ORD Domestic Flight Delays

Literature Review

Team:

Aden Ramirez

This is the start of the lit review.

Data Science in flight delays, Talk about delay segments, airports specifically.

Flight Delays Literature Review

As the air travel has grown substantially over the past 60 years, the investment in new facilities, aircraft and improved efficiency has grown along with it. Airlines in the United States pour huge amounts of time and effort into reducing delays each year. In the past 10 years we have observed delays hover around 20% in the country, with the exception of 2020. To further understand delays and what exists in the data science realm, I explore some of the research and their discoveries.

To begin identifying some areas of interest is important to further understand the topic and begin the design of a model. "As seen in Figure 1, there are six major themes regarding the flight delay: departure, arrival, propagation, airline, airport, and air system" (Carvalho, Sternberg, A., p. 502). As they point out all these areas have significant impact on a flight, from facilities to actual flying, focusing on one of these and using the rest as aspects in a prediction will allow for a good model to be developed, but be loosely applicable to the rest. Knowing areas of interest allow me to break down further and apply data science to a smaller area.

The data to be used is a good foundation for this exploration. In multiple articles *On the relevance of data science for flight delay research: a systematic review. Transport Reviews, Estimating Flight Departure Delay Distributions-A Statistical Approach With Long-Term Trend and Short-Term Pattern,* and *Predictive Modelling: Flight Delays and Associated Factors, Hartsfield–Jackson Atlanta International Airport* applicable data techniques were applied to sets sourced mainly from two organizations. These organizations being the Federal Aviation Administration (FAA) and Bureau of Transportation Statics (BTS). Both reputable government organizations, the FAA regulating all encompassing air travel, and BTS reporting data around transportation. The set we have chosen has been referenced in multiple studies, pulling from the BTS database of on-time flight statistics.

Applying methods on this data would build towards some next steps of research that provides further progress in the field. "One step in that direction would be to try to extract, from individual airline–airport models, the effects that contribute to NAS-wide delay. Such an approach would provide more insight into the general structure of delays, and also would be easier to maintain and update on a NAS-wide basis" (Tu, Ball, M. O., & Jank, p.124). This set will be used to predict delays in Chicago's O'Hare International Airport (ORD), using methods such as regression, bringing another airport model that could support or refute findings at other airports contributing to a national level of predictions.

Looking at airport specific studies some consideration must be given to the data with methods used. For example in "Since we faced a problem of unbalanced classes with 86% of flights without delay and 14% of delayed flights and since this problem can lead to a false classification accuracy [20,21] we applied a sampling technique to minimize it" (Henriques, & Feiteira, I, p.641). Classification being used in their studies had to build on top of cleaning to maximize accuracy. Methods being used include regression, classification, and neural networks. I plan to build on top of their work and with working on a specific airport and tailoring data preparation. Data is tailored to each airport or airline, or even airport and airline in these articles and examining their methods they can be tweaked to be useful for ORD.

Research in flight delays is extensive, though much of airline delay prediction is a more closely guarded secret as they use this to build their flight schedules. Looking at different reasons for delays reveals some areas of focus. My research will focus on the airport and using other identifiers for the model, but this allows the narrowing of the dataset, and can contribute to a national prediction adding an airport to extrapolate national identifiers. The data being used is reliable and referenced in multiple research articles, allowing for some questions to be expanded from others. Methods and cleaning must be tailored to the airport, but may be similar to previous research, which can help finding the most prominent variables that effect a flight delay. This research will expand the knowledge base with ORD delay research, introducing more up to date data.

References:

Carvalho, Sternberg, A., Maia Gonçalves, L., Beatriz Cruz, A., Soares, J. A., Brandão, D., Carvalho, D., & Ogasawara, E. (2021). On the relevance of data science for flight delay research: a systematic review. Transport Reviews, 41(4), 499–528. https://doi.org/10.1080/01441647.2020.1861123

Tu, Ball, M. O., & Jank, W. S. (2008). Estimating Flight Departure Delay Distributions-A Statistical Approach With Long-Term Trend and Short-Term Pattern. Journal of the American Statistical Association, 103(481), 112–125. https://doi.org/10.1198/016214507000000257

Henriques, & Feiteira, I. (2018). Predictive Modelling: Flight Delays and Associated Factors, Hartsfield–Jackson Atlanta International Airport. Procedia Computer Science, 138, 638–645. https://doi.org/10.1016/j.procs.2018.10.085

Bureau of Transport Statistics At a Glance. BTS. (n.d.). Retrieved February 3, 2023, from https://www.transtats.bts.gov/homedrillchart.asp