Visualizing Flight Delays

Team 154 Project for CSE 6242
Kaleb Boyd, Christopher Oyer, Jordan Gressel, Jacob Burdick

Introduction

Every year, flight delays cost airlines, passengers, and businesses billions of dollars. While much research has been done to predict flight delays the results of such research have often lacked a robust visual component, making it challenging for most to makes sense of the data.

We have created an interactive visual model that allows users to review flight delay probabilities across geography and time. Users can select different criteria (such as airline, origin airport, and month) to visualize flight paths and observe how the probability of flight delays changes based on different criteria. A visualization such as this is a first step toward gaining a better understanding of flight delays and provides a foundation that can be built upon to create additional, more comprehensive visualizations to aid in identifying solutions to the flight delay problem.

Approach

--Probability Model--

Our team implemented a random forest model to predict the likelihood of a flight being delayed based on a selected **Airline**, **Month of departure** and **Airport of origin**.

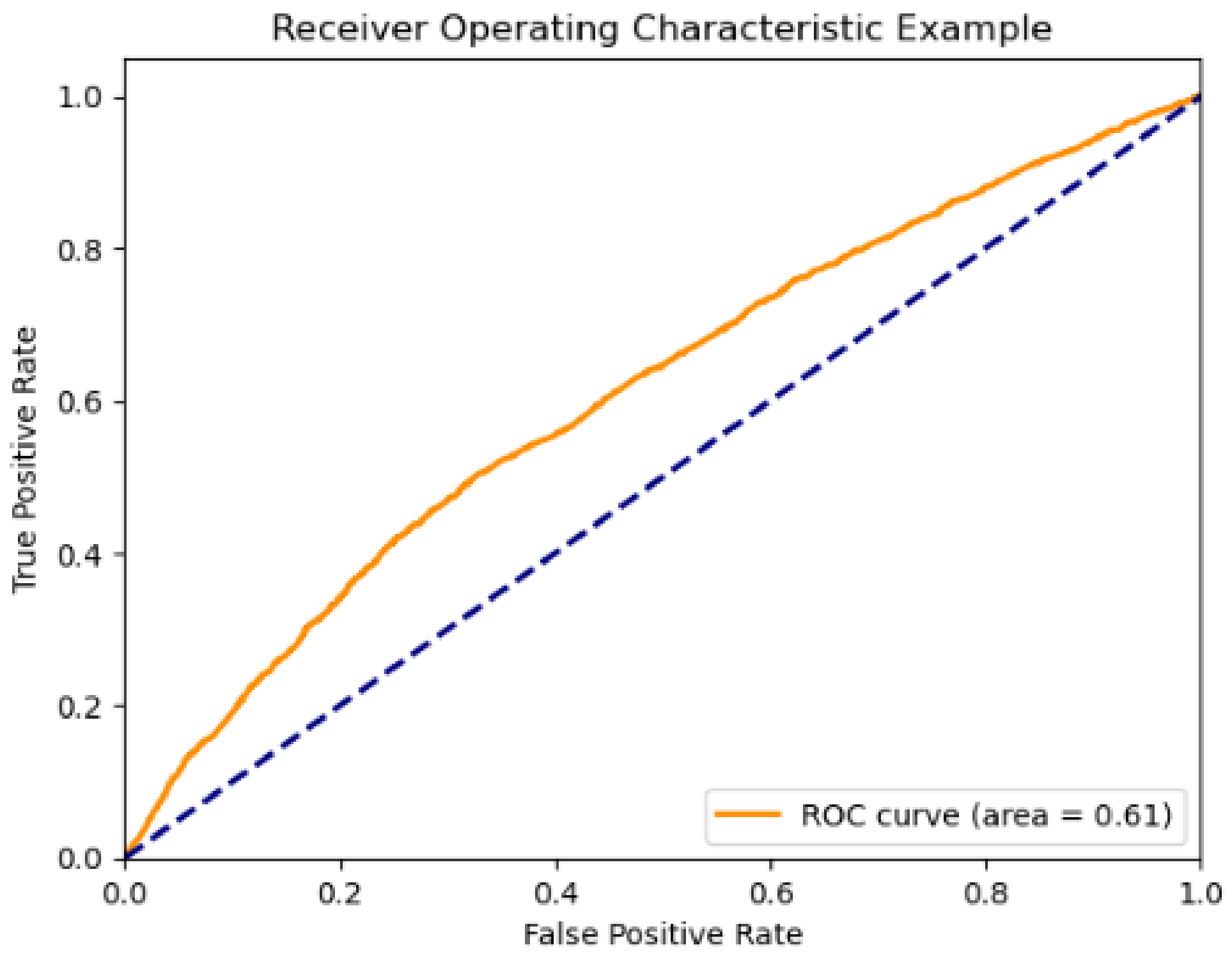
Hyper-parameter optimization was achieved via the grid search method. Different combinations of the following parameters were tested with values ranging from 5 to 30 in increments of 5:

- n_estimators (number of trees in the forest)
- max_depth (number of splits that each decision tree can make).
- min_samples_leaf (min. number of samples required at leaf nodes).

Ideal parameter settings:

- n_estimators = 10
- max_depth = 20
- min_samples_leaf = 5

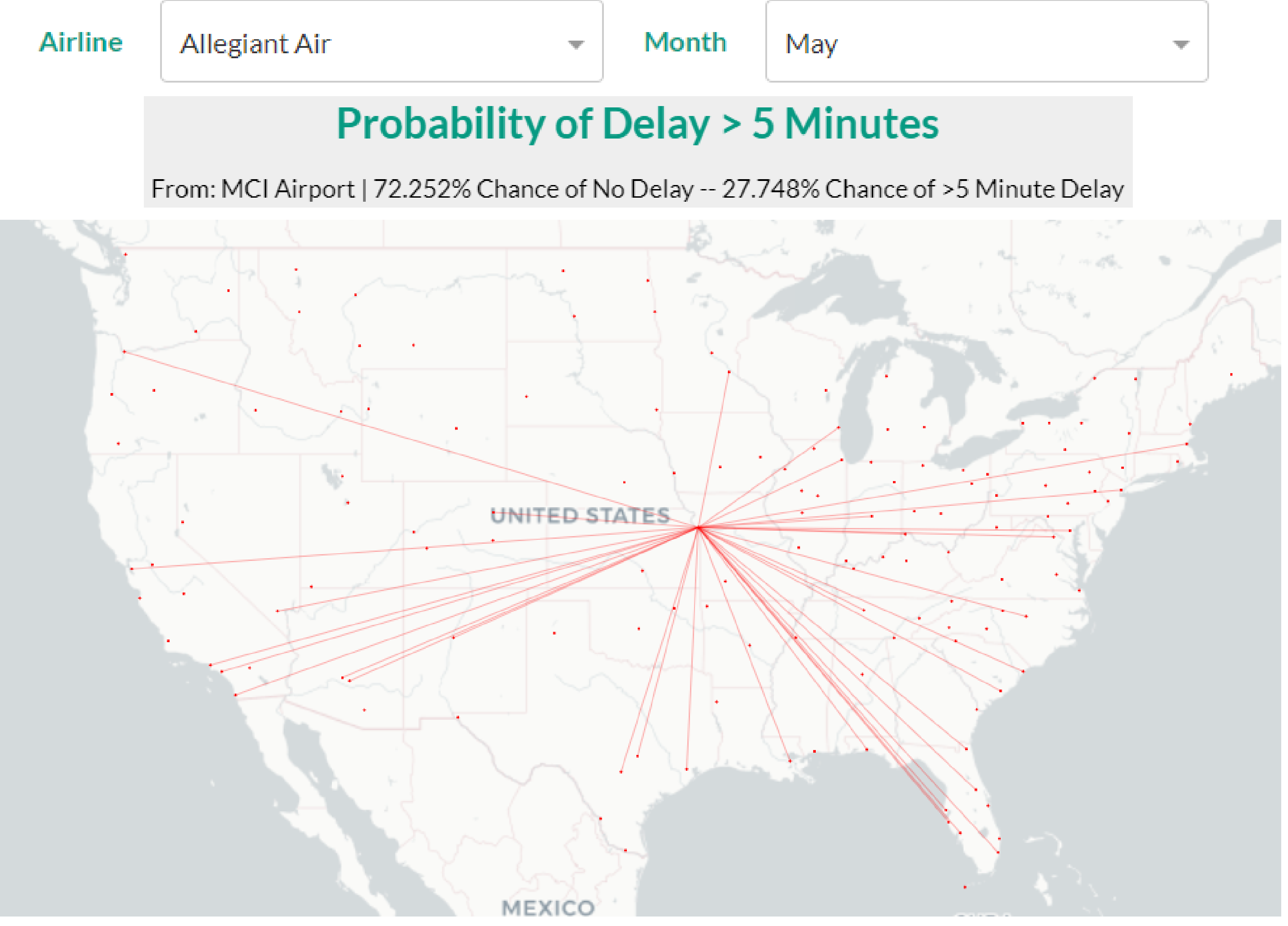
The model was evaluated using a ROC curve and had an AUC score of 0.61.



--Visualization--

Using Dash, we visualized a graph of the air traffic network overlaid on a map of the United States. Each **node** corresponds to an **individual airport** while each **edge** is a **flight path** between two airports.

The drop down menus allow a user to select an airline and a month. Once selected, the probability of delay is calculated in the background utilizing our random forest model and displayed underneath the dropdowns. When the user clicks on a specific airport, a callback function is triggered to highlight the destination airports associated with the origin via trace lines.


Airline: Allegiant Air
Month: May

Probability of Delay > 5 Minutes

From: MCI Airport | 72.252% Chance of No Delay -- 27.748% Chance of >5 Minute Delay

Flight Data Set

--Data Source--

The data used in our analysis is a compiled list of all flights in the U.S. between the years 2018 and 2022. It includes many factors providing a variety of data regarding each flight. The data was originally extracted from the Marketing Carrier On-Time Performance data table of the "On-Time" database from the TransStats data library (a library created and maintained by the U.S. Department of Transportation). We obtained the data via Kaggle using the Kaggle API.

--Data Statistics--

Years covered: 2018-2022
Number of flight records: 29,193,782
Total number of features: 61 (key features used include 'Origin', 'Dest', 'Month', 'Airline', 'Cancelled' and 'DepDelayMinutes')
Number of Airlines: 28
Number of Airports: 388

--Data Preparation and Cleaning--

Data for each year of the analysis was stored in separate files. These files were combined into a single parquet file and then subsequently used in our analysis.

For the model:

A delay indicator binary variable was created where flights delayed more than five minutes were given a value of 1. Any delay over 1,000 minutes was excluded as an outlier. Training and testing data sets were derived via random sampling from the data set. The training data consisted of one million flights while the testing data set consisted of 10,000 flights.

For the visualization:

A Dash web app was created using subsets of the overall data. Examples of these subsets included:

- Unique Origin-Destination flight paths
- Geo coordinate reference tables for the for the airports to display on the map
- A table of airlines that correlated with airport origin and destination

These tables were then utilized by the Dash interface and modified on the fly using the integration of callback functions to further delineate the data for display based on user-selected options (Airline, Month, and Airport) and input those options into our model for an estimate of delay probability.

Future Developments and Improvements

--Probability Model--

- Narrow the scope to a subset of airlines or airports.
- Employ variable selection techniques (lasso or elastic net regression).
- Try other models aside from random forest.

--Visualization--

- Add additional drop downs that incorporate additional model features.
- Call-out box that displays statistics about selected airport or flight path.
- Create a heat map to show changes in probability delay by geography.
- Create similar visualization for flight cancellations and deviations.
- heat Map for probability of delay.
- Visualize flight propagation (how one flight delay leads to others) across the air traffic network.