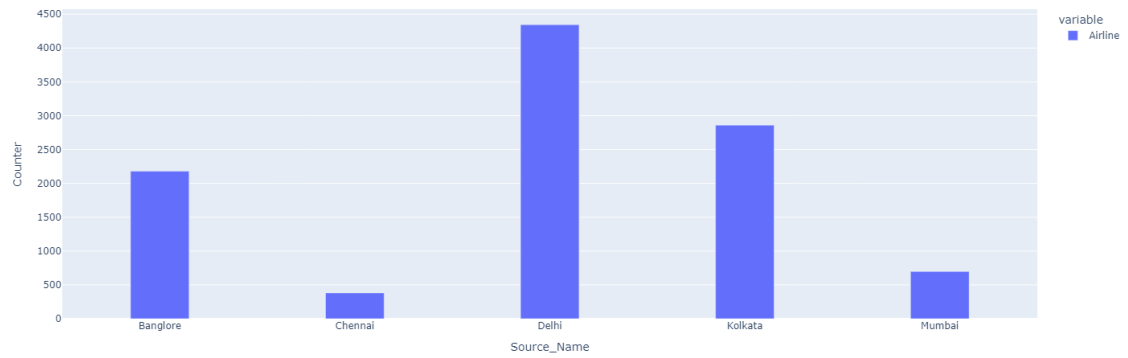


Figure 4: Most Popular Source

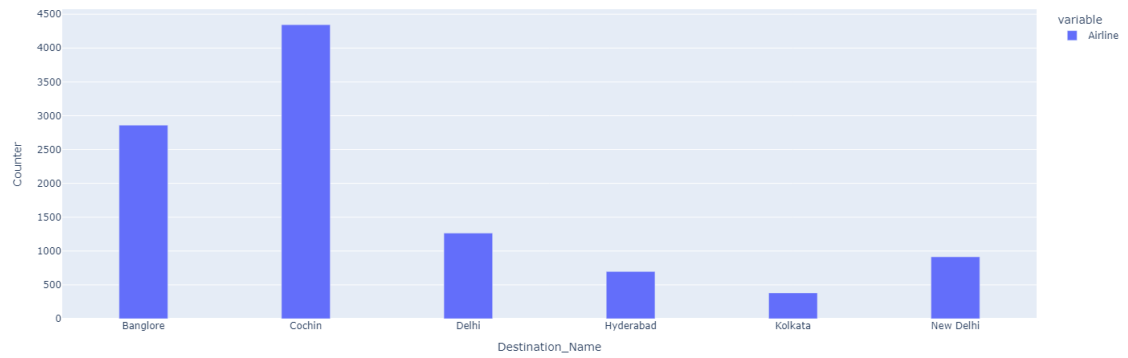


Figure 5: Most Popular Destination

## 11.5 Further Exploratory Analysis

- Do ticket prices change depending on departure and arrival times?

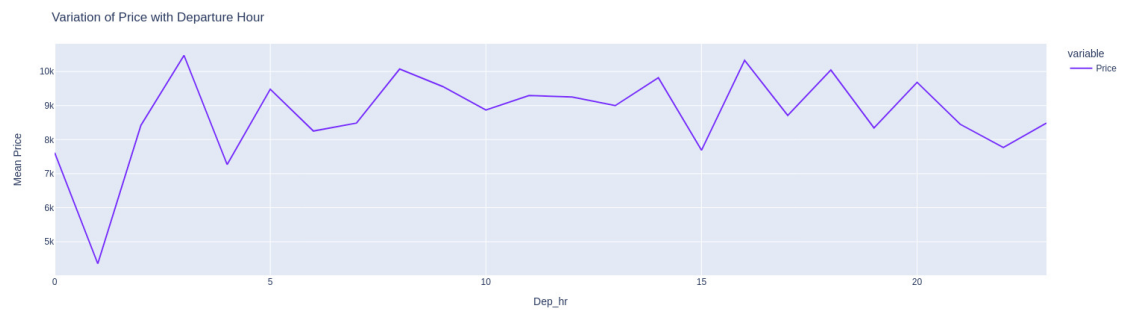


Figure 6: Variation of price on departure

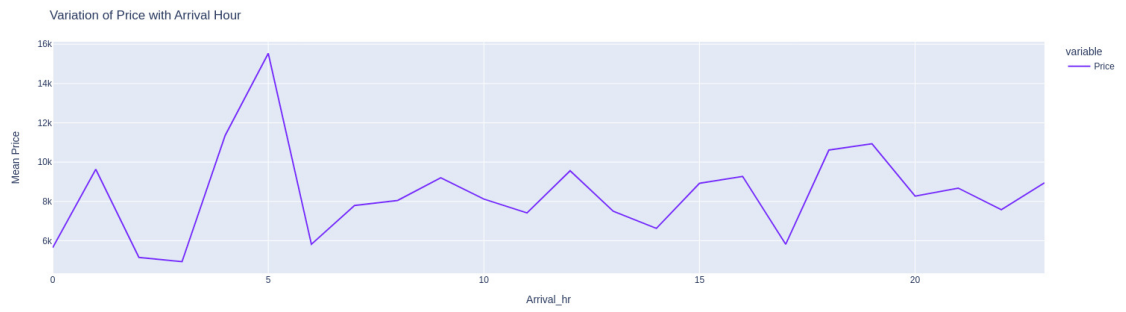


Figure 7: Variation of price on arrival

- How do price changes occur with different origins and destinations?

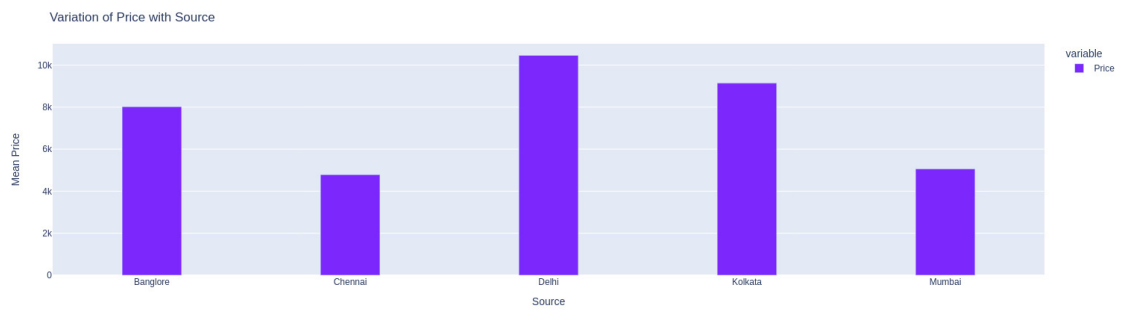


Figure 8: Variation of price on Origin

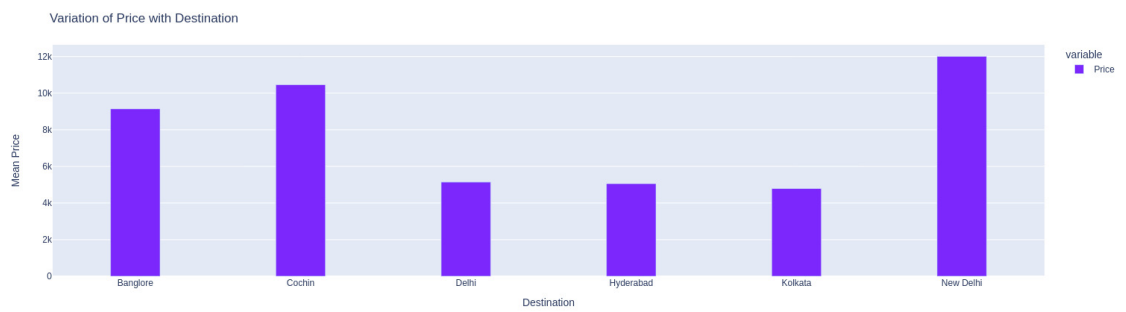


Figure 9: Variation of price on destination

- What is the price difference between Economy and Business class tickets?

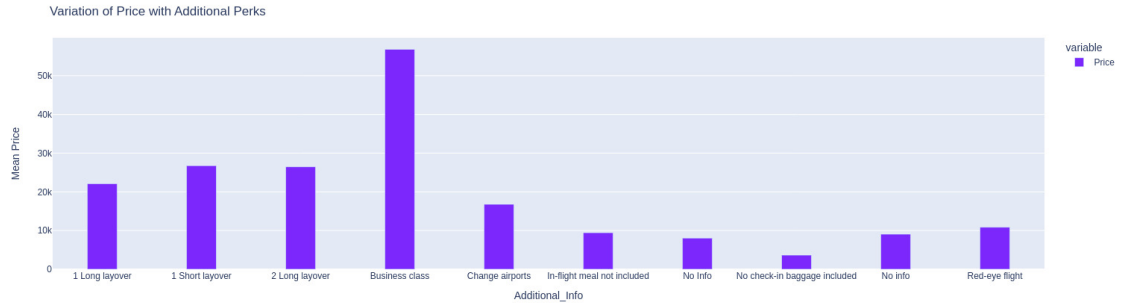


Figure 10: Variation of price on Classes

## 12 Statistical Analysis

Statistical analysis was performed on the dataset to further understand the characteristics of numerical features and their distributions. The following analyses were conducted:

### 12.1 Kurtosis

We plotted the distribution of numerical features and calculated their kurtosis values to identify heavy-tailed distributions. High kurtosis values suggest a distribution with heavy tails and a sharper peak, indicating a higher likelihood of extreme values (outliers). This analysis is crucial for understanding the risk associated with flight pricing and the potential for unexpected fare changes.

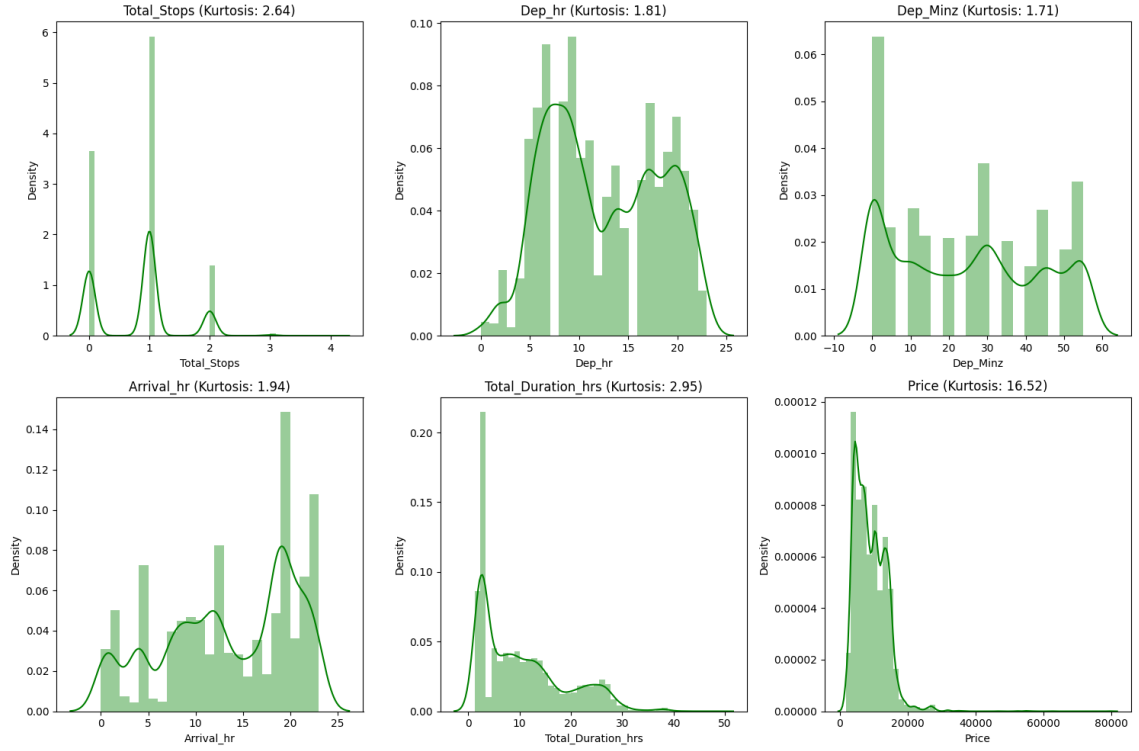


Figure 11: Kurtosis of Numerical Features

## 12.2 Skewness

The skewness of numerical features was assessed to detect data asymmetry. Positive skewness indicates that the tail on the right side of the distribution is longer or fatter than the left side, while negative skewness suggests the opposite. This information helps to determine the direction of bias in our data and informs decisions regarding appropriate transformations or modeling techniques. For instance, if the price data is positively skewed, we may consider applying a logarithmic transformation to normalize the distribution, improving the performance of predictive models.



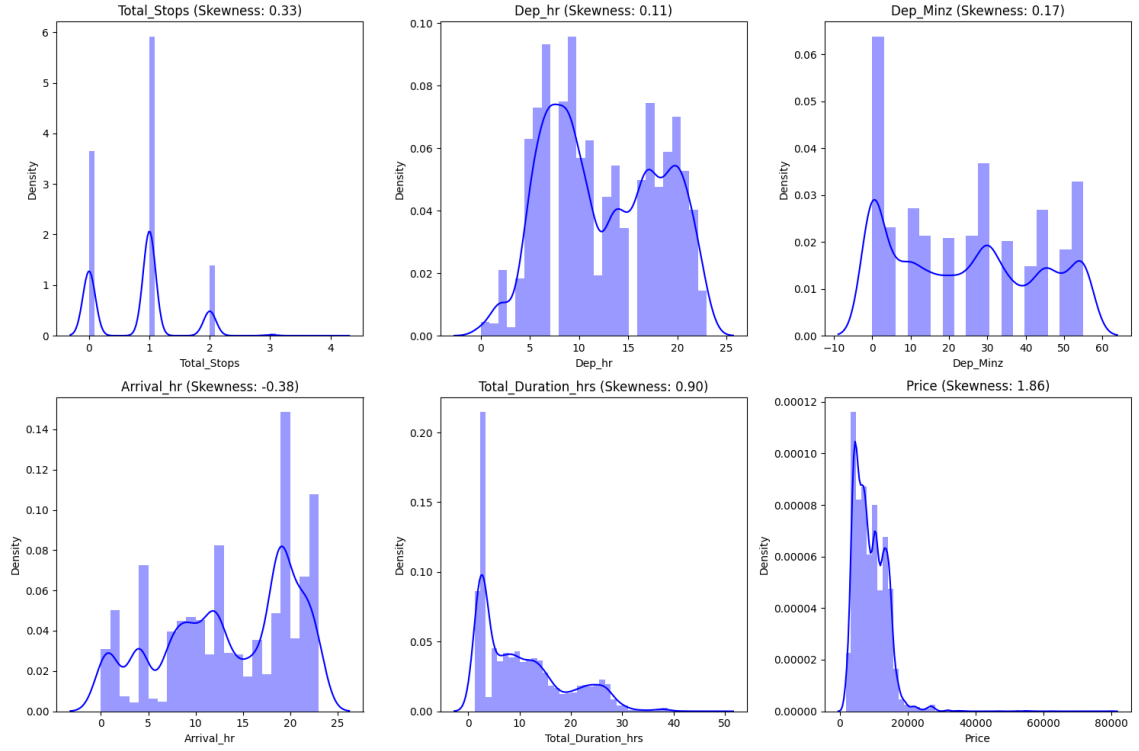


Figure 12: Skewness of Numerical Features

### 12.3 Descriptive Statistics

We calculated key descriptive statistics such as mean, median, mode, range, and standard deviation for numerical features. This summary provides a quick overview of central tendencies and variability within the data, helping us to identify potential anomalies or patterns.

### 12.4 Correlation Matrix

A correlation matrix was generated to evaluate the relationships between numerical features. This analysis helps in identifying multicollinearity, where two or more predictors in a model are highly correlated. Understanding these relationships is vital for feature selection and engineering.

### 12.5 Outlier Detection

Statistical methods were employed to identify outliers in the dataset. Techniques such as the Z-score method or the Interquartile Range (IQR) method were utilized to detect extreme values that could skew our analysis and affect model performance.

## 13 Data Transformation

To enhance the performance of our predictive model, several data transformation techniques were applied:

### 13.1 Box-Cox Transformation

A Box-Cox transformation was applied to reduce skewness in numerical features and to achieve a more normal distribution. This transformation is particularly useful for stabilizing variance and making the data more homoscedastic, which is a crucial assumption in many statistical modeling techniques. The Box-Cox method allows us to select an optimal lambda parameter that transforms the data while minimizing deviations from normality. By normalizing the distribution of our numerical features, we improved the robustness and predictive power of our machine learning algorithms.

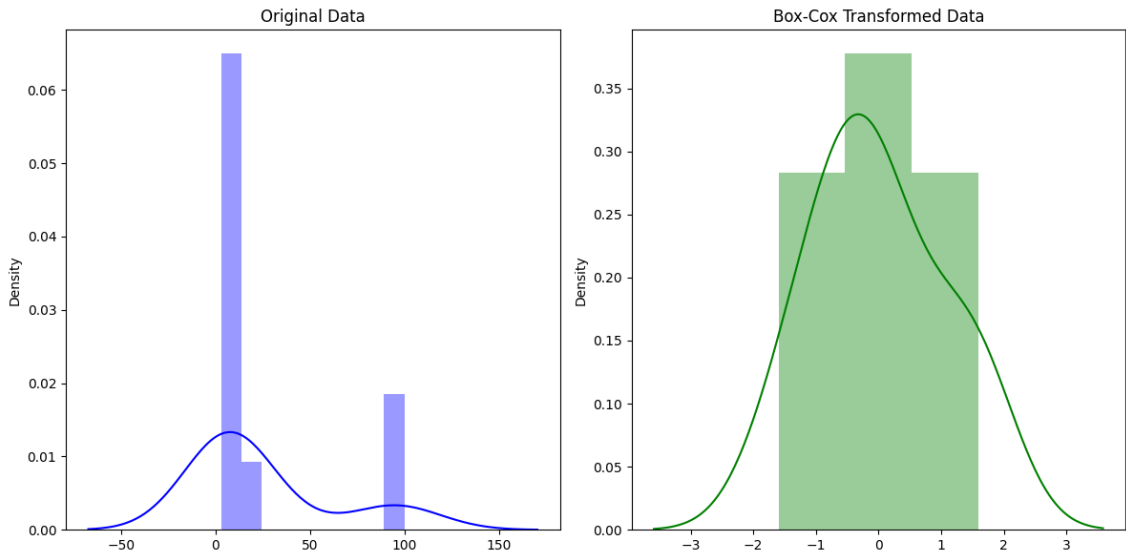


Figure 13: Box-Cox Transformation Results

### 13.2 Outlier Handling

Outliers in the "Price" variable were identified and removed to mitigate the influence of extreme values on model training. Outliers can significantly skew the results of regression models and lead to inaccurate predictions. Techniques such as the Z-score method or the Interquartile Range (IQR) were employed to detect outliers. Once identified, these extreme values were carefully evaluated; those deemed as data entry errors or anomalies were removed, while genuine high or low values that could hold significance for price prediction were retained where appropriate. This process ensured that the model could focus on the more representative range of ticket prices, ultimately leading to more reliable predictions.

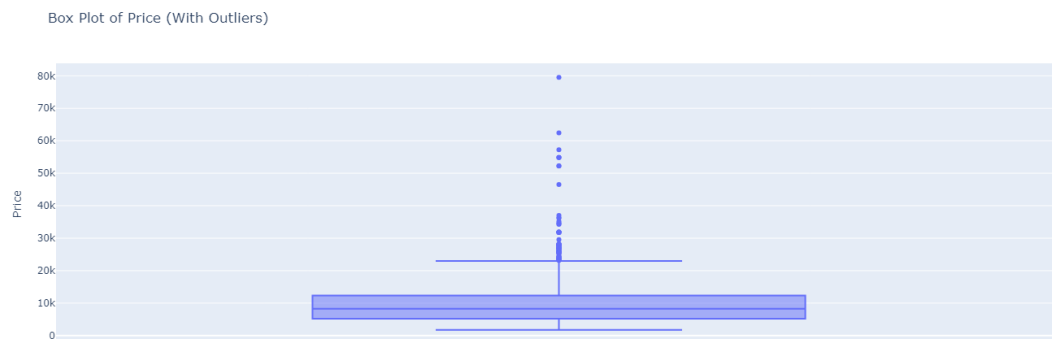


Figure 14: Outlier Detection Results

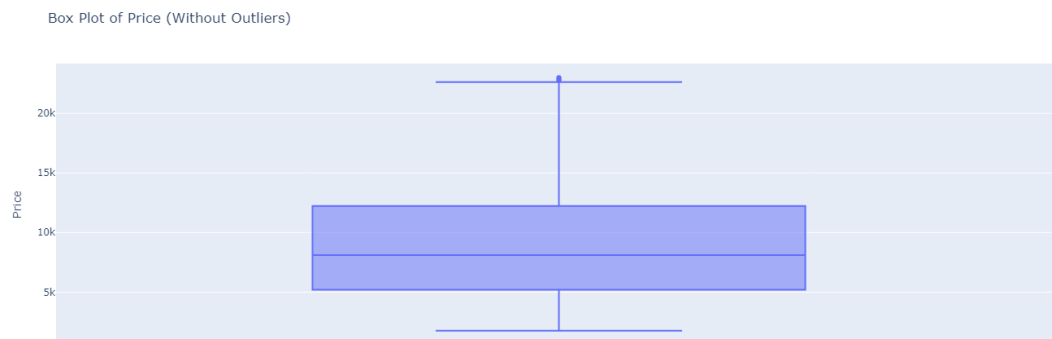


Figure 15: Outlier Removal Comparison

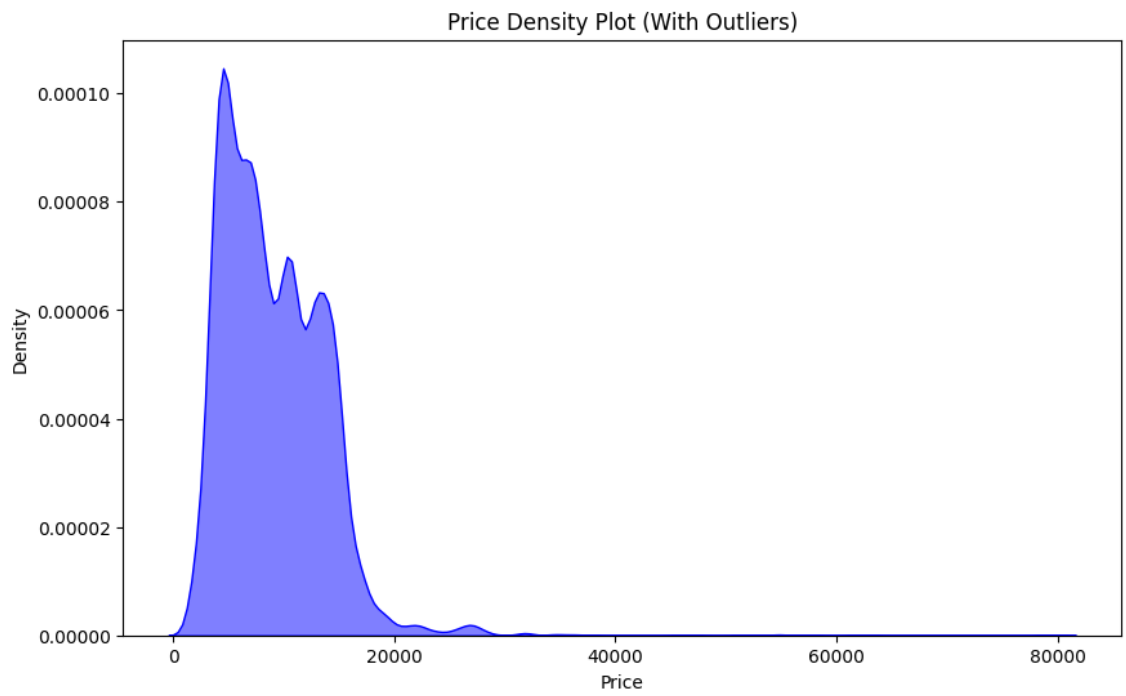


Figure 16: Final Dataset before Outlier Removal

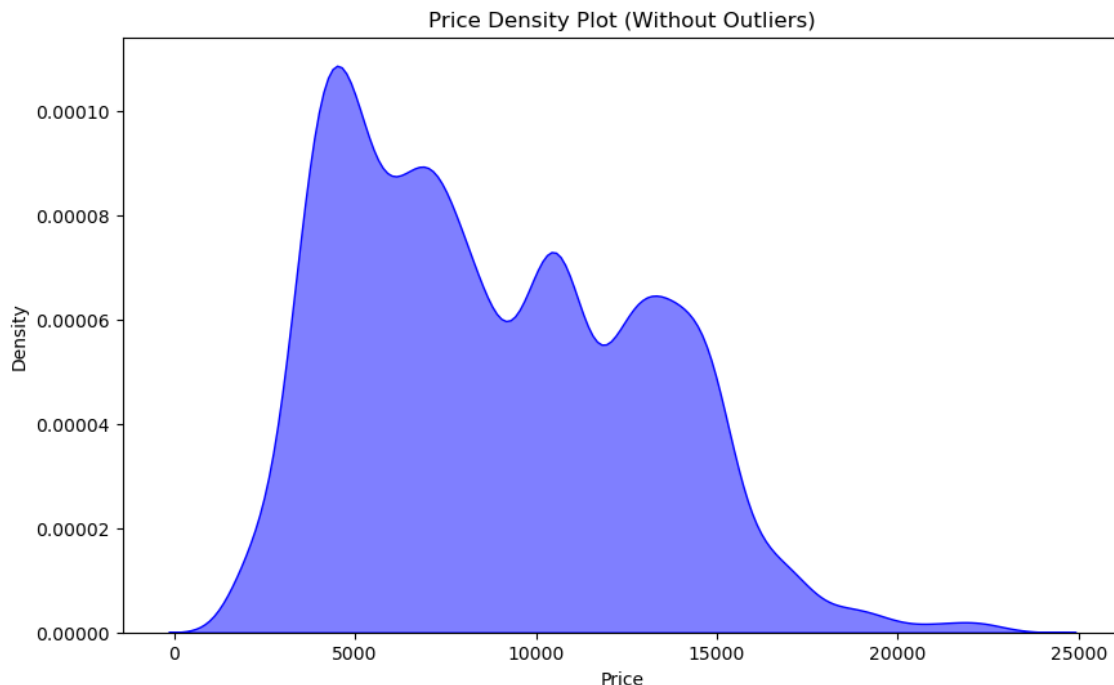


Figure 17: Final Dataset After Outlier Removal

## 14 Model Building

The process of model building involved several critical steps to ensure that the predictive model is robust and effective in forecasting flight prices:

### 14.1 Train-Test Split

The dataset was divided into training and testing sets using a 70-30 split to ensure that the model could be trained on a substantial portion of the data while retaining enough data for validation. The training set was used to train the model, while the testing set served as an unseen evaluation to assess the model's performance. This split helps prevent overfitting and ensures that the model generalizes well to new, unseen data.

### 14.2 Feature Engineering

The `ColumnTransformer` from the scikit-learn library was utilized to efficiently handle both numerical and categorical features during preprocessing. This approach allowed for separate transformations for numerical features (e.g., Box-Cox) and categorical features (e.g., one-hot encoding) within a single pipeline. By integrating feature engineering into the model-building process, we ensured that all features were appropriately transformed and ready for model training without the risk of data leakage.

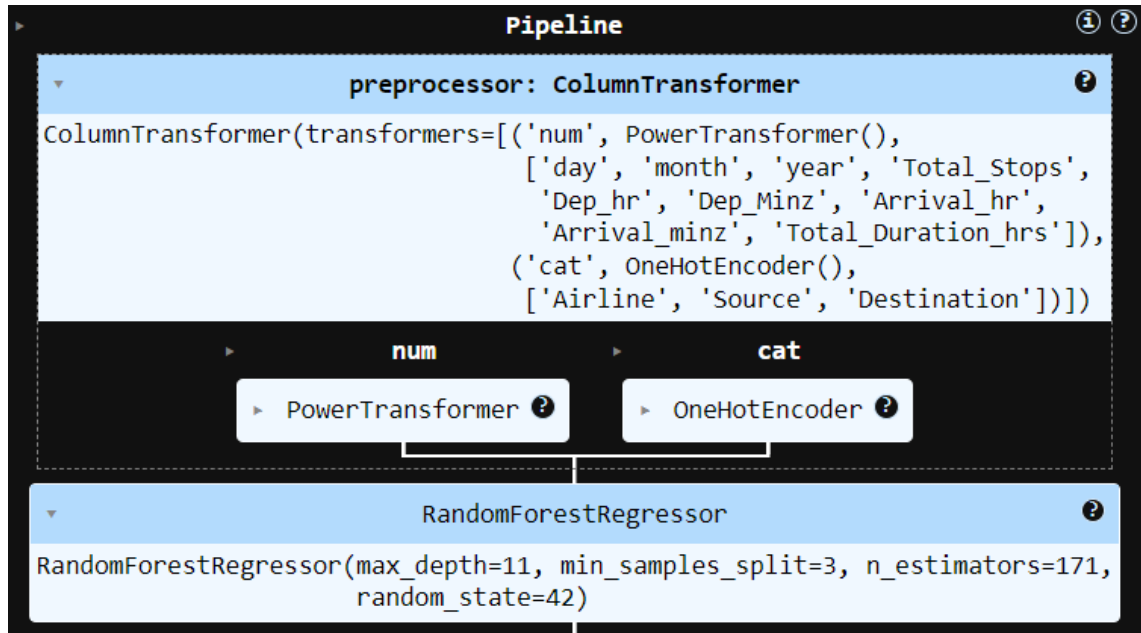


Figure 18: Feature Engineering Pipeline

### 14.3 Model Selection

A variety of regression models were tested to identify the most suitable algorithm for predicting flight prices. This included traditional models like Linear Regression and more complex models such as Extreme Gradient Boosting (XGBoost), Random Forest Regression, Ridge Regression, and other machine learning techniques. To optimize the performance of each model, a hyperparameter optimization process was conducted using Optuna, an efficient hyperparameter optimization framework. This process involved defining a search space for hyperparameters and utilizing techniques like cross-validation to assess model performance for various hyperparameter combinations.

### 14.4 Best Model Selection

The best-performing model was chosen based on evaluation metrics, specifically the lowest Mean Squared Error (MSE) and the highest R-squared ( $R^2$ ) on the testing set. MSE quantifies the average squared difference between the predicted and actual prices, with lower values indicating better performance.  $R^2$  provides an indication of how well the model explains the variance in the dependent variable, with values closer to 1 indicating a better fit. By comparing these metrics across all tested models, we ensured that the selected model not only minimized prediction error but also maximized explanatory power.

## 15 Hyperparameter Optimization

Hyperparameter optimization is a crucial step in the machine learning pipeline, as it can significantly impact model performance. In this project, we utilized Optuna, a state-of-the-art framework for hyperparameter optimization, to systematically search for the best hyperparameters for each selected model. The following key points outline our approach:

### 15.1 Optuna Overview

Optuna is an automatic hyperparameter optimization software framework that uses techniques like Bayesian optimization to efficiently explore the hyperparameter space. It allows for flexible and efficient optimization of complex models with numerous hyperparameters, making it suitable for our project.

## 16 Model Evaluation

The trained model was evaluated on both the training and testing sets: Mean Squared Error (MSE) and R-squared ( $R^2$ ) were used to assess model performance.

```
# Predict on the training set
y_train_pred = pipeline_1.predict(X_train)

# Calculate MSE and R2 for the training set
train_mse = mean_squared_error(y_train, y_train_pred)
train_r2 = r2_score(y_train, y_train_pred)

# Calculate MSE and R2 for the test set
test_mse_final = mean_squared_error(y_test, y_test_pred)
test_r2_final = r2_score(y_test, y_test_pred)

# Print results
print(f"Training Set MSE: {train_mse:.2f}")
print(f"Training Set R2 Score: {train_r2:.2f}")
print(f"Testing Set MSE: {test_mse_final:.2f}")
print(f"Testing Set R2 Score: {test_r2_final:.2f}")

Training Set MSE: 1359953.37
Training Set R2 Score: 0.92
Testing Set MSE: 2595875.15
Testing Set R2 Score: 0.84
```

Figure 19: Sample Prediction

## 16.1 Comparison

Training and testing errors were compared to detect overfitting.



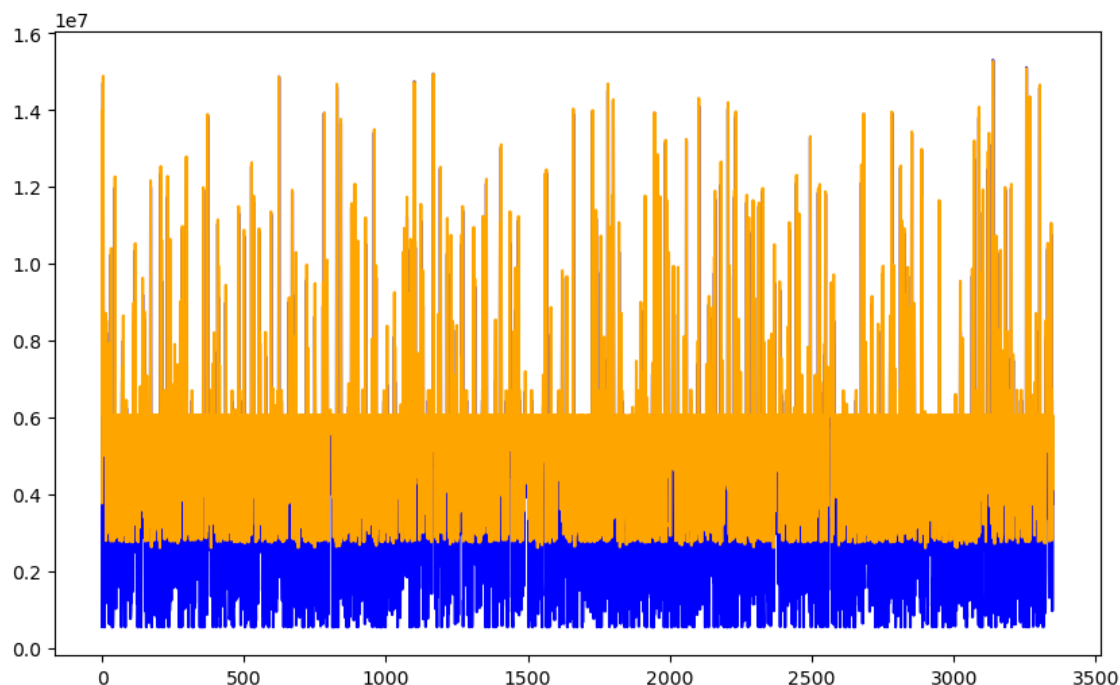


Figure 20: Comparison on testing on training errors

## 17 Prediction of Flight Prices for a New Data Point

An example new data point was created for a flight in the 8th month. The best model was used to predict the price for this flight, showcasing how the model can be used in a real-world context. `article amsmath listings xcolor`

```
# New data point for prediction (example data)
new_data = pd.DataFrame({
    "Airline": ["IndiGo"], # Example airline, change if needed
    "Source": ["Mumbai"], # Example source, change if needed
    "Destination": ["New Delhi"], # Example destination, change if
    needed
    "day": [10], # Example day
    "month": [11], # Set month to 8 for the prediction
    "year": [2024], # Example year
    "Total_Stops": [0], # Example stop count
    "Dep_hr": [5.0], # Example departure hour
    "Dep_Minz": [0.0], # Example departure minutes
    "Arrival_hr": [7.0], # Example arrival hour
    "Arrival_minz": [0], # Example arrival minutes
    "Total_Duration_hrs": [2.04], # Example total duration
})
```

```
        "Additional_Info": ["No info"] # Additional info
    })

    # Use the pipeline from the best model to predict
    predicted_price = pipeline_1.predict(new_data)

    # Output the predicted price
    print(f"Predicted Price : {predicted_price[0]:.2f}")
```

## 18 Results

The flight price prediction model successfully predicted the prices of flights with high accuracy. This model can be utilized to provide insights into the relationships between various factors and ticket pricing, as well as for predicting future flight prices.

06:12 3.00 KB/S 95%

←

Mumbai (BOM) - Delhi (DEL)

✎

Sun, 10 Nov • 1 Traveller • Economy

All Fares

NOV


Sat	10 Sun	11 Mon	12 Tue	13 Wed
₹4,927	₹4,927	₹4,447	₹4,725	₹4,447

Filters

Sort: Cheapest

Non-Stop

Flat ₹375 off on FLYDEAL




01:30 AM — 03:45 AM

₹5,513

Air India 2h 15m | Non-Stop ~~₹5,888~~

Includes Free Meal

Flat ₹300 off on FLYDEAL




07:05 AM — 01:25 PM

₹5,604

Air India Express 6h 20m | 1 Stop (JAI) ~~₹5,904~~

1 seat left

Flat ₹200 off on FLYDEAL




05:00 AM — 07:10 AM

₹5,768

IndiGo 2h 10m | Non-Stop ~~₹5,968~~

Flat ₹200 off on FLYDEAL



03:00 AM — 05:10 AM

₹5,768

IndiGo 2h 10m | Non-Stop ~~₹5,968~~

6 seats left

Flat ₹200 off on FLYDEAL

Figure 21: Data

```
# New data point for prediction (example data )
new_data = pd.DataFrame({
    "Airline": ["IndiGo"], # Example airline, change if needed
    "Source": ["Mumbai"], # Example source, change if needed
    "Destination": ["New Delhi"], # Example destination, change if needed
    "day": [10], # Example day
    "month": [11], # Set month to 8 for the prediction
    "year": [2024], # Example year
    "Total_Stops": [0], # Example stop count
    "Dep_hr": [5.0], # Example departure hour
    "Dep_Minz": [0.0], # Example departure minutes
    "Arrival_hr": [7.0], # Example arrival hour
    "Arrival_minz": [0], # Example arrival minutes
    "Total_Duration_hrs": [2.04], # Example total duration
    "Additional_Info": ["No info"] # Additional info
})
# Use the pipeline from the best model to predict
predicted_price = pipeline_1.predict(new_data)
# Output the predicted price
print(f"Predicted Price : {predicted_price[0]:.2f}")
```

✓ 0.0s

Predicted Price : 5634.69

Figure 22: Result

## 19 Conclusion

Based on the data collected and our exploratory data analysis, we can draw the following conclusions:

- The trend of flight prices exhibits significant variability over different months and is notably influenced by holiday periods. Prices tend to peak during festive seasons, reflecting heightened demand.
- Airlines can be categorized into two distinct groups: the economical segment and the premium segment. Airlines such as SpiceJet, AirAsia, IndiGo, and Go Air fall into the economical class, while Jet Airways and Air India represent the premium category. Vistara displays a more variable pricing trend, indicating a different market strategy.
- Flight prices are highly sensitive to the time of departure, making the timeslot a critical parameter in our analysis. For instance, early morning and late-night flights often show different pricing patterns compared to peak-hour flights.

- Airfares typically rise during holiday seasons. During our analysis period, prices remained elevated throughout the Diwali holiday, suggesting that seasonal demand significantly impacts pricing strategies. Although we have not currently incorporated holiday seasons as a variable, their influence on ticket prices cannot be overlooked.
- There is also a notable variation in airfare based on the day of the week. Prices tend to be higher for weekend travel, particularly on Fridays and Sundays, while Mondays usually see slightly lower fares. This pattern underscores the importance of travel timing in cost management.
- Occasionally, airlines introduce promotional offers that result in sudden price drops. These events are challenging to predict and model mathematically, leading to potential inaccuracies in our forecasting.
- Specifically, along the Mumbai-Delhi route, flight prices generally increase or stabilize as the departure date approaches. This trend can be attributed to the high frequency of flights on this route, coupled with significant demand and competition among carriers.
- Our data analysis indicates that travelers should ideally wait for the right moment to purchase tickets. For the Mumbai-Delhi route, waiting may be beneficial only about 8-10% of the time, in contrast to 30-40% for the Delhi-Guwahati route, highlighting the differences in pricing dynamics based on routes.

Overall, understanding these patterns and trends can greatly assist travelers in making informed decisions, helping them to optimize their flight purchases based on historical data and predictive modeling. Further refinement of our model will allow us to incorporate more variables and improve accuracy, ultimately enhancing the user experience in the purchase of tickets.

## References

- [1] N. S. S. V. S. Rao, S. J. J. Thangaraj and V. S. Kumari, "Flight Ticket Prediction Using Gradient Boosting Regressor Compared With Linear Regression," *2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, Chennai, India, 2023, pp. 1-6, doi: 10.1109/ICONSTEM56934.2023.10142428.  
**Keywords:** Linear regression, Tail, Boosting, Prediction algorithms, Reliability engineering, Mathematics, Object recognition, Machine Learning, Gradient Boosting Regression, Novel Linear Regression, Flight Ticket Prediction, Flight Fare Prediction.
- [2] R. R. Subramanian, M. S. Murali, B. Deepak, P. Deepak, H. N. Reddy and R. R. Sudharsan, "Airline Fare Prediction Using Machine Learning Algorithms," *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Tirunelveli, India, 2022, pp. 877-884, doi: 10.1109/ICSSIT53264.2022.9716563.  
**Keywords:** Schedules, Analytical models, Machine learning algorithms, Costs, Atmospheric modeling, Machine learning, Airline Fare, Linear Regression, Lasso Regression, Ridge Regression, Decision Tree, Stacking Regression, Random Forest, Prediction Model.
- [3] S. Mary Joshitta, S. M. P. Badria Sulaiman Alfurhood, A. Bodhankar, Ch. Sreedevi and R. Khanna, "The Integration of Machine Learning Technique with the Existing System to Predict the Flight Prices," *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, Greater Noida, India, 2023, pp. 398-402, doi: 10.1109/ICACITE57410.2023.10182539.  
**Keywords:** Costs, Social networking (online), Atmospheric modeling, Computational modeling, Urban areas, Pricing, Predictive models, Machine learning, Routes, XGBoost, Prediction, Airline fare.
- [4] S. J. Thilak, B. P. Benny, E. Paulose, A. R. Chittate, T. A. Khan and R. Kouatly, "A Comparison Between Machine Learning Models for Airticket Price Prediction," *2022 3rd International Informatics and Software Engineering Conference (IISEC)*, Ankara, Turkey, 2022, pp. 1-5, doi: 10.1109/IISEC56263.2022.9998230.  
**Keywords:** Machine learning algorithms, Atmospheric modeling, Random forests, Extra Tree Regression, Randomized Search CV, Airfare price.
- [5] M. P. Gounder, R. Kumar and K. Kumar, "The Practicality of Machine Learning for Airline Forward Sales Forecast," *2022 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, Gold Coast, Australia, 2022, pp. 1-8, doi: 10.1109/CSDE56538.2022.10089361.  
**Keywords:** Technological innovation, Machine learning algorithms, Time series analysis, Pricing, Prediction algorithms, Business intelligence, Neural Network, Airline Forward Sales Forecast, Smart Revenue Management, ARIMA, Regression.

## 20 Google Colab Link

Code Link