Flight Delays: A Feature Importance Analysis

Lythia Amoakon and Maria Casas

Faculty Advisor: Dr. Sherry Ni

KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING
School of Data Science and Analytics

INTRODUCTION

Every day, the Federal Aviation Administration's (FAA) Air Traffic Organization (ATO) provides service to more than 45,000 flights and 2.9 million airline passengers.

This rather vast and complex system of daily operations and the continual increase in air travel demand alongside the capacity limitations to accommodate these flights lead to one of the inevitable pains of air transportation experienced by many, delays. Flight delays significantly impact airports' on-time performance and airline operations, which are tightly interconnected with passenger satisfaction.

Delays are a huge inconvenience for passengers and airlines. A better understanding of flight delays and the variables that drive them may help airports in developing strategic decisions to minimize delays and inform consumers on how they can avoid them.

METHODS

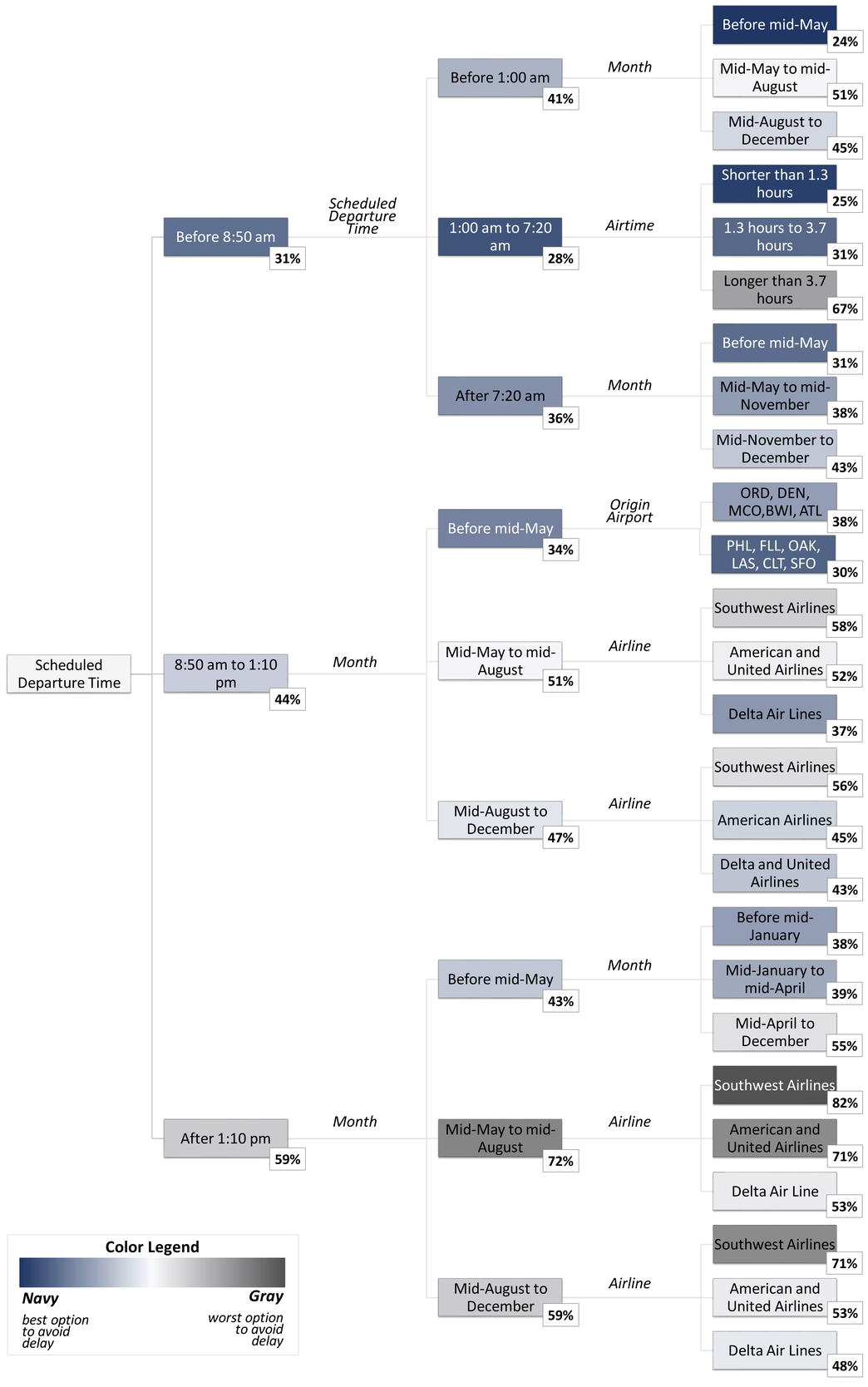
- **Data Collection:** Compiled monthly 2021 csv files from the Bureau of Transportation Statistics using Pandas append in Python. *https://www.transtats.bts.gov/*

- **Data Preprocessing:** Original dataset included 6.3M records with 120 variables.
  - Narrowed our scope to only the top 4 airlines by market share (American, Delta, Southwest, and United Airlines) and cities with at least 10% of the largest city's percent of the dataset.
  - Removed irrelevant, redundant, and unary variables.
  - Imputed missing values with appropriate values.
  - Created categorical variables for departure times arrival times, and flight day of month.
  - After cleaning, preprocessing and random sampling for a more workable dataset, we worked with a 193,654 records and 30 variables.

- **Assumption Testing:** Checking for multicollinearity as well as influential outliers was important for both classification models. Linearity was checked as a requirement for the logistic models.

- **Decision Tree Modeling:** Three decision trees were modeled:
  - Default binary-split tree with time/date quantitative variables as processed from the Bureau of Transportation Statistics site.
    - i.e., Scheduled Departure and Arrival Times, Flight Day of Month.
  - Default binary-split tree with time/date quantitative variables converted to ordinal categorical.
    - i.e., Time of Day Scheduled Departure and Arrival (morning, afternoon, sundown) and Flight Part of Month (early in month, middle of month, end of month).
  - Three-way tree with original quantitative variables for time and date variables.

- **Logistic Regression:**
  - Variables skewed to the right were transformed (Log10).
  - Variables with a more uniform distribution were binned using optimal binning or the categorical variables created were used.
  - Because SAS EM performs dummy coding through the regression, categorical variables did not need further transformations.
  - Three logistic regression models with different selections (Forward, Stepwise, Backward) were run. The stepwise model was retained as the best.



RESULTS

Model	ROC Index	Misclassification Rate	Average Squared Error
3-Way Quantitative Tree	0.715	0.341	0.215
Stepwise Logistic Reg.	0.701	0.354	0.220
Quantitative Default Tree	0.685	0.356	0.224
Categorical Default Tree	0.682	0.358	0.222

Model Comparison

The Model Comparison for the three decision trees and the logistic regression model indicates the **3-Way Quantitative Tree is best.**

Variable Importance (Rank)	3-Way Quantitative Tree	Stepwise Log. Regression	Quantitative Default Tree	Categorical Default Tree
1	Departure Time	Departure Time	Departure Time	Arrival Time
2	Month	Month	Month	Month
3	Airline	Distance	Airline	Airline
4	Day of Month	Airline	Origin City	Origin State
5	Origin State	Origin Airport	Origin Airport	Origin City

Top 5 Important Variables by Model

CONCLUSION

Despite the 3-way decision tree being the best at classifying, all four models resulted in similar conclusions on how to avoid flight delays.

Take early morning flights: Scheduled departure time proved to be the most important variable here. As the hours of the day go by, the odds of a flight being delayed increase. There is a 31% of chance of a flight being late if departing before 8:50 am compared to ~60% when traveling after 1:10 pm.

Limit traveling in the summer: The time of the year is the second most important variable. When traveling in the summer, the probability of a delayed flight is highest. Meanwhile, traveling before mid-May is the best option followed by traveling after mid-August.

The choice of airlines matters: The operating airline is the third most important variable. Travelling with Delta appears to have the lowest probability followed by United Airlines. Southwest has the highest probability of delayed flights.

To increase your chances of an on-time flight, traveling for short flights before the summer and before 8:50 am, especially before 7:20 am, is the best path to avoid delays.

The worst combination would be traveling on a summer afternoon flight with Southwest Airlines.

Opportunities for expanding the analysis would include:
- Using the carrier delay flag as target variable to understand the drivers of delays when the carrier is at fault.
- Feature processing: Add more variables and better feature processing
  - Categorizing states and/or cities into a regional variable.
  - Including more years in the scope of the project. Scope was limited to 2021 due to processing constraints.
- Splitting analysis for larger volume airports/cities.