Inah Lee
Data Analytics
4/20/2025

Flight Delay Pattern Project:
Predicting Flight Delays Using Data-Driven Modeling

The objective of this project is to predict flight delays, a problem that extends beyond minor inconveniences to cause significant disruptions for passengers and considerable time and cost losses for airlines and airports. According to the U.S. Federal Aviation Administration (FAA), the annual cost of flight delays is approximately 33 billion dollars. If potential delays can be predicted in advance, passengers may select alternative flights, and airlines and airports can improve operational efficiency through proactive responses or schedule adjustments.

In this project, a binary classification model was developed to predict whether a flight would be delayed by more than 15 minutes. The target variable, is_delayed, was defined as 1 if the arrival delay exceeded 15 minutes, and 0 otherwise. The initial hypothesis was that flight delays follow identifiable patterns based on factors such as departure time, airline, and flight distance. Based on this hypothesis, data visualization, analysis of relationships among key variables, and predictive modeling were conducted to demonstrate the practical value of data-driven decision-making.

The dataset used in this project is the Flight Delay and Cancellation Dataset (2019–2023) provided by Kaggle. This large-scale dataset, in CSV format, covers U.S. domestic flights with approximately 3 million records. It includes various features such as scheduled and actual departure and arrival times, airline codes, flight distances, delay durations, and cancellation information.

Key variables:

CRS_DEP_TIME: Scheduled departure time

ARR_DELAY: Arrival delay time

AIRLINE: Airline name

DISTANCE: Flight Distance

CANCELED, DIVERTED: Cancellation and diversion

ORIGIN, DEST: Departure and Destination Airport

I prepared the dataset for modeling with the following steps.

- `filter(CANCELLED == 0, DIVERTED == 0) %>%`
  Removed flights that were cancelled or diverted.

- `drop_na(ARR_DELAY, CRS_DEP_TIME, DISTANCE)`
  Dropped rows with missing values in key fields.

- `mutate(is_delayed = ifelse(ARR_DELAY > 15, 1, 0))`
  Created the binary target variable.

- `model_data <- balanced_flights %>%`
  `select(is_delayed, AIRLINE, CRS_DEP_TIME, DISTANCE)`
  Selected features for modeling and removed high cardinality fields.

Rationale

- Cancelling and diverting indicate non comparable operational outcomes, so those records were excluded.
- The target variable is_delayed captures the FAA threshold for a meaningful delay, defined as arrival delay greater than 15 minutes.
- High cardinality location fields can inflate model complexity without proportional performance gains, so a lean feature set was retained.

Exploratory Patterns

To inspect delay patterns by carrier, I visualized the class proportions for the top airlines.

```
flights %>%
  filter(AIRLINE %in% top_airlines) %>%
  ggplot(aes(x = AIRLINE, fill = as.factor(is_delayed))) +
  geom_bar(position = "fill") +
  labs(title = "Top 5 Airlines by Delay Rate", y = "Proportion", fill = "Delayed") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

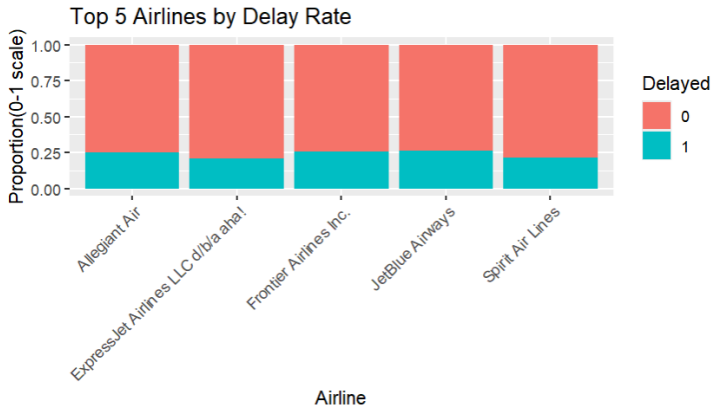

Figure 1. EDA delay rate by airlines

Figure 1 illustrates that Allegiant Air, ExpressJet Airlines LLC d/b/a aha!, Frontier Airlines, JetBlue Airways, and Spirit Airlines exhibit relatively high delay rates. Because the dataset contains a large number of airlines, including all of them in a single visualization made interpretation difficult. Therefore, only the top five airlines with the highest delay rates were selected for clearer comparison and analysis.

```
flights <- flights %>%
  mutate(dep_hour = floor(CRS_DEP_TIME / 100))
ggplot(flights, aes(x = dep_hour, fill = as.factor(is_delayed))) +
  geom_bar(position = "fill") +
  labs(title = "Departure Hour vs Delay Rate", x = "Scheduled Hour", y =
"Proportion", fill = "Delayed")
```

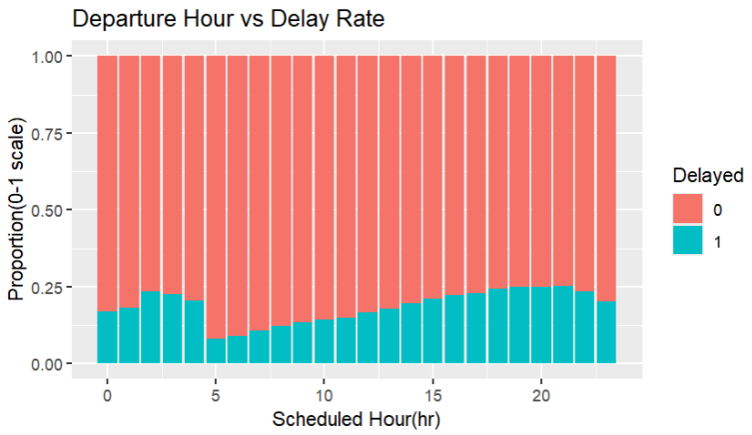

Figure 2. EDA delay rate by departure time

Visualization for delay rate by departure time zone:
Figure 2 presents the delay rate by departure time zone, derived by grouping scheduled departure times (CRS_DEP_TIME) into hourly intervals. The results show a clear variation in delay rates across different time zones. Flights departing early in the morning (around 00:00) exhibited the lowest delay rates, while a gradual increase was observed from the afternoon through the evening (up to 15:00). This pattern may be attributed to cumulative delays building throughout the day, increased air traffic congestion, and the cascading effect of arrival delays from earlier flights.

These findings indicate that departure time has a strong correlation with the likelihood of flight delays and is therefore an important feature in the prediction model.

Correlation Analysis and Feature Selection

The Pearson correlation coefficients between key variables were calculated using the Cor() function from the DescTools package. The variables examined were CRS_DEP_TIME, DISTANCE, and ARR_DELAY.

```
cor_data <- flights %>%
  select(CRS_DEP_TIME, DISTANCE, ARR_DELAY)
cor_matrix <- Cor(cor_data, method = "pearson", use = "complete.obs")
print(cor_matrix)
```

```
              CRS_DEP_TIME     DISTANCE   ARR_DELAY
CRS_DEP_TIME   1.00000000 -0.007825040 0.089852946
DISTANCE      -0.00782504  1.000000000 0.001883837
ARR_DELAY      0.08985295  0.001883837 1.000000000
```

Figure 3. Correlation

Figure 3 shows that CRS_DEP_TIME and ARR_DELAY have a weak positive correlation ($r \approx 0.09$), indicating that later scheduled departures are slightly more likely to be delayed. DISTANCE and ARR_DELAY show virtually no correlation ($r \approx 0.002$), suggesting that flight distance alone does not meaningfully explain delay duration.

The Pearson correlation coefficient between key variables was calculated using the Cor() function from the DescTools package. The correlation between CRS_DEP_TIME (scheduled departure time) and ARR_DELAY (arrival delay) was approximately 0.0899, indicating a weak positive correlation and suggesting that the likelihood of a delay increases slightly for later departures. The correlation between DISTANCE and ARR_DELAY was only 0.0018, indicating virtually no relationship between flight distance and arrival delay. There was also no significant relationship between departure time and distance.

These findings suggest that departure time may provide some predictive value for delays, while distance contributes very little. When combined with earlier visualization results, airline and departure time emerged as variables with a more meaningful impact on flight delays. These insights informed the model design and feature selection process.

Modeling Approach

For the prediction of flight delays, two classification models were developed: a Random Forest classifier and a Logistic Regression model. The prediction target variable was is_delayed, defined as 1 if ARR_DELAY exceeded 15 minutes, and 0 otherwise, framing the task as a binary classification problem.

Data Preprocessing Summary

- Excluded cancelled (CANCELLED == 1) or diverted (DIVERTED == 1) flights.
- Removed rows with missing values for ARR_DELAY, CRS_DEP_TIME, or DISTANCE.
- Converted CRS_DEP_TIME to numeric format for modeling.
- Removed high-cardinality variables such as ORIGIN and DEST to prevent performance and complexity issues in the Random Forest model.
- Final modeling variables: AIRLINE, CRS_DEP_TIME, and DISTANCE, converted to appropriate numeric or categorical types.

```
model_data <- balanced_flights %>%
  select(is_delayed, AIRLINE, CRS_DEP_TIME, DISTANCE) %>%
  mutate(
    CRS_DEP_TIME = as.numeric(CRS_DEP_TIME),
    is_delayed = as.factor(is_delayed)
  )
```

Addressing Class Imbalance

Initially, using the raw dataset led to a class imbalance problem: the majority of flights were classified as "not delayed" (0). As a result, the model achieved a deceptively high accuracy of about 82% but performed poorly in detecting delayed flights.

To address this, a random undersampling technique was applied. A subset of non-delayed flights was randomly selected to match the number of delayed flights, creating a balanced 1:1 class ratio in the training data. This ensured that both classes were equally represented during model training.

```
      # Undersampling
      delayed <- flights %>% filter(is_delayed == 1)
      not_delayed <- flights %>% filter(is_delayed == 0) %>%
sample_n(nrow(delayed))
      balanced_flights <- bind_rows(delayed, not_delayed)
```

Random Forest Model Training and Evaluation

The dataset was split into 80 percent for training and 20 percent for testing. A Random Forest model with 100 trees was trained to predict the binary target variable is_delayed. Model performance was evaluated using the confusionMatrix() function from the caret package.

```
      # Train and test set
      set.seed(123)
      train_index <- createDataPartition(model_data$is_delayed, p = 0.8, list =
FALSE)
      train <- model_data[train_index, ]
      test <- model_data[-train_index, ]
```

```
# Model 1 RandomForest
rf_model <- randomForest(is_delayed ~ ., data = train, ntree = 100)
rf_pred <- predict(rf_model, newdata = test)
rf_result <- confusionMatrix(rf_pred, test$is_delayed)
print(rf_result)
```

```
> print(rf_result)
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 63124 42202
         1 39933 60855

               Accuracy : 0.6015
```

Figure 4. Confusion Matrix for Random Forest

Figure 4 shows the confusion matrix for the Random Forest model. The overall accuracy was approximately 60.15 percent. While this accuracy is above the baseline, it indicates that there is substantial room for improvement, particularly in correctly identifying delayed flights.

Variable Importance Analysis and Logistic Regression Model

Figure 5 presents the variable importance plot generated using the varImpPlot() function for the Random Forest model. The x-axis represents MeanDecreaseGini, which indicates each variable's contribution to model performance. CRS_DEP_TIME (scheduled departure time) emerged as the most influential variable for delay prediction. This aligns with the earlier EDA finding that flights departing later in the day tend to have higher delay rates, reinforcing the significance of departure time as a predictor. DISTANCE (flight distance) ranked second, followed by AIRLINE. These results are consistent with the earlier observation of a weak correlation between distance and delay, and the variation in delay rates across airlines.

```
# Visualize variable importance
varImpPlot(rf_model)
```

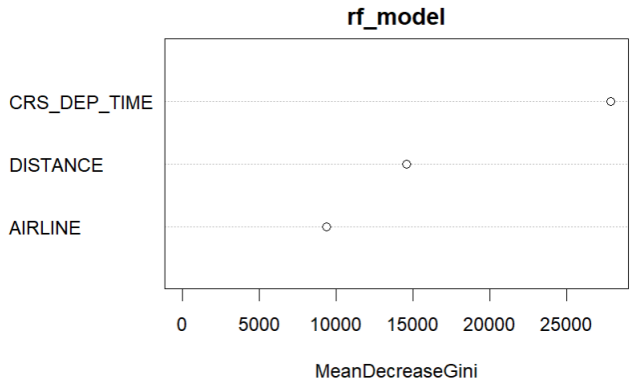

Figure 5. Variable Importance for Random Forest

As a second model, a Logistic Regression classifier was trained. Predictions were classified as delayed if the estimated probability exceeded 0.5. As shown in Figure 6, logistic regression achieved an overall accuracy of approximately 58.9 percent, slightly lower than that of the Random Forest model.

```
# Model 2 logistic Regression
glm_model <- glm(is_delayed ~ ., data = train, family = "binomial")
glm_probs <- predict(glm_model, newdata = test, type = "response")
glm_pred <- ifelse(glm_probs > 0.5, 1, 0)
glm_pred <- as.factor(glm_pred)
glm_result <- confusionMatrix(glm_pred, test$is_delayed)
print(glm_result)
```

```
> print(glm_result)
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 60195 41833
         1 42862 61224

               Accuracy : 0.5891
```

Figure 6. Confusion Matrix for Logistic Regression

The logistic regression model has limitations in capturing complex patterns and nonlinear relationships, whereas the Random Forest model can better account for nonlinearity and feature interactions. As a result, the Random Forest model achieved higher accuracy compared to the logistic regression model.

Although the accuracy of both models was lower than before, this represents a significant improvement in that it was possible to develop a model that predicts both classes in a balanced manner without being biased toward non-delayed flights. Figures 7 and 8 illustrate this improvement.

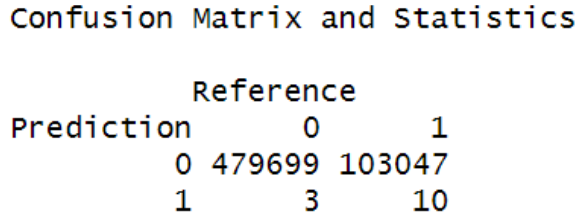

Figure 7. Confusion matrix before the undersampling

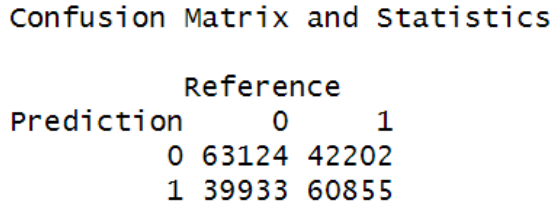

Figure 8. Confusion matrix after the undersampling

Key Insights and Future Directions

Through this project, I found that there are significant differences in the likelihood of delays depending on the departure time, which can lead to practical insights. For example, the observed pattern of a higher probability of delays for flights departing from the afternoon to the evening can serve as valuable information for passengers when choosing flights and for airlines when scheduling departures. In addition, by identifying which factors influence predictions through the variable importance analysis, the results can be used as a basis for more sophisticated modeling in the future. However, the overall accuracy of the model was not high, which I believe is due to the inherent complexity of the flight delay phenomenon. Flight delays are influenced by various external factors, such as weather conditions, aircraft maintenance status, airport congestion, and delays from previous flights, which are difficult to collect in a dataset. As a result, it is challenging to achieve a certain or higher level of accuracy using only schedule-based variables.

This project was meaningful in that it demonstrated the potential for data-driven decision-making by analyzing the complex problem of flight delays using simple variables. However, there were several limitations in the modeling process, indicating the need for careful consideration regarding practical applicability.

The most significant limitation was the constraint of the dataset. The data was primarily concentrated on flight schedules and a limited set of operational indicators, lacking important external variables such as weather, maintenance issues, airport congestion, boarding delays, and connectivity from previous flights, all of which have a strong direct impact on delays. Without these variables, it is difficult for the model to fully capture the causes of delays. The second limitation was model accuracy. In both the Random Forest and Logistic Regression models, the balanced accuracy remained around 60 percent, which is insufficient for use as a reliable predictive tool in real-world operations. Although an undersampling technique was applied to address the class imbalance problem, the models still struggled to explain the complex delay phenomenon using only simple variables.

Nevertheless, the insights gained through this analysis are valuable. The pattern of increasing delay probability with later departure times and the tendency for certain airlines to have significantly higher delay rates can be used as actionable information for both passengers and airport operations planning. Additionally, the variable importance analysis helped identify which features most influenced prediction performance, providing direction for future model improvements through the integration of more external data.

In conclusion, this project showed that it is possible to predict flight delays to a certain degree even with limited data, and that the statistical analyses and visualizations from the modeling process hold potential for development into practical, data-based decision support tools. In the future, more sophisticated models can be designed by incorporating expanded datasets that include external factors, thereby improving both accuracy and applicability.

Reference

"Evaluating Classification & Clustering Models." Rensselaer Polytechnic Institute, Spring 2025. PDF, Data_Analytics2025Fall_Group3_Evaluating_Classification_Clustering_Models.pdf.

"Linear Models & Random Forest." Rensselaer Polytechnic Institute, Spring 2025. PDF, Data_Analytics2025Fall_Group2_Linear_Models_Random_Forest_0.pdf.

"Generalization, Validation, & Optimization." Rensselaer Polytechnic Institute, Spring 2025. PDF, Data_Analytics2025Spring_Group2_Generalization_Validation_Optimization.pdf.

“What Is Undersampling? | Master's in Data Science.” MastersInDataScience.org, https://www.mastersindatascience.org/learning/statistics-data-science/undersampling/

Rathi, Daksh. “Handling Imbalanced Data Part 2: Under-sampling.” Medium, 15 Oct. 2024, https://medium.com/@dakshrathi/handling-imbalanced-data-under-sampling-473dd4e35e8c

"Distributions and Hypotheses." Rensselaer Polytechnic Institute, Spring 2025. PDF. Data_Analytics2025Spring_group1_module3_distributions_hypotheses_0.pdf

"Introduction to Analytics." Rensselaer Polytechnic Institute, Spring 2025. PDF, Data_Analytics2025Spring_group1_module4_introanalytics_1.pdf

“Variable importance plot using random forest package in R.” GeeksforGeeks, https://www.geeksforgeeks.org/variable-importance-plot-using-random-forest-package-in-r/