Project Proposal

Flight Arrival Delay Analysis

Mary Silva Wyara Moura Silva

The Data

The U.S. Department of Transportation's *Airline On-Time Performance Data*[1] is a database that contains flight records spanning the past 20 years. Records of flight departures from up to 13 major airline carriers and 306 origin cities across the United States, including the arrival delay in minutes and the distance of each flight are contained in the database. For January 2017 alone, this is approximately 450,000 recorded observations of flight delays.

We intend to combine this dataset with weather data to be collected from National Oceanic and Atmospheric Administration[2]. The first choice of climatic data would be precipitation data. These dataset are hourly precipitation where they provide hourly amounts of precipitation per hour for a network of more than 7000 stations located mainly in the United States. In addition, we also have the option to use data of 15 minutes of precipitation, so we will decide which variable best fits the model, which we will use. The other weather data options are temperature, wind speed, which can also be collected in [2].

The objective for this project is both inference and prediction. Ideally we would like to determine which variables (i.e. precipitation, temperature, wind speed, distance of flight, date of flight, etc.) effect flight delays. This will be able to answer our main question of what factors contribute to delays in flight. Additionally, we would like to be able to produce a model which can predict delays based on statistically significant variables. To reduce the size of the dataset, we may only span one year and between 10-12 popular U.S. airports.

Methodology

We have not decided if a linear regression or a logistic regression approach will be appropriate. It's possible we may look into both methods. If we were to take a linear regression approach, the response variable will be the delay in minutes. Since predicting the exact delay in minutes doesn't seem useful, this approach may only be used for determining statistically significant variables. If we were to take a logistic regression approach, the response will be treated as binary, separating the data into delays less than 30 minutes from delays greater than or equal to 30 minutes.

Additional Comments

Because the *Airline On-Time Performance Data* is popular dataset, similar analysis of this data may have been produced. However, we only plan to use a specific subset of the dataset; for instance, we may only use the past year. Also, by combining this dataset with weather data which our conclusions are expected to be unique.

References

[1] United States Department of Transportation. (2016). *Airline On-Time Performance Data*. Retrieved from `https://www.transtats.bts.gov`

[2] National Oceanic and Atmospheric Administration. (2016). *National Climatic Data Center*. Retrieved from `https://www.ncdc.noaa.gov`