Logo, company name

Description automatically generated

**The Final Project report: Flight Price Prediction**

Submitted in partial fulfillment of the requirement for project deliverable.

**PYTHON PROGRAMMING**

**SPRING SEMESTER 2024**

**Submitted by Group - 7**

**BOMMINENI ALEKHYA**

**YASHWANTH THALLA**

**MAMIDALA RUTHWIK**

Supervised by

**Professor Kshitij Sharma**

**PACE UNIVERSITY**

**SEIDENBERG SCHOOL OF COMPUTER SCIENCE AND INFORMATION SYSTEMS**

Contents

[Abstract: 3](#_Toc165469210)

[Introduction 3](#_Toc165469211)

[Project Workflow 3](#_Toc165469212)

[Dataset: 3](#_Toc165469213)

[Data Pre-processing and EDA: 3](#_Toc165469214)

[Model Training and Evaluation: 13](#_Toc165469215)

[Linear Regression: 13](#_Toc165469216)

[Decision Tree: 14](#_Toc165469217)

[Random Forest: 15](#_Toc165469218)

[Results: 16](#_Toc165469219)

[Conclusion 17](#_Toc165469220)

[Recommendations: 17](#_Toc165469221)

[Future Work: 17](#_Toc165469222)

[References and Links: 18](#_Toc165469223)

# 

# Abstract:

This study undertakes a comprehensive evaluation of three prevalent machine learning models—Linear Regression, Decision Tree, and Random Forest—by analyzing their performance on a standardized dataset preprocessed for optimal analysis. The research incorporates exploratory data analysis (EDA) to identify underlying patterns, distributions, and anomalies within the dataset, followed by rigorous data preprocessing techniques such as handling missing values, feature scaling, and encoding categorical variables. Performance metrics including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R-squared) are employed to compare the accuracy, consistency, and reliability of each model. Results indicate that Decision Trees provide high accuracy but may risk overfitting, Linear Regression offers clarity in linear relationships but is limited by its assumptions, and Random Forests achieve a balance of accuracy and stability, demonstrating robustness across complex datasets.

# Introduction

Predictive modelling is a cornerstone of machine learning, enabling effective interpretation of data for decision-making. Among the diverse array of algorithms, Linear Regression, Decision Trees, and Random Forests stand out for their unique capabilities and widespread application across various domains. Before applying these models, exploratory data analysis (EDA) is essential to understand the dataset's characteristics, such as its central tendencies, dispersion, and potential correlations between variables. This initial phase helps in uncovering insights, guiding the subsequent data preprocessing steps.

Data preprocessing is a critical step that ensures the quality and suitability of data for modelling. It involves cleaning the data, dealing with missing values, normalizing or scaling features, and encoding categorical variables to convert them into a machine-readable format. This study integrates these preliminary steps to prepare the dataset for the application of three machine learning models, assessing each model's performance based on accuracy, error metrics, and explanatory power.

This paper details the methods and results of EDA and data preprocessing, followed by a comparative analysis of the models using quantitative performance metrics. The aim is to identify which model performs best under various conditions, offering insights into their applicability for specific analytical tasks. By examining Linear Regression, Decision Trees, and Random Forests within this framework, the paper seeks to provide a comprehensive guide to selecting appropriate modelling techniques based on the specific needs and characteristics of the dataset, ultimately enhancing the effectiveness of predictive analytics in real-world applications.

# Project Workflow

Dataset: [Flight Price Prediction (kaggle.com)](https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction)

## Data Pre-processing and EDA:

**Removing unwanted Columns:**

A close-up of a computer screen

Description automatically generated

**Changing datatypes of numerical features to 32 bit for reducing memory usage:**

A screenshot of a computer code

Description automatically generated

**Descriptive Statistics:**

**A screenshot of a computer

Description automatically generated**

**Checking for missing values:**

**A screenshot of a computer code

Description automatically generated**

**Checked for Duplicated values:**

**A screen shot of a computer code

Description automatically generated**

**Data Distributions:**

A graph showing a distribution of flight prices

Description automatically generated

**Observations from the Graph:**

**Skewness:** The distribution of flight prices is right-skewed, indicating that most of the flight prices are clustered towards the lower end of the scale, but there are a few relatively expensive flights.

**Primary Mode:** The most frequent price range appears to be just under 10,000, suggesting this is the most common price for flights within the dataset.

**Secondary Peaks:** There are noticeable secondary peaks around 40,000andjustunder100,000. These peaks could represent premium or luxury flight options, or perhaps flights to particularly distant or popular destinations.

**Tail of Distribution:** The long tail extending towards the higher price range ($100,000 and beyond) suggests that while expensive flights are less common, there is a significant variety in how high flight prices can go.

A graph of a distribution of duration

Description automatically generated

A graph of a number of days left

Description automatically generated

A graph of a distribution of price

Description automatically generated

**Distribution of Flight Duration**

**Shape and Trend:** The histogram shows a right-skewed distribution with most flights having shorter durations, peaking at 10-15 hours. The frequency then gradually decreases as the duration increases.

**Implications:** This suggests that shorter flights are more common, which could be due to the prevalence of domestic or short-haul international flights. Airlines may prioritize these routes due to higher demand or profitability.

**Distribution of Days Left Before Flight**

**Shape and Trend:** This distribution is somewhat uniform but slightly right-skewed, with frequencies gradually increasing from 0 days left to about 30-35 days, then stabilizing around 25,000 to 30,000 flights until it peaks again slightly around 40-45 days.

**Implications:** This might indicate booking patterns where a significant number of bookings occur both well in advance (30-45 days) and at the last minute. This can help airlines adjust pricing strategies or promotions based on booking timelines.

**Distribution of Price**

**Shape and Trend:** The price distribution is strongly right-skewed with a significant peak at lower prices (below 20,000) and smaller peaks around 40,000 and 80,000.

**Implications:** The large peak at the lower price range suggests that most flights are relatively affordable, with fewer flights in the higher price ranges, likely corresponding to longer or more premium service offerings. This could reflect a broader market focus on budget to mid-range travel options.

**Combined Insights**

**Flight Duration and Price Relationship:** Shorter, more frequent flights are generally cheaper, aligning with the peak of lower prices in the price distribution. Longer flights, which are less frequent, may correspond to the higher price outliers.

**Booking Patterns and Flight Prices:** With a stable booking rate around 30-45 days before flight (as shown in the days left distribution), this might be an optimal time for promotions or price adjustments to capture bookings before traveler commitment drops.

**Strategic Implications:** Airlines can use this data to better understand consumer preferences and adjust their route planning, pricing strategies, and promotional efforts to maximize occupancy and revenue.

A pie chart with different colored circles

Description automatically generated

**Analysis of the Airline Market Shares:**

**Vistara:** Holds the largest share of the market at 42.6%. This dominance suggests Vistara is a leading choice among travelers, possibly due to its routes, service quality, or pricing strategies.

**Air India**: The second largest, with 27.0% of the market share, indicating that it remains a significant player in the airline industry, possibly due to its historical presence and government backing.

**Indigo:** Captures 14.4% of the market, placing it as a substantial but smaller competitor compared to Vistara and Air India. Its positioning might appeal to specific consumer segments.

**SpiceJet and Others:** SpiceJet, at 5.4%, followed by AirAsia at 3.0%, and GO\_FIRST at 7.7%, represent smaller segments of the market. These shares suggest they cater to niche markets or compete on specific routes or price points.

A graph of different colored bars

Description automatically generated

**Observations from the Chart:**

**Delhi and Mumbai:** These two cities have the highest number of flights, both close to 60,000. This high volume likely reflects their roles as major economic and transportation hubs in the country.

**Bangalore and Kolkata:** Both cities have a significant number of flights, roughly around 50,000 each. Bangalore's prominence as a tech hub and Kolkata's historical significance and population might contribute to these numbers.

**Hyderabad and Chennai:** These cities show a slightly lower flight volume compared to the others, each with around 40,000 flights. These figures still indicate their importance as central travel points in their respective regions.

A graph showing a number of days left before flight

Description automatically generated

**Key Observations from the Plot:**

**Price Variability:** There is a wide range of prices across the number of days left before the flight. Prices vary from low to high without a clear pattern of increase or decrease as the departure date approaches.

**High Price Points:** There are flights with very high prices scattered throughout the range of days, suggesting that certain flights remain expensive regardless of how far in advance they are booked.

**Clusters:** Prices are densely packed at certain levels, indicating common pricing tiers utilized by airlines. These clusters do not show a definitive trend related to the days left, which implies that other factors such as flight demand, route popularity, or airline pricing strategies might be influencing the pricing more than the booking lead time.

A graph showing a number of stops

Description automatically generated

**Key Observations from the Plot:**

**Non-Stop Flights (Zero Stops):** These flights have the lowest price range and the smallest interquartile range (IQR), indicating less variability in prices. The majority of prices are concentrated at the lower end, with some outliers indicating more expensive direct flights.

**One Stop:** This category shows a much higher median price and a larger IQR, suggesting greater variability in the prices of one-stop flights. The range extends significantly higher than the zero-stop flights, with several outliers indicating that some one-stop flights can be very expensive.

**Two or More Stops:** Interestingly, the median price for flights with two or more stops is lower than for one-stop flights, and the IQR is also narrower, suggesting less variability. However, there is a notable spread of lower prices and a few high-priced outliers.

**Corelation Plot:**

A screenshot of a graph

Description automatically generated

**Analysis of the Correlation Matrix**

1. **Duration and Days Left (-0.039)**:
   * This value indicates a very weak negative correlation between **duration** and **days\_left**. This suggests that there is almost no linear relationship between how long a duration is and how many days are left. This implies that the duration of whatever is being measured does not significantly change as the event or deadline approaches.
2. **Duration and Price (0.2)**:
   * There is a weak positive correlation between **duration** and **price**. A positive correlation of 0.2 suggests that as the duration increases, the price tends to slightly increase as well, but the relationship is not strong. This could indicate that longer durations might be associated with higher costs, but other factors likely have a more substantial impact on the price.
3. **Days Left and Price (-0.092)**:
   * There is a very weak negative correlation between **days\_left** and **price**. This suggests that as the number of days left decreases, the price might slightly decrease, but again, the relationship is very weak. This could be relevant in scenarios such as booking or reservations, where prices might decrease as the event date approaches to encourage last-minute sales, but the effect here is minimal.

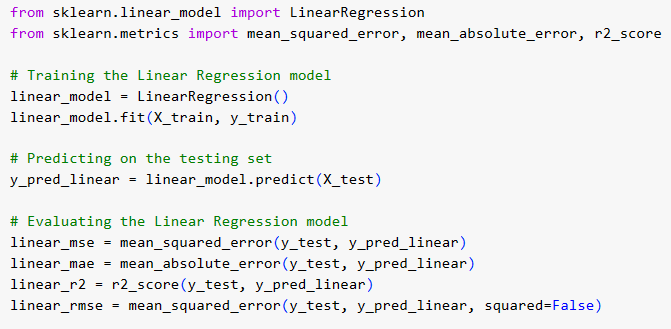
**Train and Split:**

**A screen shot of a computer code

Description automatically generated**

# Model Training and Evaluation:

## Linear Regression:



A screenshot of a computer error

Description automatically generated

## Decision Tree:

A screenshot of a computer program

Description automatically generated

A number and numbers on a white background

Description automatically generated

## Random Forest:

**Garbage Collection and Sampling Data for random forest for reducing time complexity:**

A screen shot of a computer code

Description automatically generated

**Algorithm:**

A screenshot of a computer program

Description automatically generated

# Results:

**Linear Regression**

* **MAE**: 4254.83 indicates that, on average, the predicted values deviate from the actual values by about 4254 units.
* **MSE**: 38409361.97 is quite high, suggesting that there are significant variations in the errors of predictions.
* **RMSE**: 6197.53, also a large value, indicates a high level of error in the currency of the output variable.
* **R-squared**: 0.9255 shows that approximately 92.55% of the variance in the dependent variable is predictable from the independent variables. This is fairly high, suggesting good model performance in terms of explained variability, though the error metrics suggest the predictions are off by a considerable margin.

**Decision Tree**

* **MAE**: 880.62 is much lower than that of Linear Regression, indicating more precise predictions.
* **MSE**: 8889112.54, while lower than Linear Regression’s MSE, still indicates a notable variance in prediction errors.
* **RMSE**: 2981.46, significantly lower than that of Linear Regression, indicates improved accuracy in predictions.
* **R-squared**: 0.9828 is very high, suggesting that the model explains nearly 98.28% of the variance. This indicates excellent model performance.

**Random Forest**

* **MAE**: 1438.97 is higher than the Decision Tree but substantially lower than the Linear Regression model.
* **MSE**: 10051078.12, which is higher than the Decision Tree but lower than Linear Regression, suggesting moderate error variance.
* **RMSE**: 3170.34, also indicating better performance than Linear Regression but not as good as the Decision Tree.
* **R-squared**: 0.9803, very close to the Decision Tree, indicating that the model explains about 98.03% of the variance, which is excellent.

**Comparison and Conclusion**

* **Accuracy**: The Decision Tree model provides the most accurate predictions (lowest MAE and RMSE) among the three, followed by the Random Forest and then Linear Regression.
* **Consistency**: Despite the Decision Tree having the lowest error metrics, Decision Trees are generally more prone to overfitting compared to Random Forests. The Random Forest model, which is an ensemble of decision trees, tends to provide more stable and consistent predictions across different datasets.
* **Model Fit**: The R-squared value is very high for both the Decision Tree and Random Forest, indicating that they both capture the variability of the target variable very well. Linear Regression, while still showing a good fit (92.55%), is outperformed in terms of both error metrics and the proportion of variance explained.
* **Suitability**: Decision Trees and Random Forests are more suitable for complex datasets where relationships between features are non-linear. Linear Regression, while useful for understanding relationships between variables, may not capture complex patterns as effectively as tree-based methods.

# Conclusion

This study evaluated three widely-used machine learning models—Linear Regression, Decision Tree, and Random Forest—through exploratory data analysis (EDA), data preprocessing, and a series of performance metrics (MAE, MSE, RMSE, and R-squared). The analysis highlighted the strengths and weaknesses of each model in handling a standardized dataset, which underwent comprehensive preprocessing to ensure data quality and suitability for modelling.

The Decision Tree model exhibited the highest accuracy but raised concerns about potential overfitting. Linear Regression, while providing valuable insights into linear relationships, was limited by its assumptions and was outperformed by the more complex models in terms of error metrics. The Random Forest model demonstrated a strong balance between accuracy and model stability, proving to be robust across various scenarios and capable of handling complex, non-linear data structures effectively.

## Recommendations:

Based on the findings from this study, the following recommendations are proposed:

1. **Model Selection**: For datasets characterized by non-linear relationships and high complexity, Random Forest is recommended due to its robust performance and generalizability. However, for simpler or well-defined linear relationships, Linear Regression can provide transparency and ease of interpretation.
2. **Data Preprocessing**: Continued emphasis should be placed on thorough data preprocessing, including the handling of missing values, normalization of data, and encoding of categorical variables. This foundational work is crucial for the successful application of any machine learning model.
3. **Feature Selection and Engineering**: Invest efforts in feature selection and engineering to enhance model performance. Identifying and engineering features that directly influence the target variable can lead to more accurate and efficient models.
4. **Avoid Overfitting**: Particularly with Decision Trees and complex models, implement strategies to prevent overfitting, such as setting maximum depth for trees, increasing the number of trees in Random Forests, and using cross-validation techniques.

# Future Work:

To build on the findings of this study, the following areas for future work are identified:

1. **Incorporating Additional Data**: Expanding the dataset to include more variables that might affect the target variable could provide deeper insights and improve model accuracy.
2. **Advanced Modelling Techniques**: Exploring advanced machine learning techniques such as Gradient Boosting Machines (GBMs) or neural networks could offer improvements over the models tested, especially in handling non-linear and complex data patterns.
3. **Real-time Data Processing**: Investigating models that can adapt to real-time data and offer dynamic predictions could be highly beneficial, especially in rapidly changing environments like finance or social media analytics.
4. **Hybrid Models**: Combining the strengths of different models through techniques such as stacking or blending might yield better performance than any single model, especially in complex predictive tasks.
5. **Longitudinal Study**: Conducting a longitudinal study to understand the impact of temporal dynamics on model performance, which could be particularly relevant for industries like retail or seasonal services.

# References and Links:

**References:**

1. https://stackoverflow.com/questions/
2. Lectures Notes