**Machine Learning Project Assessment**

**Presented By:**

Prashamsa Rijal

Student Id:23140743

Word Count:1421

**Table of Contents**

[**I.** **Introduction** 3](#_Toc183095875)

[**II.** **Problem Description** 3](#_Toc183095876)

[**III.** **Objective** 3](#_Toc183095877)

[**IV.** **Dataset** 3](#_Toc183095878)

[**V.** **Data Preprocessing** 4](#_Toc183095879)

[**1.1.** **Feature Engineering** 4](#_Toc183095880)

[**1.2.** **Data Splitting** 4](#_Toc183095881)

[**VI.** **Exploratory Data Analysis (EDA)** 4](#_Toc183095882)

[**1.3.** **Insights from EDA** 4](#_Toc183095883)

[**A.** **Airline Influence on Flight Prices** 4](#_Toc183095884)

[**B.** **Impact of Stops on Flight Prices** 4](#_Toc183095885)

[**C.** **Duration as a Price Determinant** 5](#_Toc183095886)

[**VII.** **Modeling Approach** 5](#_Toc183095887)

[**1.4.** **Model Selection** 5](#_Toc183095888)

[**A.** **Model Performance Results** 5](#_Toc183095889)

[**B.** **Analysis of Model Results** 5](#_Toc183095890)

[**1.5.** **Model Tuning** 6](#_Toc183095891)

[**1.6.** **Evaluation Metrics** 6](#_Toc183095892)

[**VIII.** **Model Interpretation** 7](#_Toc183095893)

[**IX.** **Conclusion** 7](#_Toc183095894)

[**X.** **References** 8](#_Toc183095895)

# **Introduction**

Airfare pricing has become more intricate than ever, and I wanted to dive deeper into the factors that drive flight ticket costs. To do this, I decided to build a predictive model using a dataset from Ease My Trip, available on Kaggle. This dataset offers a wealth of information, including the airline, departure and arrival cities, travel route, total flight duration, number of stops, and even details like meal availability. By analyzing these features, I aimed to uncover the key variables that influence how airlines set their prices.

# **Problem Description**

Flight ticket prices often fluctuate significantly, influenced by various factors such as the airline, travel route, flight duration, and more. Predicting these prices is not only challenging but also highly valuable—helping travelers make smarter booking decisions and enabling airlines to better understand pricing trends. By analyzing this dataset, my goal was to uncover the key elements driving airfare costs and develop a predictive model to forecast ticket prices with greater accuracy.

# **Objective**

The primary goal of this project was to identify the best machine learning model for predicting flight ticket prices. After experimenting with various algorithms, I selected **XGBoostRegressor** as the final model due to its ability to effectively handle the data and deliver accurate predictions. It captured the subtle relationships between different features and their impact on ticket pricing. This project aims to offer valuable insights to help travelers make informed booking decisions while also shedding light on the factor’s airlines consider when determining ticket prices.

# **Dataset**

The dataset used for this study is sourced from Kaggle, under the title [Flight Price Prediction Dataset](https://www.kaggle.com/datasets/jillanisofttech/flight-price-prediction-dataset). It contains information about 10,683 flight bookings, including the following columns:

1. **Airline**: The airline operating the flight.
2. **Date\_of\_Journey**: The date when the flight was booked.
3. **Source**: Departure city.
4. **Destination**: Arrival city.
5. **Route**: Flight route including intermediate stops.
6. **Dep\_Time**: Time of departure.
7. **Arrival\_Time**: Time of arrival.
8. **Duration**: Total flight duration.
9. **Total\_Stops**: The number of stops during the flight.
10. **Additional\_Info**: Additional flight information, such as the presence of an in-flight meal.
11. **Price**: The ticket price.

# **Data Preprocessing**

To prepare the dataset, missing values in the *Route* and *Total\_Stops* columns were filled with their most frequent values, and duplicate entries were removed to ensure data quality

## **Feature Engineering**

Several transformations were applied to enhance the dataset for modeling:

* **Date Columns**: Date\_of\_Journey was converted to datetime format, allowing extraction of Day\_of\_Week and Month to capture booking trends.
* **Time Information**: The Dep\_Time column was split into hour and minute components, adding time-related insights.
* **Categorical Data**: Columns like Airline, Source, and Destination were encoded (label or one-hot) for compatibility with the model.
* **Flight Duration**: Duration was transformed from a mixed format (e.g., "2h 50m") into total minutes for consistency and simplicity.

## **Data Splitting**

An 80-20 split was applied to divide the data into training and testing sets. This approach ensured that the model trains on a substantial portion of the data, with 20% reserved for evaluating generalizability and performance.

# **Exploratory Data Analysis (EDA)**

EDA helps uncover patterns, correlations, and anomalies in the data that impact flight prices. This section presents key visualizations and insights to highlight influential trends in the dataset.

## **Insights from EDA**

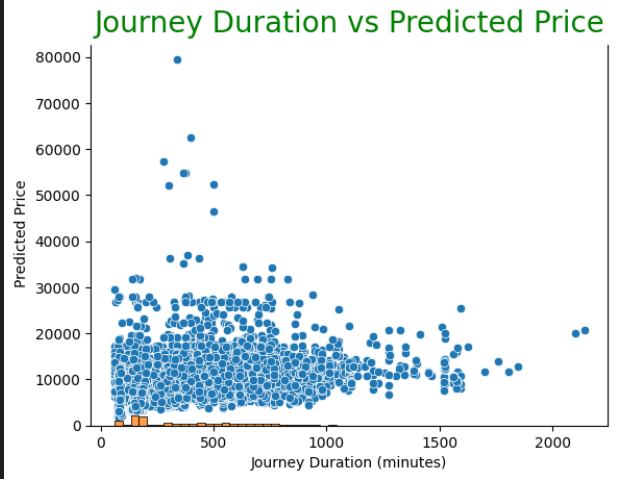
## **Most Popular Flight Route**

* **Most Popular Route**: Delhi → Mumbai → Cochin (DEL → BOM → COK) is by far the most frequent route.
* **Other Busy Routes**: Delhi → Bangalore, Kolkata → Bangalore, and Mumbai → Hyderabad also have high numbers of flights.
* **City Hubs**: Delhi, Mumbai, and Cochin are major travel hubs, with many routes passing through them.
* **Variety in Routes**: While a few routes dominate, there are also many other city combinations showing diverse travel preferences.

A graph showing a number of routes

Description automatically generated with medium confidence

## **Jounery Duration Impact on Price**

* Longer flights = higher fares: Most data points show that longer journeys are linked to higher prices.
* Duration increases from 100 minutes to 33 hours, and prices follow a noticeable upward trend.

## **Seasonal Price Variation**

A graph showing a line

Description automatically generatedThe sharp drop in early March, as seen in the graph, suggests that there may be seasonal patterns influencing flight pricing, with airlines offering discounts or experiencing lower demand during this period.

# **Modeling Approach**

## **Model Selection**

## **Model Performance Results**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Train Accuracy** | **Test Accuracy** | **CV Mean Score (Train)** |
| LinearRegression (Tuned) | 0.72 | 0.69 | 0.71 |
| SGDRegressor | 0.35 | 0.33 | 0.34 |
| RandomForestRegressor | 0.93 | 0.84 | 0.85 |
| MLPRegressor (Tuned) | 0.50 | 0.48 | 0.46 |
| XGBRegressor | 0.95 | 0.84 | 0.84 |
| Ridge | 0.72 | 0.69 | 0.71 |
| Lasso | 0.72 | 0.69 | 0.71 |
| KNeighborsRegressor | 0.98 | 0.60 | 0.57 |
| DecisionTreeRegressor | 0.93 | 0.80 | 0.79 |
| GradientBoostingRegressor | 0.95 | 0.84 | 0.84 |
| RidgeCV | 0.72 | 0.69 | 0.71 |

## **Analysis of Model Results**

* **Top Performers:** XGBoost and Random Forest delivered the best results, with both achieving high training accuracy (95% and 93%, respectively) and strong test accuracy (84%), making them reliable choices. XGBoost maintained an excellent balance between train and test accuracy.
* **Underperformers:** MLP Regressor and SGD Regressor struggled, with accuracies below 50%, indicating they are not suitable for this task.
* **Overfitting Concerns:** K-Nearest Neighbors (98% train accuracy) and Decision Tree Regressor (93% train accuracy) showed significant drops in test accuracy (60% and 80%, respectively), suggesting overfitting.
* **Consistent Alternatives:** Ridge and Lasso Regression provided stable performance, with similar train and test accuracies (72% and 69%, respectively), making them good alternatives for simpler, interpretable models.

XGBoost and Random Forest stood out as the most reliable models for predicting flight prices due to their combination of high accuracy, consistency, and robustness.

## **Model Tuning**

I applied **GridSearchCV** to fine-tune hyperparameters across all the models. This process involved exploring combinations of parameters such as learning rates, tree depths, and the number of estimators. The tuning enhanced performance, with XGBoost and Random Forest showing the most significant improvements, further solidifying their status as the best-performing models for predicting flight prices.

## **Evaluation Metrics**

To evaluate model performance, the following metrics were used:

1. **Mean Absolute Error (MAE)**

This metric provides the average magnitude of prediction errors, giving an overall sense of how much the model's predictions deviate from actual prices.

1. **Median Absolute Error**

Unlike MAE, which is influenced by extreme outliers, this metric uses the median error to offer a more robust measure of model performance

1. **Train/Test Split Accuracy**

This compares the model’s performance on both training and test datasets, ensuring that it generalizes well to new, unseen data.

1. **Cross-Validation Scores**

K-fold cross-validation was used to validate performance across different data splits, helping to prevent overfitting and ensure the model's robustness.

# **Model Interpretation**

LIME (Local Interpretable Model-agnostic Explanations) was used to explain how different features influenced the XGBoost model's flight price predictions. Key findings include:

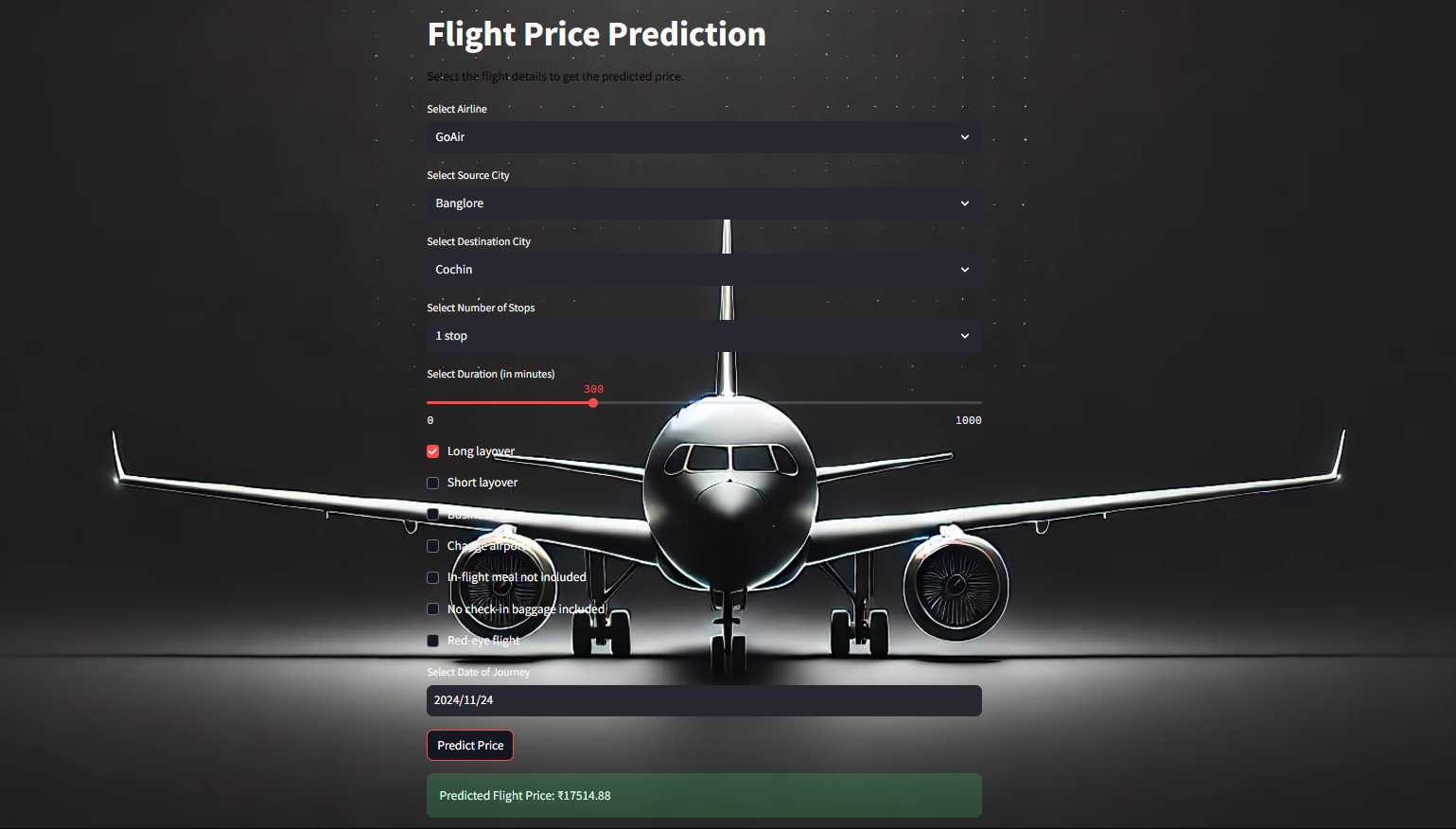
1. **Airline:** Premium airlines, such as Jet Airways and Air India, generally have higher ticket prices due to enhanced services and comfort, whereas budget airlines like IndiGo and SpiceJet offer more economical options.
2. **Flight Duration:** Longer flights or those with layovers tend to have higher costs, reflecting increased operational expenses and added travel time.
3. **Total Stops:** Non-stop flights are typically more expensive because they offer greater convenience and shorter travel durations compared to flights with multiple stops.
4. **Source and Destination:** Flights to popular or high-demand destinations, as well as routes from busy hubs, often command higher fares due to elevated demand and operational complexities.
5. **Date of Journey:** Flight prices increase as the departure date approaches, aligning with typical airline pricing strategies where early bookings are incentivized with lower fares, and last-minute bookings incur premium charges.

A screenshot of a computer

Description automatically generated

# **Web Application Using Streamlit**

The Streamlit app predicts flight prices based on user inputs like destination and stops. It compares the inputs with precomputed predictions from a CSV file and displays the result in a simple, user-friendly interface with a custom background image for a polished look.



# **Conclusion**

This project explored the factors that influence flight ticket prices, using data from the "Ease My Trip" dataset. Variables like airline, flight duration, number of stops, travel route, and journey date proved to be key in determining ticket costs.

Through testing various machine learning models, **XGBoost** stood out as the top performer, delivering the most accurate predictions with the lowest error rates and the highest reliability.

**Key Takeaways:**

* Premium airlines and non-stop flights are pricier due to added convenience and superior services.
* Flights with longer durations or multiple stops tend to be cheaper, offering a cost-saving option for those with flexible schedules.
* The starting and ending points of a journey play a critical role in price fluctuations.
* Ticket prices increase as the departure date approaches, making early bookings a smarter choice for budget-conscious travelers.

By uncovering these insights, this study not only aids travelers in planning better but also offers a clearer perspective on the factors airlines consider when pricing tickets.

# **References**

* Kaggle. (n.d.). Flight Price Prediction Dataset. Retrieved from <https://www.kaggle.com/datasets/jillanisofttech/flight-price-prediction-dataset>
* scikit-learn developers. (n.d.). GridSearchCV. Retrieved from <https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html>
* scikit-learn developers. (n.d.). Random Forest Regressor. Retrieved from <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>
* scikit-learn developers. (n.d.). XGBoost. Retrieved from <https://scikit-learn.org/stable/modules/generated/xgboost.XGBRegressor.html>