STA 325: Final Project Report

Calleigh Smith, Hannah Bogomilsky, Hugh Esterson, Maria Henriquez, Mariana Izon

November 22, 2020

Introduction

The importance of air travel in the United States is unparalleled, connecting distant parts of the country with the aviation industry's hallmarks of efficiency, safety, and reliablility. This mode of transportation offers citizens to conduct business, visit loved ones, and travel for pleasure, and the number of Americans flying is widely on the climb. In fact, in 2019, U.S. airlines carried a staggering 925.5 million passengers, a record-setting number, and a healthy increase of 4.1% over the previous year. However, what is the most prominent complaint from these 925.5 million clients? Perhaps unsurprisingly, the answer is delayed flights.

Simply, flight delays brings into question a given airline's devotion to efficiency and reliability, and when such efforts are not met, disgruntled passengers are sure to become an issue. Arrival delays do not occur all too seldomly, with 19.95% of flights incurring arrival delays in 2019, according to the Department of Transportation's Bureau of Statistics. A slew of research has also shown that flight delays, and the ensuing negative reactions by passengers, have consequential effects for all involved, affecting customers' airline choice, as well as their spending habits at a given airport. Thus, it is in the best interest of all parties (customers, airlines, and airport management) to ensure that the maximum number of flights are completed without delay. This goal, of course, is not realistically achieveable, 100% of the time. Yet a model in which to predict arrival delays could benefit all parties involved, offering a better understanding of the duration of any delay and allowing customers and providers to plan to optimize the situation at hand for their collective benefit.

With this thought in mind, our group has taken on the task of using machine learning methods to form a model that accurately predicts arrival delays for real-world flights. By focusing on a popular flight route, within a specific interval of time, our team hopes to accurately predict arrival delays, while also [keeping in mind] the interpretability of available predictors, which could help in an explanation in the primary factors of flight delays. Using publicly-accessible data, the team also aims to provide findings that are readily reproducible and interpretable for all audiences, whether it be fellow passengers or airline executives.

In order to form such a model, we will use various model-building techniques, including multiple linear regression, generalized additive models, and tree-based regression. Informed by statistical measures of goodness-of-fit, diagnostic checks, and in-depth exploratory data analysis, the project will help to develop a choice of a specific model across the viable options. Specifically by comparing relative error metrics across the different types of models, a final machine learning model will be fully explained and interpreted, weighing the relative pros and cons of each statistical decision. Future directions of the project will also be discussed, hoping to draw generalized, yet accurate, conclusions from our dataset and model to the large-scale topic of flight delays across the American aviation industry.

Data

Data Background & Cleaning

The data used within this final project originates from the United States Department of Transportation's Bureau of Transportation Statistics. Specifically, the team has downloaded the publicly-accessible, government data from their Airline On-Time Performance[link: https://www.transtats.bts.gov/Tables.asp?DB_ID=120] database, using a subset of the data entitled “Reporting Carrier On-Time Performance.” This portion of the database records all relevant data for all non-stop flights of major U.S. airlines. It is updated monthly, dating back to 1987, and includes a plethora of informative variables.

The Bureau’s website allows for a direct download of the dataset for a given month and year by means of a .CSV file. For purposes of this project, the team opted to choose January 2020 as our time period of interest. Several considerations were involved in this decision, including the choice of a recent month that was not severely affected by the COVID-19 pandemic. Thus, the data collected from this month will not showcase the drastic and devastating effect that the pandemic has had on air travel traffic. We also chose to focus on a specific non-stop route within this month of data. Us

Methods

Results

Future Directions

- could expand airports, years, COVID effects, etc.

Notes & References