Project 1 – Front Left

Scott Kaufman, Dennis Boston, Todd FitzGerald, John Ekblad, Peter Resnick

Flight Delay Analysis and Model Construction

Overview:

As a group, we decided that an investigation into flight delay data and corresponding weather observations may allow us to create a predictive model to understand future delays based on quantitative inputs. We believed the combination of transportation and weather data was robust enough to allow us to build a substantial data set that could be used to develop a series of meaningful insights. This analysis was based in two parts: initial data cleaning and exploration, followed by a model building exercise.

Main Data Sources:

Our data comes from a combination of publicly available .csv files and calls to public APIs. We first believed we could use FlightStats to obtain all of our flight delay data, and we were initially successful. We pulled a broad range of quantitative weather observations from a large amount of flights and were able to put together the beginning of a strong data frame, before realizing we had hit out API call limits. We did not fully understand the restrictions at the beginning of the project and were therefore forced to explore other sources. We opted to use government backed data as these measurements were consistent across the country and could be obtained easily via a .csv download. After some slight cleaning, we had a set of 458,727 flights in September of 2017 and began to analyze. Later on in the project, we would use Weather Underground to relate delays with weather observations at that time and location by manipulating timestamps.

Initial Analysis:

To begin our analysis of flight delay data, we created a visualization showing flight delay lengths for each airline carrier in our data set. This plot encompassed all 458,727 flights, showcasing general performance for each airline, relative to each other. As we continued this analysis, we began to notice that we were incorrectly comparing airlines with unrelatable data. For instance, Hawaiian Airlines has substantially fewer flights at fewer airports than larger carriers, such as United or Delta. In order to combat this, we normalized the data by comparing z-scores of the delays, from only the 10 largest airports in the country. This allowed us to better understand how airlines truly compared with each other when it came to flight delays.

Weather-based analysis:

When cleaning our flight delay data set, we uncovered that there are 5 categories of flight delay: Carrier Delay, Weather Delay, NAS Delay (air-traffic control), Security Delay, and Late Aircraft Delay. In order to build our delay prediction model, we determined that we would focus on only weather delays. This allowed us use API calls to try to bring in as many data points as possible, while the other 4 delay types would be much harder to quantify. If we could correlate quantitative weather observations to flight delay lengths, there was a better chance of us building an accurate model. We wanted to select data that was non-binary, as binary data tends to make it harder to predict a length of delay. Knowing this, we selected 4 criteria to build our model: wind speed, pressure, visibility, and temperature.

We were able to examine and extract the timestamp of each flight delay and use this timestamp (with some manipulation) to call the weather underground API. This means we were essentially able to determine the weather for each flight delay at there given time and location.

Our first step was to plot each of these criteria against the length of the delay and see if there were any obvious contributors to flight delay. None of our 4 plots showed clear relationships between measured observations and flight delay lengths, there seemed to be some slight trends, such as shorter flight delays as wind speeds increased, and longer flight delays when visibility in maximized. However, our limited domain knowledge allows us to rationalize that these trends do not make sense. There is no logical reason that increased visibility would create longer delays. This anomaly is likely seen because there is a maximum value of 10 miles in the data set, meaning that even times when visibility was much longer than 10 miles, the observation set denoted a value of 10. This should be dealt with in our analysis, however we did not take those steps. There are statistical procedures that can be used to quantify these unique cases, and should we reevaluate the project, we would likely take these steps.

Because we did not see any direct correlations between the measured weather variables and the flight delay lengths, we attempted to compare variables to each other, in hopes of seeing correlations there. To do this, we created a scatter matrix, but our results were not promising. Our visualizations were chaotic and hard to follow, there were no obvious patterns within these plots. We then attempted to color-code our “bins” of delay lengths in an attempt to visualize where other possible trends may have lied. If we were able to see that long delays most often occurred when Pressure and Temperature were both high, or short delays most often occurred when Wind and Visibility were low, we would have been confident going into our model building process. However, this is not at all what we saw, leading to some worries are we began our attempt to build a model.

Model Building:

We chose to experiment with 6 different models, some using similar techniques, others using very different techniques. The chosen models were: Logistic Regression (LR), Linear Discriminant Analysis (LDA), k-Nearest Neighbors (KNN), Classification and Regression Tree (CART), Gaussian Naïve-Bayes (NB), and Support Vector Machine (SVM). Our biggest issue as we began building models, was lack of useful data. As we called the weather underground API, we quickly reached our limit at 784 flights. This limited set meant we would have a hard time optimizing our model. One of our biggest decisions was how many cross-validations to have in the models. There can be lots of work done to determine the appropriate number, when we run only 2 cross validations, our model has a larger data set to learn from, but not very long to learn. Conversely, if we cross validate 10 times, there a lot more time for the model to learn and adjust, but the data it is learning from is so slim that we can’t be confident in it’s outputs. The analysis that can be done on model optimization in generally is significant and out of scope for this particular project.

We also need to understand the assumptions that these models are making. For instance, the NB model assumes a Gaussian distribution of our data, which frankly, ours is not. This is a reason as to why our accuracy measurements are so low. Should we have had more time, we would have gone more in depth into the underlying assumptions and mathematical methods. These models, as noted in our presentation, use complex mathematical methods that are difficult to comprehend in 2- or even 3-dimensional space. We also wanted to take some logical steps to see whether or not our models were behaving as we expected them too. To do this, we reevaluated our classifications such that there were only 5 classes, instead of 8. When rerunning our models using 5 bins, we saw accuracy improvements across the board, meaning we can be more confident that are model is on the right track.

Conclusion:

If we had a much larger data set (10s of millions of flights), we would have been able to optimize more of our model while also seeing the benefits of larger training vs testing splits. Additionally, we would have benefitted from a deeper understanding of each models’ methods and how those methods may or may not be appropriate for our specific application. Knowing this, it is not surprising that we were able to determine that some airlines perform better or worse than others based on normalized scores, while also determining some interesting observations regarding extreme weather (hurricanes, wildfires), but we were unable to construct a model that predicts flight delay lengths based on weather observations with a reasonable degree of accuracy.