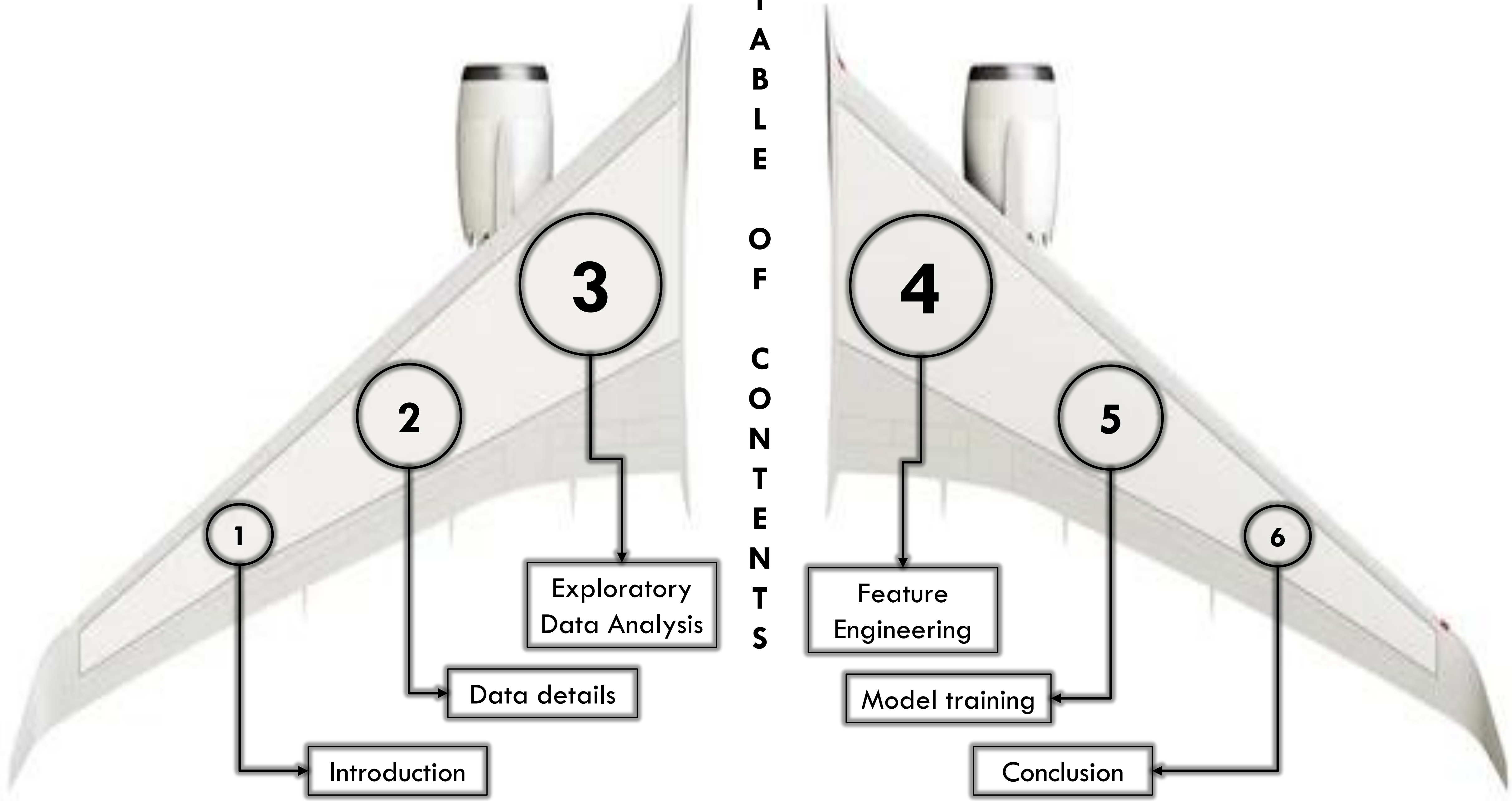
Airfare Trends and Ticket Pricing

- Predictive Modelling and Analysis

- Janardhan Reddy Illuru
- Advanced Data Science and AI

TABLE OF CONTENTS



1 Introduction

2 Data details

3 Exploratory Data Analysis

4 Feature Engineering

5 Model training

6 Conclusion

✈ 1. INTRODUCTION

✈ The aviation industry in India is experiencing unprecedented growth, with an increasing number of airlines offering extensive flight networks across the country. However, one of the most complex challenges in this sector is the dynamic nature of air ticket pricing. Ticket prices fluctuate widely due to various factors, including demand, timing, airline reputation, competition, and economic conditions. This volatility makes it difficult for both airlines and passengers to predict prices effectively.

✈ For airlines, accurate price prediction is essential for maximizing revenue, optimizing flight occupancy, and maintaining competitive edge. For passengers, understanding price trends can lead to more informed purchasing decisions, ensuring affordability and value. To address this challenge, our project focuses on developing a machine learning model that accurately predicts air ticket prices for domestic flights within India.

✈ By leveraging historical data from multiple airlines and flights, our model will identify and analyze the key factors influencing ticket prices. The insights derived from this model will empower stakeholders to make data-driven decisions, enhance pricing strategies, and ultimately contribute to a more efficient and customer-centric aviation market in India. This presentation will walk you through the process of building this model, from data collection to model evaluation, and highlight the potential impact on the industry.

✈ 2. DATA DETAILS

✈ The dataset consists of 10,683 rows and 11 columns, each representing a key aspect of flight details.

✈ **Airline:** Different airlines have different pricing strategies, so this feature will help in understanding the impact of the airline on ticket prices.

✈ **Date_of_Journey:** Ticket prices vary depending on the date of travel, influenced by factors like holidays, weekends, and seasonal trends.

✈ **Source:** The departure city can influence prices due to factors like airport taxes, demand, and competition.

✈ **Destination:** Similar to the source, the destination city/airport impacts the pricing based on demand, distance, and connectivity.

✈ **Route:** The specific route taken can affect pricing, especially if there are multiple stops or connections.

✈ **Dep_Time:** Departure time can influence demand, with prices varying based on whether a flight is during peak or off-peak hours.

✈ **Arrival_Time:** Arrival time might also affect demand and price, particularly for business travelers.

✈ **Duration:** Longer flights may be more expensive, depending on the distance and the number of stops.

✈ **Total_Stops:** Non-stop flights are generally more expensive than flights with stops.

✈ **Additional_Info:** This feature may include important details like meal availability, in-flight entertainment, or baggage policies that could affect pricing.

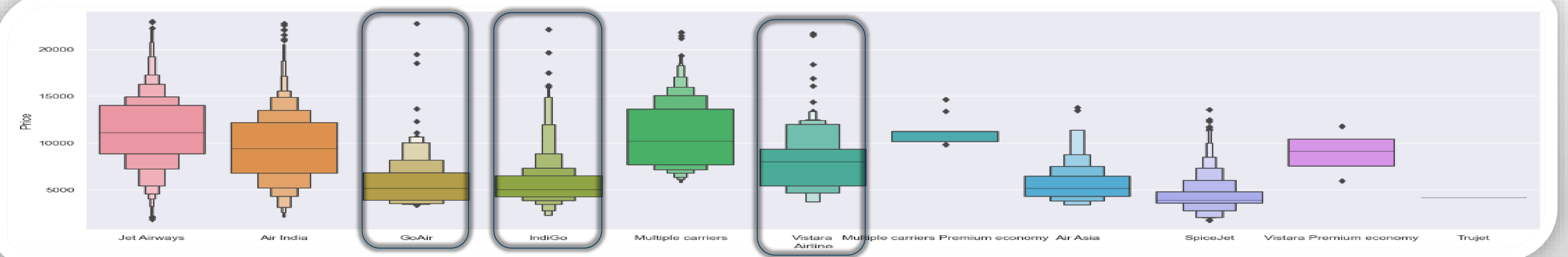
✈ **Price:** This is the target variable, which the model will predict.

3. DATA ANALYSIS

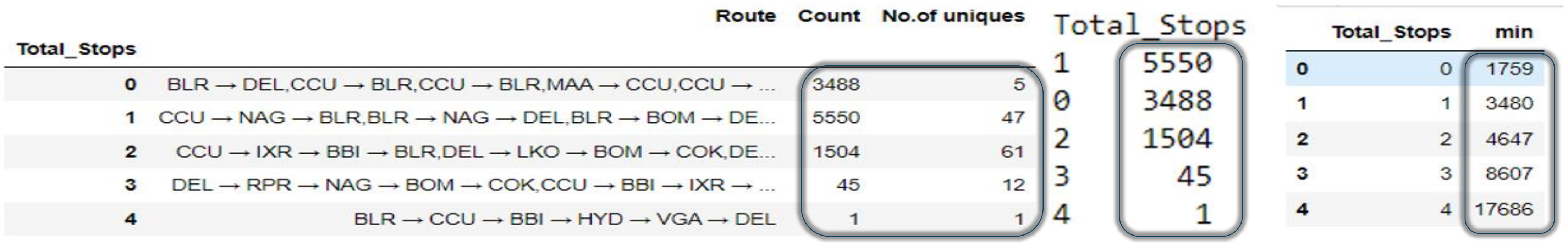
✈ **Null Values:** We identified a single missing entry in the dataset, so we removed it.

✈ **Outliers:** We identified 95 outliers, which account for **0.889%** of the dataset, and proceeded to remove them using the IQR method.

✈ **Airline:** We observed that Jet Airways, IndiGo, and Air India are the top three airlines preferred by passengers. In terms of pricing, GoAir, IndiGo, and Vistara Airlines exhibit significant fluctuations.



✈ **Route & Total_Stops :** We found that 'Route' and 'Total_stops' are closely related to each other. Additionally we observed that as the number of stops increases, the minimum fare also tends to increase.



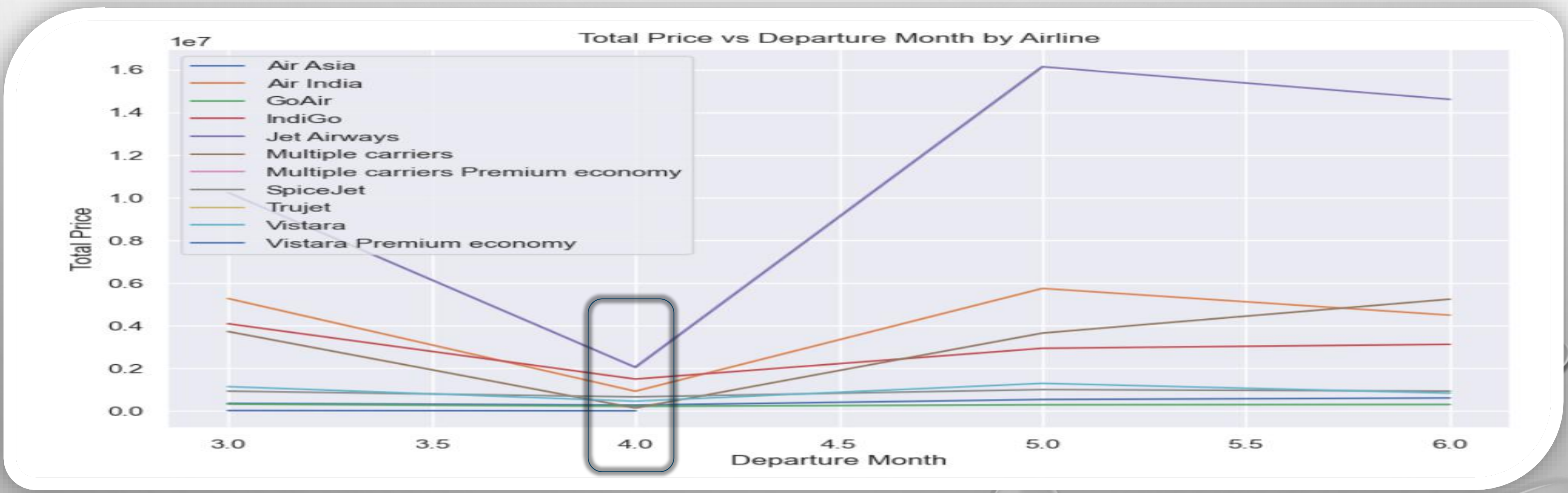
Total_Stops	Route	Count	No.of uniques
0	BLR → DEL,CCU → BLR,CCU → BLR,MAA → CCU,CCU → ...	3488	5
1	CCU → NAG → BLR,BLR → NAG → DEL,BLR → BOM → DE...	5550	47
2	CCU → IXR → BBI → BLR,DEL → LKO → BOM → COK,DE...	1504	61
3	DEL → RPR → NAG → BOM → COK,CCU → BBI → IXR → ...	45	12
4	BLR → CCU → BBI → HYD → VGA → DEL	1	1

```
Total_Stops
1    5550
0    3488
2    1504
3      45
4       1
```

	Total_Stops	min
0	0	1759
1	1	3480
2	2	4647
3	3	8607
4	4	17686

✈ 3. DATA ANALYSIS

- **Source:** There are a higher number of flights departing from Delhi, Kolkata, and Bangalore, ranking as the top three cities respectively.
- **Destination:** For destinations, the top three cities are Cochin, Bangalore, and Delhi, respectively.
- **Date_of_Journey:** We observed that Jet Airways, IndiGo, and Air India are the top three airlines preferred by passengers. In April, there was a decrease in revenue for each airline.



- **Additional_Info:** We found that most of the data lacks additional information.

4. FEATURE ENGINEERING

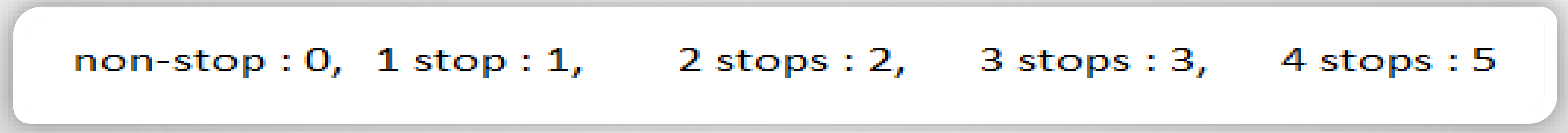
✈ **Airline:** Since ‘Airline’ is a nominal variable, we applied one-hot encoding to it.

✈ **Date_of_Journey:** We extracted the departure day and departure month from the date, while ignoring the departure year, as all data is from 2019

✈ **Source:** Since ‘Source’ is a nominal variable, we applied one-hot encoding to it.

✈ **Destination:** Since ‘Destination’ is a nominal variable, we applied one-hot encoding to it.

✈ **Route & Total_Stops:** We dropped the ‘Route’ feature and retained ‘Total_Stops’ due to their closely relationship. We used Label Encoding, assigning lower weights to non-stop flights and higher weights to flights with more stops.



non-stop : 0, 1 stop : 1, 2 stops : 2, 3 stops : 3, 4 stops : 5

✈ **Dep_Time:** We extracted the departure hours and minutes by modifying the data to include ‘0 hours’ for entries that only had minutes and ‘0 minutes’ for entries that only had hours.

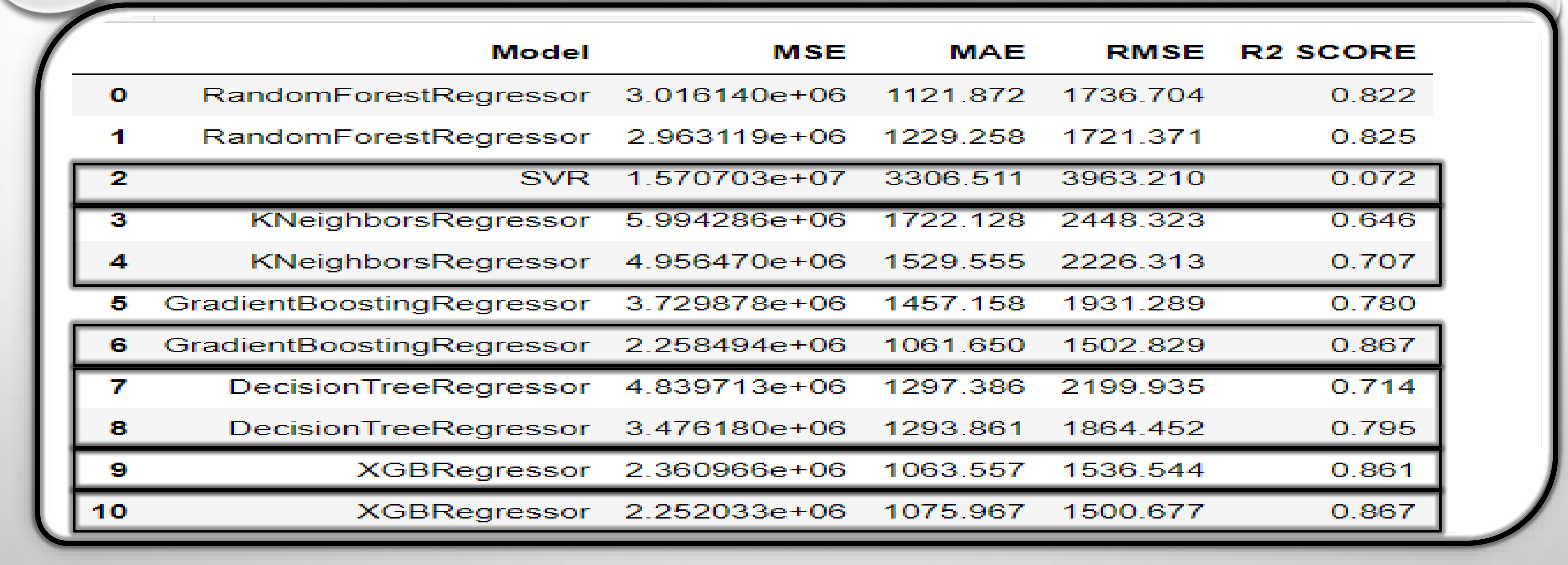
✈ **Arrival_Time:** We extracted the Arrival hours and minutes, assuming that, as domestic flights, the duration would typically be less than 24 hours. However, we found that 10% of the flights took more that 24 hours.

✈ **Duration:** We extracted the duration in hour and minutes.

✈ **Additional_Info:** We dropped this feature because 78.09% of the data in it was null.

✈ **Price:** This is the target variable, which the model will predict.

5. MODEL TRAINING



	Model	MSE	MAE	RMSE	R2 SCORE
0	RandomForestRegressor	3.016140e+06	1121.872	1736.704	0.822
1	RandomForestRegressor	2.963119e+06	1229.258	1721.371	0.825
2	SVR	1.570703e+07	3306.511	3963.210	0.072
3	KNeighborsRegressor	5.994286e+06	1722.128	2448.323	0.646
4	KNeighborsRegressor	4.956470e+06	1529.555	2226.313	0.707
5	GradientBoostingRegressor	3.729878e+06	1457.158	1931.289	0.780
6	GradientBoostingRegressor	2.258494e+06	1061.650	1502.829	0.867
7	DecisionTreeRegressor	4.839713e+06	1297.386	2199.935	0.714
8	DecisionTreeRegressor	3.476180e+06	1293.861	1864.452	0.795
9	XGBRegressor	2.360966e+06	1063.557	1536.544	0.861
10	XGBRegressor	2.252033e+06	1075.967	1500.677	0.867

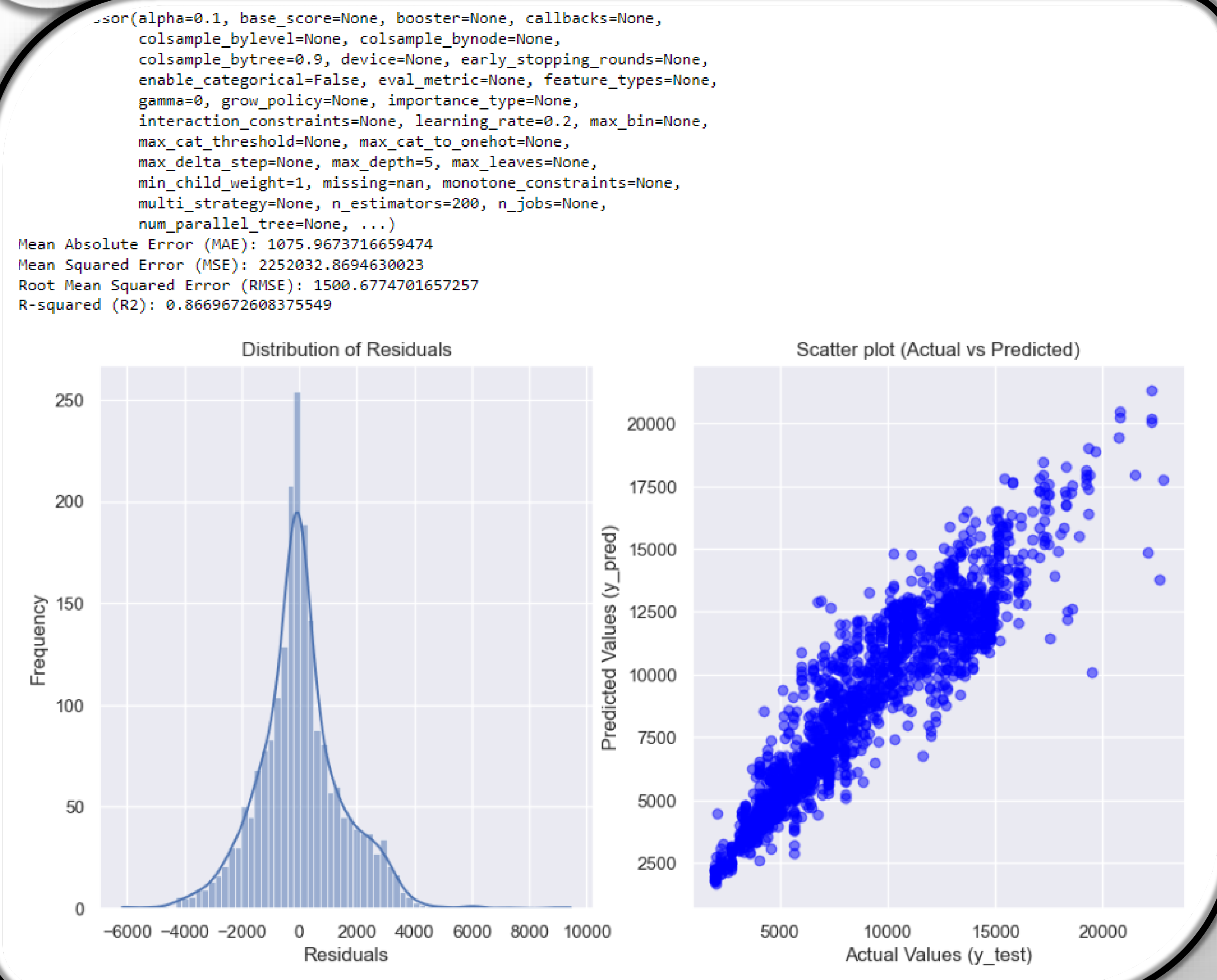
✈ **Tuned GradientBoostingRegressor** and **Tuned XGBRegressor** are the top performers with an R² score of **0.867** and low RMSE values. These models provide the best balance between prediction accuracy and generalization.

✈ **XGBRegressor (Default)** is very close to the tuned version and also performs well, with an R² score of **0.861.**

✈ **DecisionTreeRegressor** and **KNeighborsRegressor** showed lower R² scores and higher RMSEs, indicating possible overfitting.

✈ **SVR** was the least effective model, with an R² score of **0.072**.

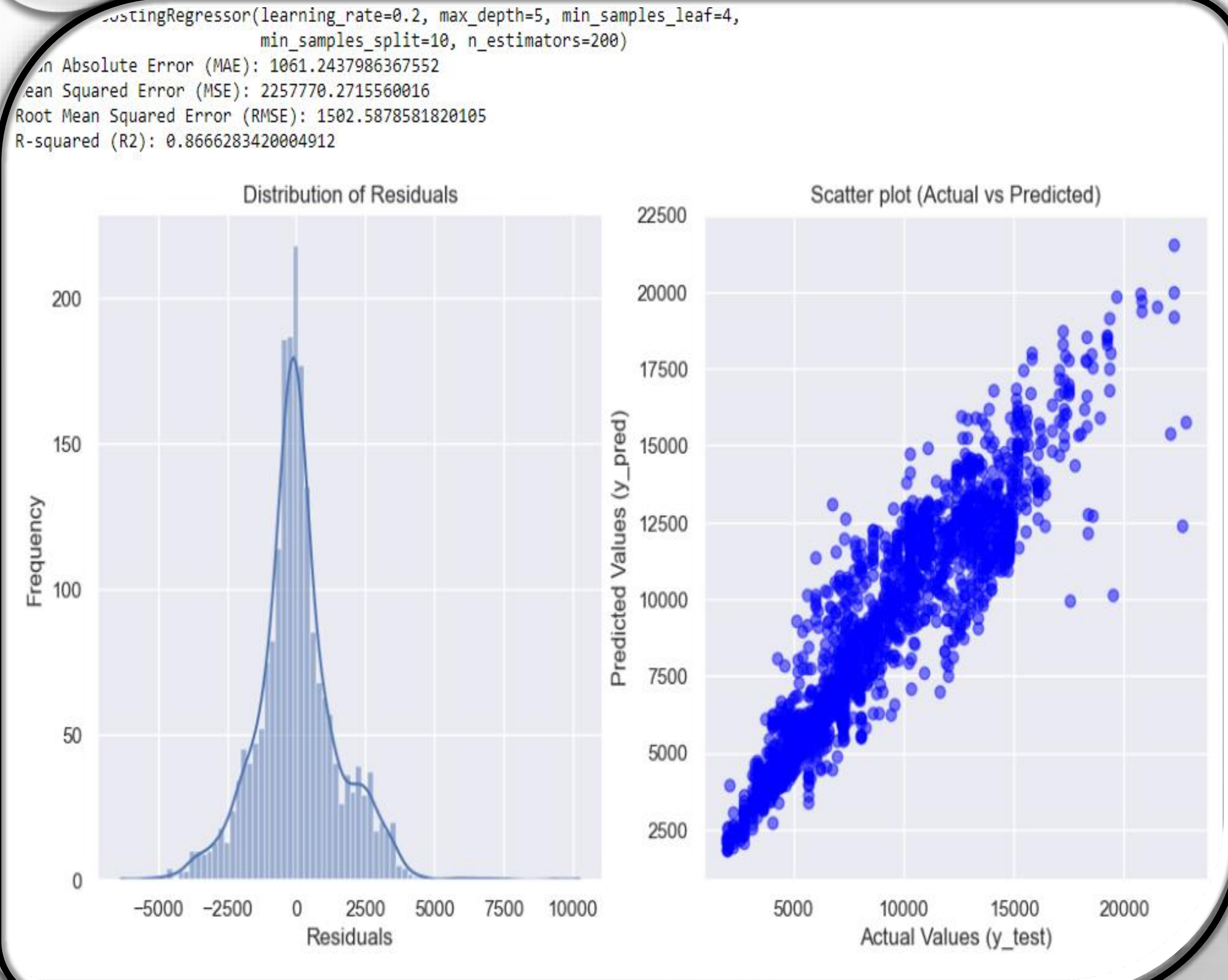
6. CONCLUSION



```
sor(alpha=0.1, base_score=None, booster=None, callbacks=None,
             colsample_bylevel=None, colsample_bynode=None,
             colsample_bytree=0.9, device=None, early_stopping_rounds=None,
             enable_categorical=False, eval_metric=None, feature_types=None,
             gamma=0, grow_policy=None, importance_type=None,
             interaction_constraints=None, learning_rate=0.2, max_bin=None,
             max_cat_threshold=None, max_cat_to_onehot=None,
             max_delta_step=None, max_depth=5, max_leaves=None,
             min_child_weight=1, missing=nan, monotone_constraints=None,
             multi_strategy=None, n_estimators=200, n_jobs=None,
             num_parallel_tree=None, ...)
Mean Absolute Error (MAE): 1075.9673716659474
Mean Squared Error (MSE): 2252032.8694630023
Root Mean Squared Error (RMSE): 1500.6774701657257
R-squared (R2): 0.8669672608375549
```

- **Tuned XGBRegressor** is recommended for the best overall performance.

✈ 6. CONCLUSION



```
…ostingRegressor(learning_rate=0.2, max_depth=5, min_samples_leaf=4,
                          min_samples_split=10, n_estimators=200)
…n Absolute Error (MAE): 1061.2437986367552
…ean Squared Error (MSE): 2257770.2715560016
Root Mean Squared Error (RMSE): 1502.5878581820105
R-squared (R2): 0.8666283420004912
```

✈ **Tuned GradientBoostingRegressor** is also highly recommended.

✈ These models demonstrate both strong prediction accuracy and good generalization, making them ideal choices for your problem.

Thank You