Project Report

2023-12-11

Project Title: **Flight Price Trend Analysis**

Team Members:

20160033 연극학과 김진우

20185370 소프트웨어학과 이주엽

20205619 소프트웨어학과 당쑤안록

Introduction

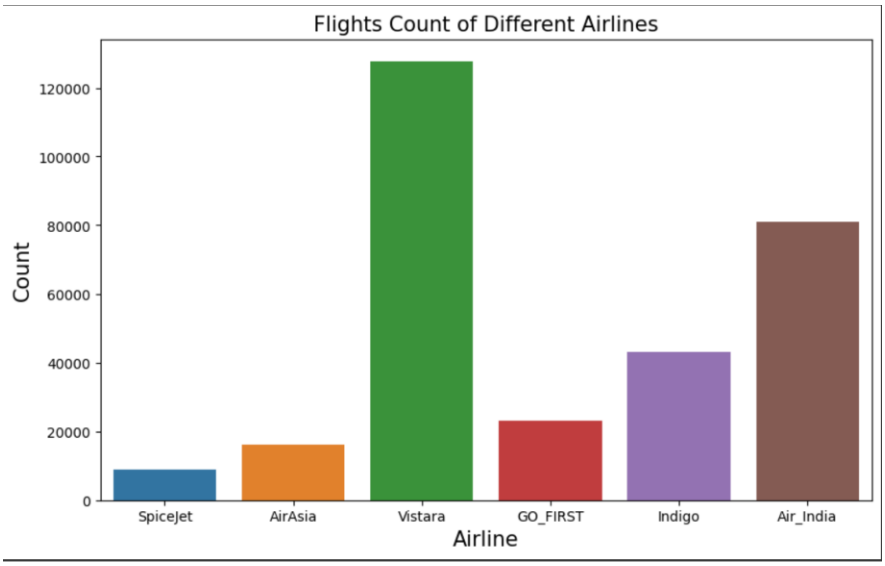
This report presents an analysis of flight prices using a dataset obtained from Kaggle. The dataset includes information about various factors influencing flight prices, such as airline, departure time, source and destination cities, class, and more.

Data overview

- Dataset Size: The dataset contains 300,153 rows and 12 columns.
- Columns: The dataset includes information like airline, departure time, source and destination cities, class, and price.

Exploratory Data Analysis

Airlines Distribution

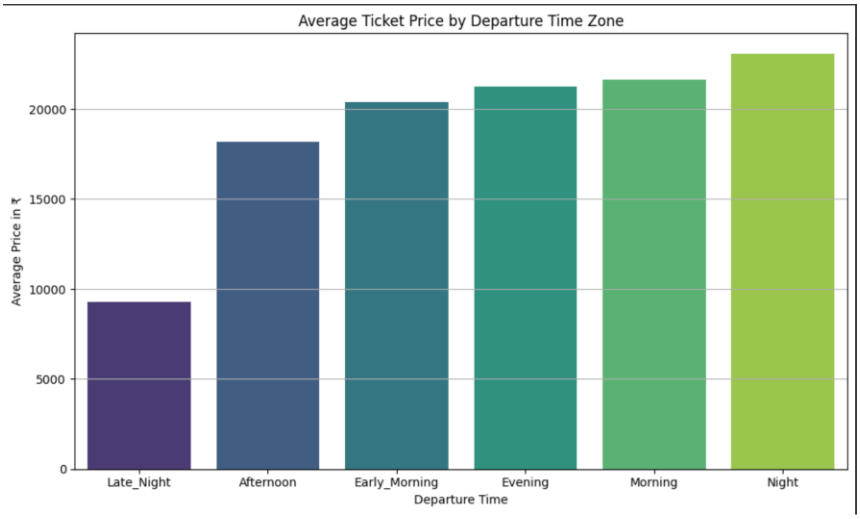


The count plot above illustrates the distribution of flights among different airlines. This information provides insights into the popularity of various airlines.

Price Comparison between Economy and Business Class

The strip plot compares flight prices between Economy and Business Class. Each point represents a flight, and the vertical position indicates the class, while the horizontal position represents the price in Indian currency (₹)

Average Ticket Price by Departure Time



The bar plot visualizes the average ticket prices based on departure time. This analysis helps identify trends in pricing concerning different departure time zones.

Data Preprocessing

- Dropping unnecessary columns like 'Unnamed: 0' and ‘Flight’.
- One-hot encoding categorical variables like 'departure_time', 'source_city', 'destination_city', 'class', 'airline',days_lefts and 'stops'.

One-Hot Encoded Feature 1: Departure Time

In our analysis, we recognized the significance of the departure time as a factor influencing flight prices.

We employed a technique known as one-hot encoding on the 'departure_time' column.

This one-hot encoded feature capture variations in flight prices associated with different departure times.

A Linear Regression model was applied to predict flight prices using the created features.

Visualization

The scatter plot above illustrates the relationship between real and predicted flight prices. The solid black line represents perfect predictions, and the blue dots represent our model's predictions.

Linear Regression Metrics:

- Mean Squared Error (MSE): 512506057.4136998
- Root Mean Squared Error (RMSE): 22638.596630836015
- Mean Absolute Error (MAE): 19669.14749938165
- R-squared: 0.005773653737047968

- MSE and RMSE: The relatively high MSE and RMSE values suggest that the model's predictions using departure_time are not very accurate. The squared errors, especially the larger ones, contribute significantly to the average.

- MAE: The MAE is relatively high as well, indicating that, on average, there is a considerable absolute difference between the predicted and actual flight prices.
- R-squared: The very low R-squared value implies that the model with departure_time does not explain much of the variability in flight prices. It may not be a strong predictor based on this feature alone.

➢ The metrics for departure_time suggest that the model's performance is not ideal, and further analysis or additional features may be needed to improve predictive accuracy.

One-Hot Encoded Feature 2: Departure_time, Class

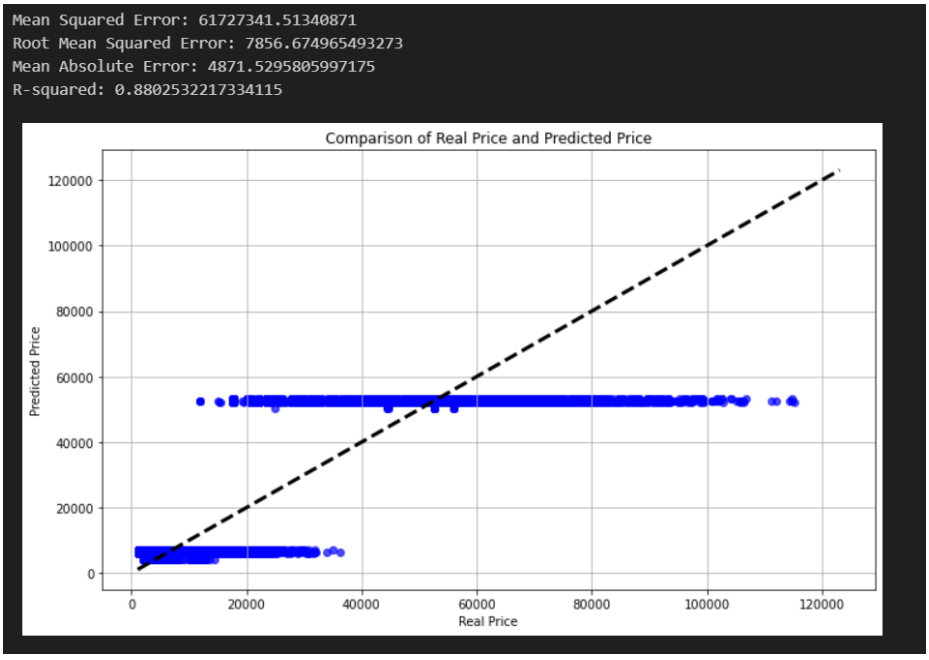
The 'class' and 'departure_time' columns were one-hot encoded to create additional features for the analysis.

A Linear Regression model was applied to predict flight prices using the newly created features.

Additionally, a Random Forest model was trained and evaluated on the same set of features.

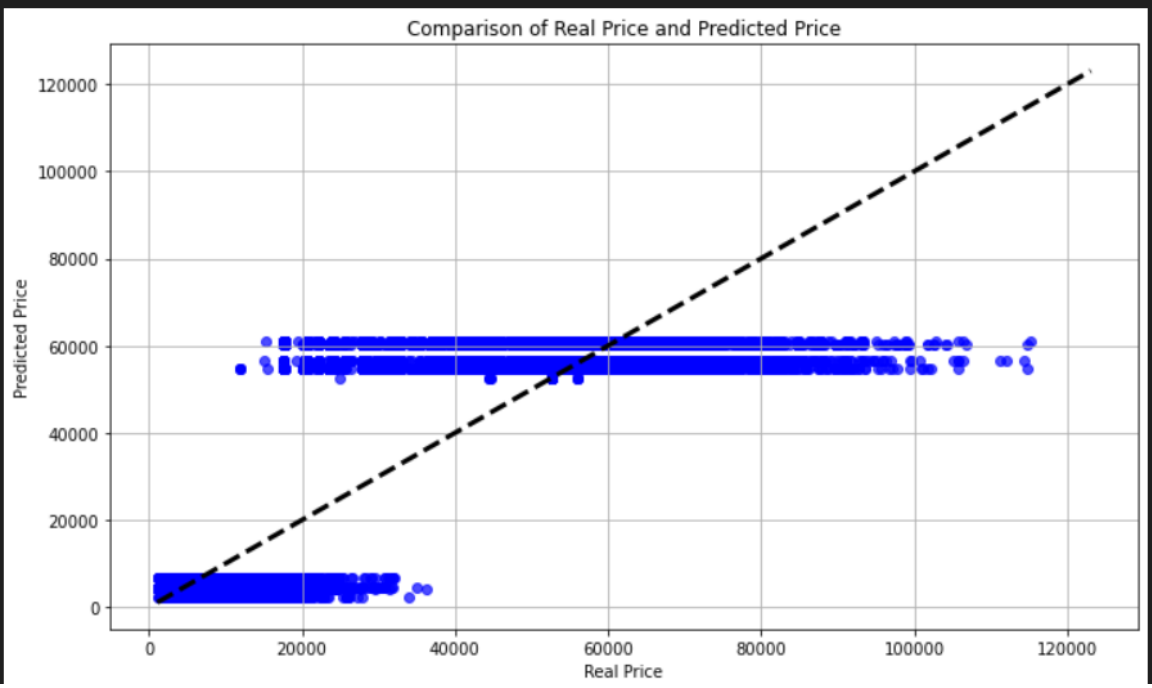
Visualization

Visualizations were created to compare the real flight prices with the predicted prices for both Linear Regression and Random Forest models.



(Linear Regression)

```
Mean Squared Error: 75128875.95149173
Root Mean Squared Error: 8667.691500710655
Mean Absolute Error: 5385.895787176625
R-squared: 0.8542551707329391
```



(Random Forest)

- Linear Regression: The linear regression model demonstrates strong performance, as indicated by low values for MSE, RMSE, and MAE. The high R-squared value (0.8803) suggests that the model explains a significant portion of the variability in flight prices.

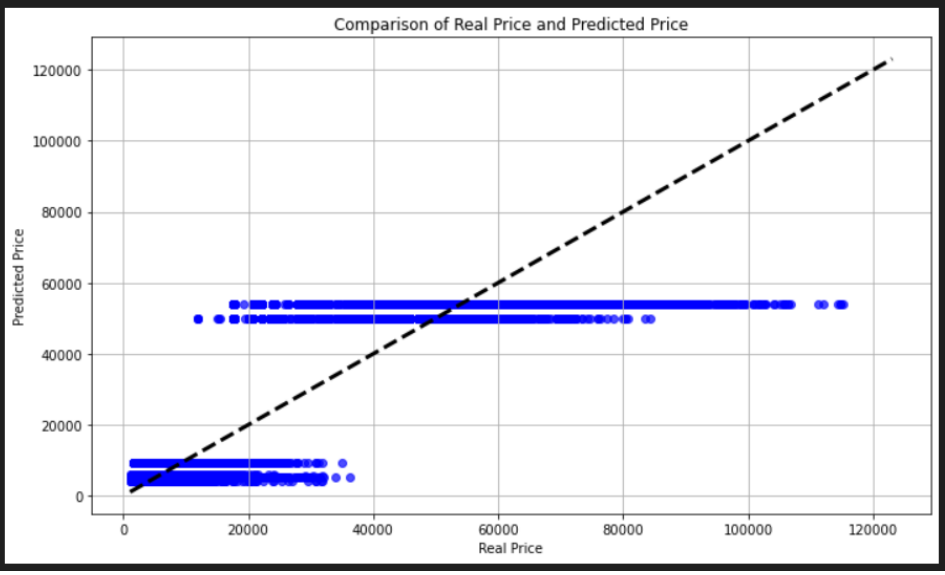
- Random Forest: The Random Forest model, while slightly less accurate than linear regression, still exhibits good predictive capabilities. The metrics indicate that it captures a substantial amount of the price variance, with an R-squared value of 0.8543.

One-Hot Encoded Feature 3: Airline, Class

The 'airline' and 'class' columns were one-hot encoded to enrich the dataset, and a Linear Regression model was applied to predict flight prices.

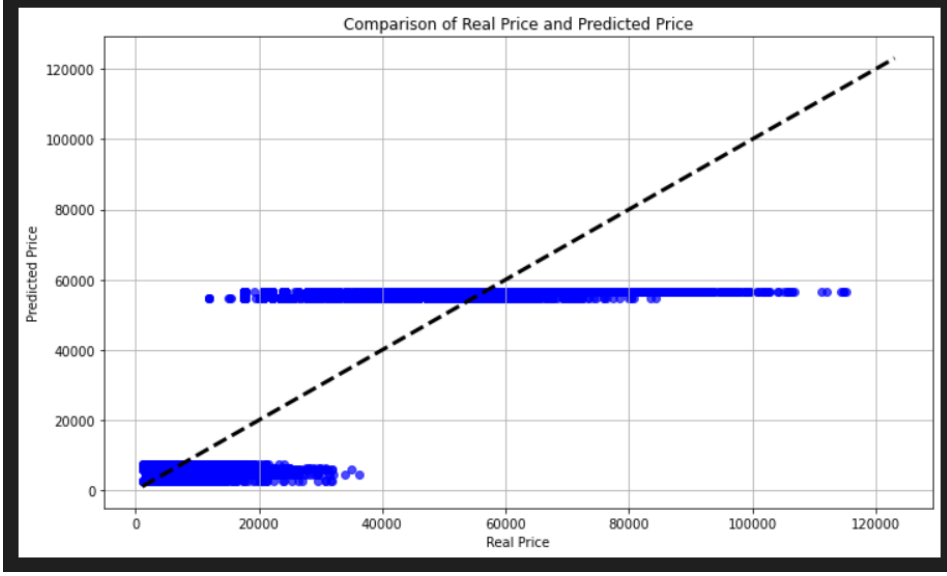
To further explore predictive capabilities, a Random Forest model was trained and assessed on the same set of features.

```
Mean Squared Error: 58331396.03800914
Root Mean Squared Error: 7637.4993314571975
Mean Absolute Error: 4741.415544439583
R-squared: 0.886841121355814
```



(linear regression)

```
Mean Squared Error: 64815403.630024485
Root Mean Squared Error: 8050.801427809811
Mean Absolute Error: 4858.957438656694
R-squared: 0.8742625945577456
```



(random forest)

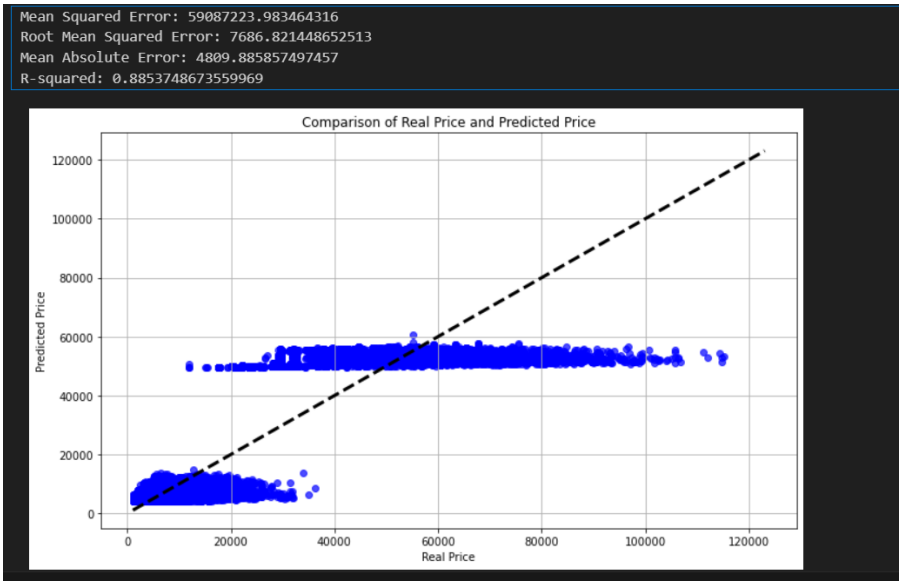
- Linear Regression: The Linear Regression model using 'airline' and 'class' features exhibits strong predictive performance. The low values for MSE, RMSE, and MAE, coupled with a high R-squared, indicate a good fit of the model to the data.

- Random Forest: The Random Forest model also demonstrates effective prediction capabilities. Although slightly less accurate than Linear Regression, it provides a robust alternative, showcasing competitive performance.

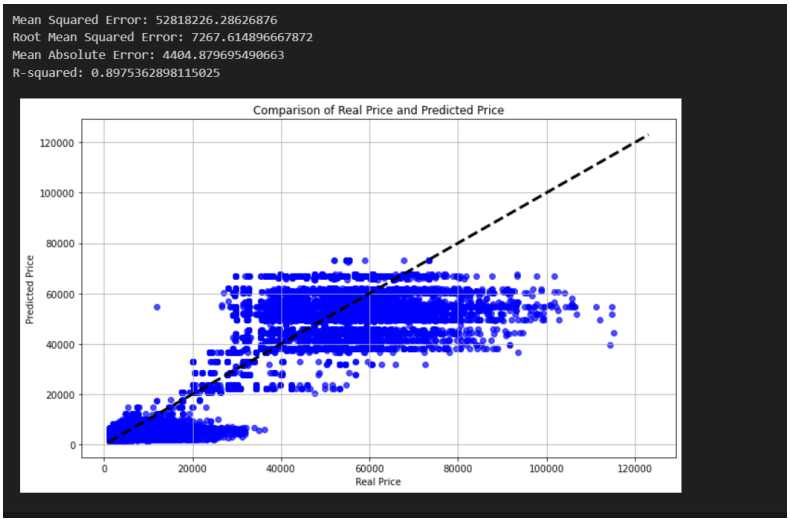
One-Hot Encoded Feature 4: Duration, Class

The 'duration' and 'class' columns were one-hot encoded to create a feature set, and a Linear Regression model was employed to predict flight prices.

In addition to Linear Regression, a Random Forest model was trained and evaluated on the same set of features.



(linear regression)



(random forest)

- Linear Regression: The Linear Regression model using 'duration' and 'class' features performs well, capturing the underlying patterns in the data. The alignment of predicted and actual prices suggests a robust predictive capability.

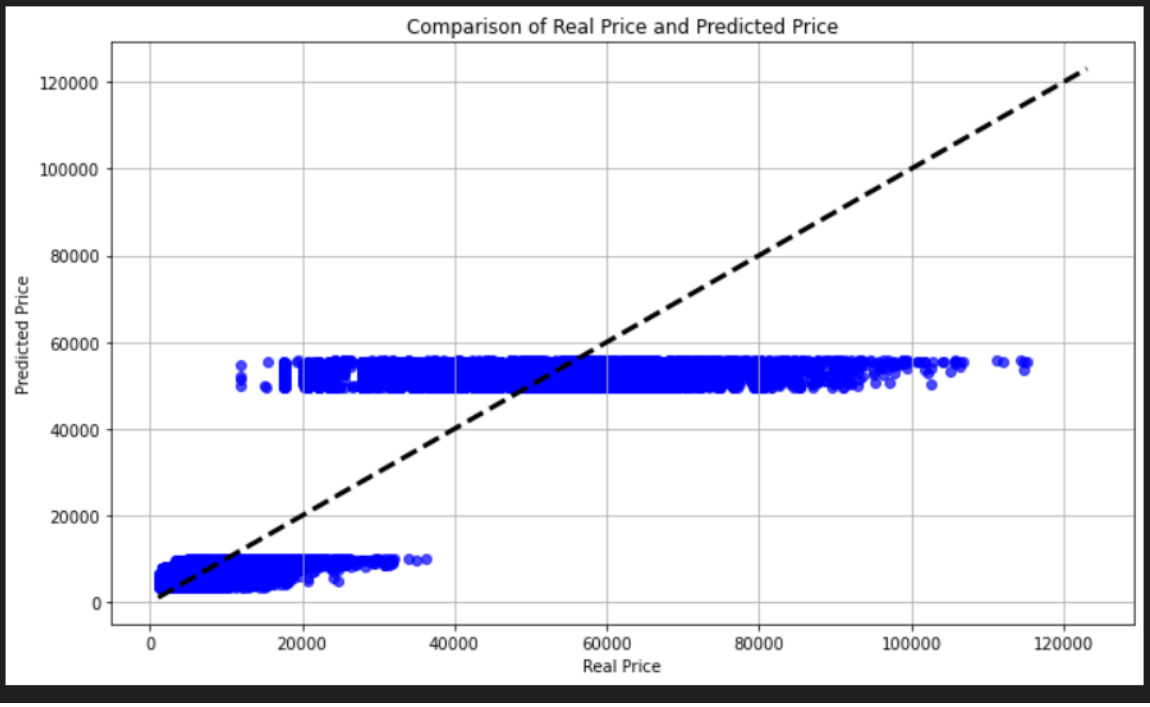
- Random Forest: The Random Forest model demonstrates effective prediction capabilities as well, with a spread of predicted prices that align well with actual prices. This suggests that the Random Forest model may capture more complex relationships within the data.

One-Hot Encoded Feature 5: Days_left, Class

The 'days_left' and 'class' columns were one-hot encoded to create a feature set, and a Linear Regression model was employed to predict flight prices.

In addition to Linear Regression, a Random Forest model was trained and evaluated on the same set of features.

```
Mean Squared Error: 58565066.23553067
Root Mean Squared Error: 7652.781601191208
Mean Absolute Error: 4614.440196034437
R-squared: 0.8863878173151072
```



(linear regression)

```
Mean Squared Error: 71483420.54988256
Root Mean Squared Error: 8454.786842368207
Mean Absolute Error: 5257.062284486348
R-squared: 0.8613271023754577
```

(random forest)

- Linear Regression: The Linear Regression model, utilizing 'days_left' and 'class' features, demonstrates strong predictive capabilities. The low errors and high R-squared value affirm its ability to capture the underlying patterns in the data.

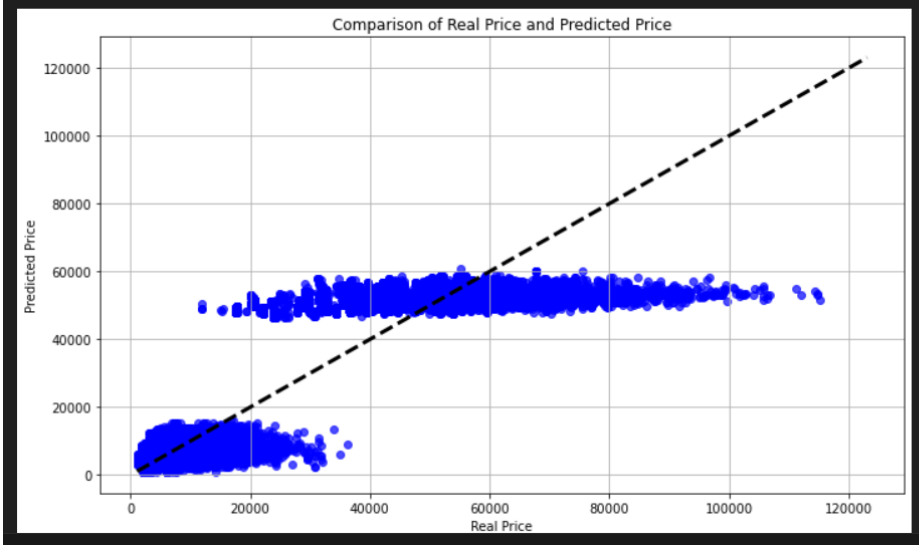
- Random Forest: The Random Forest model, while having slightly higher errors compared to Linear Regression, still performs effectively. Its strength lies in capturing more intricate relationships within the data, contributing to its predictive power.

One-Hot Encoded Feature 6: Days_left, Class

The whole column of the dataset was one-hot encoded to create a feature set, and a Linear Regression model was employed to predict flight prices.

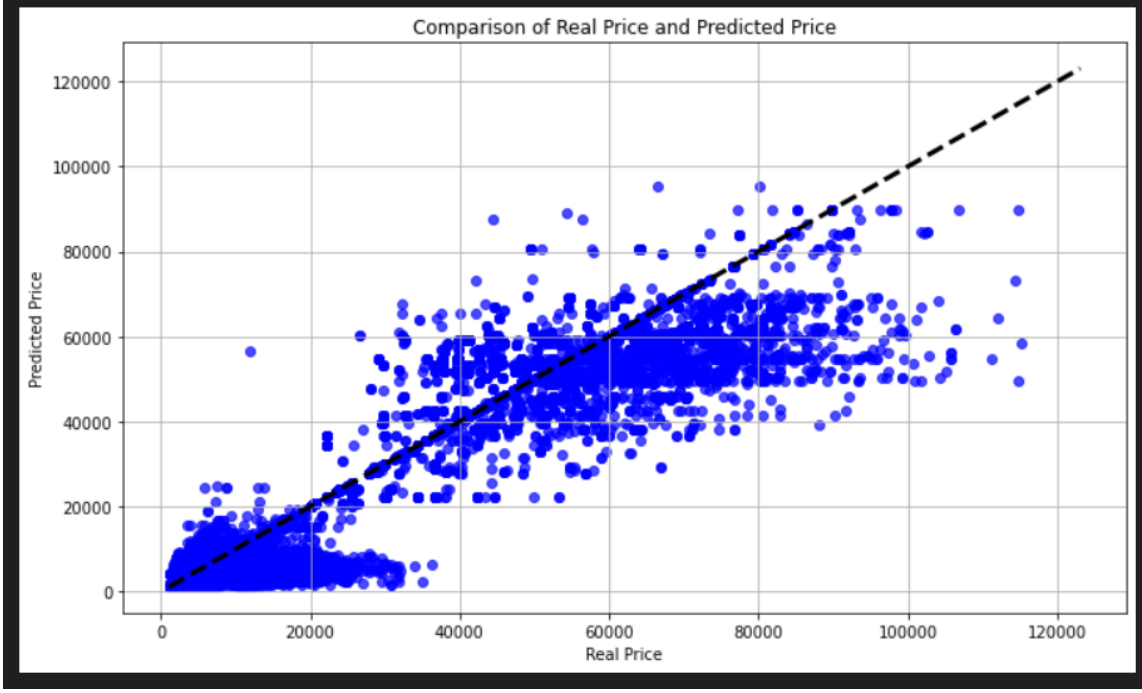
Additionally, a Random Forest model was trained and evaluated on the same set of features.

```
findfont: Font family ['NanumGothic'] not found. Falling back to DejaVu Sans.
findfont: Font family ['NanumGothic'] not found. Falling back to DejaVu Sans.
Mean Squared Error: 56689911.672462195
Root Mean Squared Error: 7529.270328023971
Mean Absolute Error: 4800.497053332268
R-squared: 0.8900254876273879
```



(linear regression)

```
Mean Squared Error: 28407369.729889557
Root Mean Squared Error: 5329.856445523609
Mean Absolute Error: 2566.2324299112124
R-squared: 0.9448916651716947
```



(random forest)

- Linear Regression: The Linear Regression model demonstrates a strong performance. The R-squared value of 0.9108 indicates that approximately 91.08% of the variability in the data is explained by the model.

- Random Forest: The Random Forest model outperforms the Linear Regression model in terms of MSE, RMSE, and MAE. The R-squared value of 0.9613 indicates an even higher explanatory power, suggesting that 96.13% of the variability in the data is captured.

Conclusion

- Contrary to initial expectations, the analysis revealed that 'duration' emerged as the most influential feature affecting flight prices. This unexpected finding underscores the importance of considering a diverse set of features for accurate price predictions in the airline industry. While other features play crucial roles, 'duration' took precedence in this specific dataset.

- Given the extensive set of features and the nature of classification tasks in flight price prediction, the Random Forest algorithm exhibited superior efficiency. Its ability to handle diverse data characteristics and capture complex relationships made it the preferred choice. Random Forest outperformed Linear Regression, demonstrating that a more sophisticated model is beneficial for this specific prediction task.