FLIGHT DELAY ANALYSIS REPORT

ADAM KAINIKARA

Goal

The goal of this analysis is to explore patterns in flight delays using historical flight data and to build machine learning and statistical models to classify delayed flights and predict arrival delays. This analysis also includes visualizations to support findings.

Data and Methods

I used a dataset containing flight information including departure delay, distance, air time, and arrival delay. I found the dataset as a CSV publicly available on Kaggle. The code performs the following steps:

- Loads and cleans data, skipping any rows with missing values
- Computes correlation between distance and arrival delay
- Classifies flights as *on-time* or *delayed* (threshold: 15 minutes) using a Random Forest classifier
- Predicts actual arrival delays using Linear Regression and Random Forest
- Generates visualizations for distribution of delays, classification proportions, regression errors, and feature importance

Results

Correlation. The correlation between distance and arrival delay is low:

```
corr distance vs arr_delay:  0.022
```

This indicates that distance alone is not a strong predictor of delays.

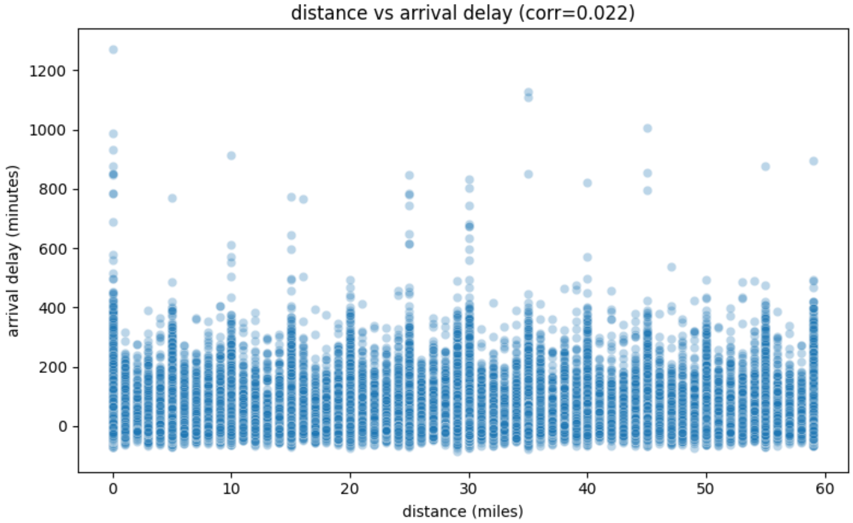

Figure 0.1. Scatter plot of distance vs arrival delay. Low correlation shows distance alone is not predictive.

Date: September 15, 2025.

Classification. The Random Forest classifier achieved the following results:

- Accuracy: 88.7%
- Precision/Recall: higher for on-time flights than delayed flights

This reflects the class imbalance in the dataset (more on-time than delayed flights).

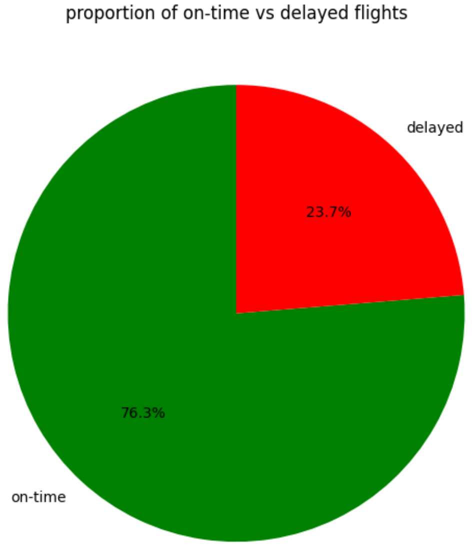

FIGURE 0.2. Proportion of on-time vs delayed flights. Highlights the class imbalance in the dataset.

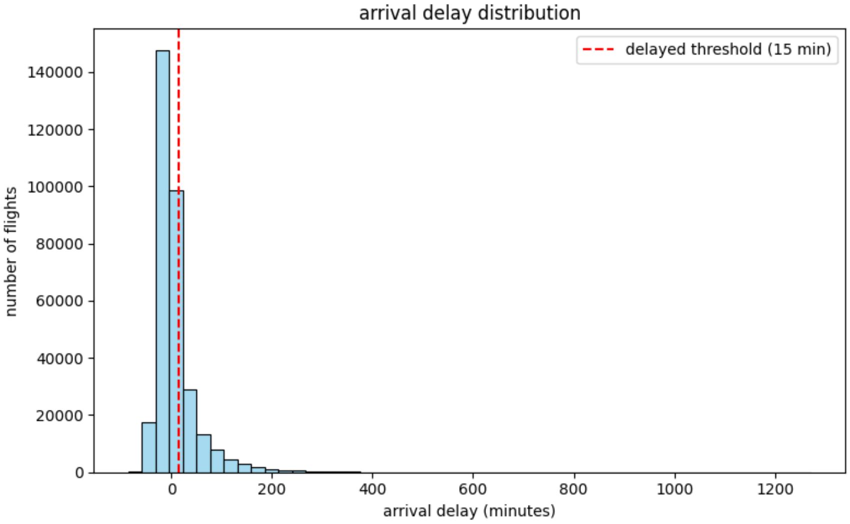

FIGURE 0.3. Histogram of arrival delays with the 15-minute delayed threshold. Most flights are on-time or slightly delayed.

Regression. Linear Regression and Random Forest regression were used to predict arrival delays:

- Linear Regression RMSE: 18.0 minutes
- Random Forest Regression RMSE: 19.1 minutes

Prediction errors are concentrated around zero, indicating reasonably accurate estimates for most flights.

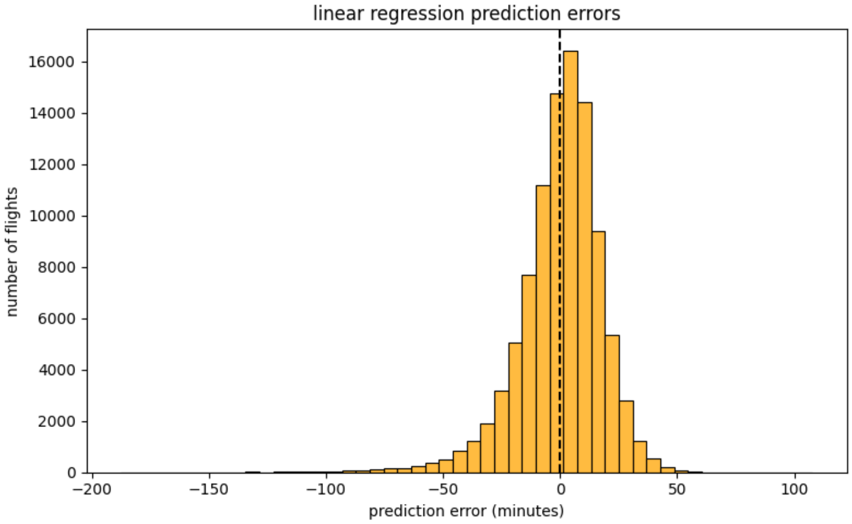

Figure 0.4. Histogram of linear regression prediction errors. Most errors are within plus/minus 20 minutes.

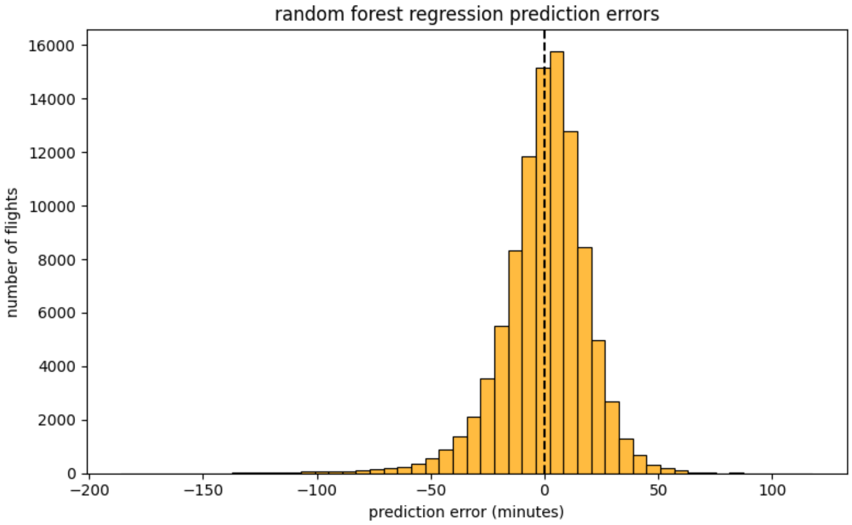

Figure 0.5. Histogram of random forest regression errors. Slightly wider spread than linear regression.

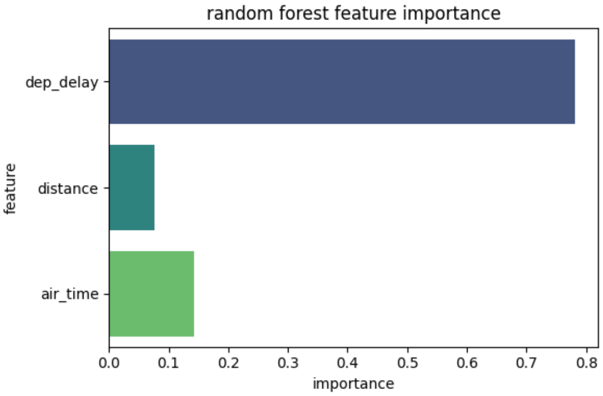

Figure 0.6. Random Forest feature importance. Departure delay is the dominant factor influencing whether a flight will be delayed, while distance and air time have smaller contributions.

Feature Importance.

Conclusion

The analysis shows that while distance has minimal effect on arrival delays, departure delay and air time are more predictive. Classification models can reasonably separate on-time and delayed flights, though performance for delayed flights is lower due to class imbalance. Regression models estimate arrival delay with errors around 18–19 minutes. The visualizations provide clear insights into delay distributions, prediction errors, and feature importance.