Flight Delay Predictions
Capstone Project 2

Every day millions of people plan trips across the United States that involve commercial airline flights. It can be a major inconvenience when those flights run late and impact the trip. The passengers can be a make better decisions when planning for their trips by knowing how much of a delay they can expect for their flights.

With all of the data that is made publically available by the Bureau of Transportation Statistics (https://www.transtats.bts.gov) as well as the weather data from the National Oceanic and Atmospheric Administration (https://www.ncdc.noaa.gov/), it seems that a reasonable estimation for a flight delay could be made.

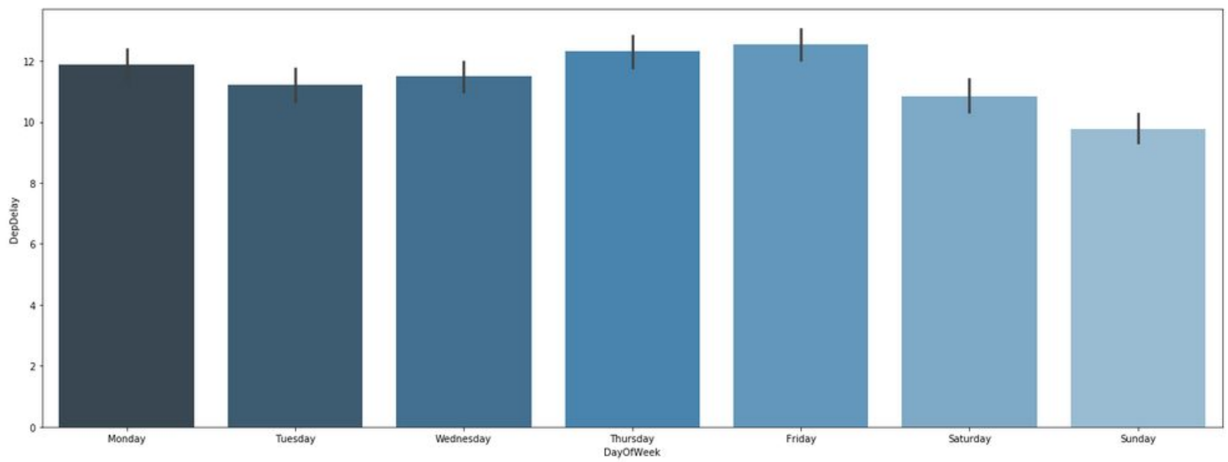
An issue with analyzing flight delays is the wealth of information that is available. There are tens of thousands of flights per day in the United States. For this project it was not reasonable to look at all of the flight data due to disk space and processing power. Instead the focus will be on flight departing Dulles International Airport (IAD), the major airport outside Washington, DC. Narrowing our scope to just one airport allowed us to collect the data from the last 5 years to help make predictions.

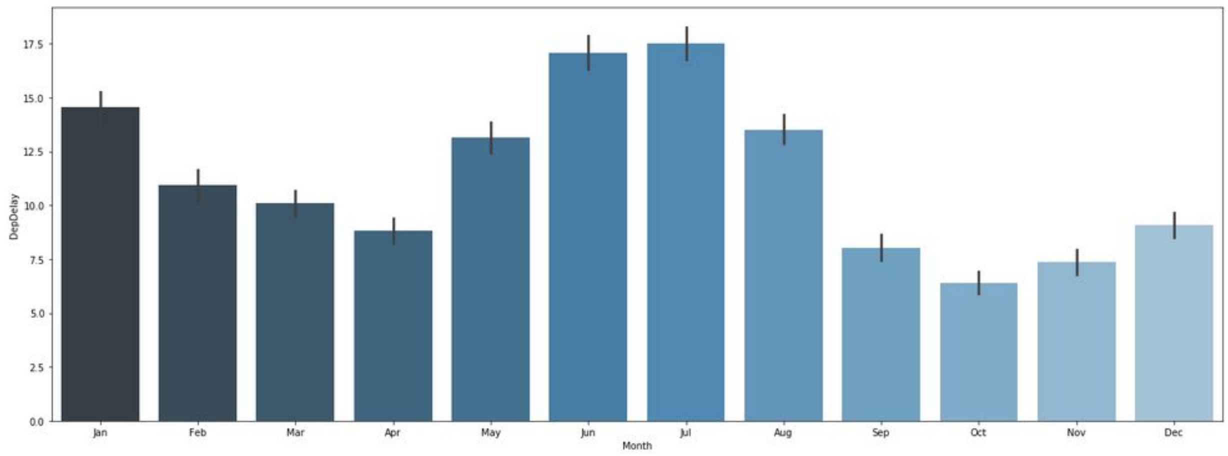
The data from the Bureau of Transportation Statistics (BTS) as well as the weather data from the National Oceanic and Atmospheric Administration (NOAA) is very well maintained. There was surprisingly little that needed to be done to clean and wrangle the data into a format that could be further analyzed. One topic that did arise was the situation with Cancelled flights. These are decidedly an inconvenience to the passengers much like a flight delay. However, we are trying to predict a precise number of minutes for a flight delay. Cancellations do not have a precise delay associated with them and were therefore removed from the analysis. Likewise, some of the flight observations did not have any data for their delay. These flight observations were also removed because they could not add value to the analysis.

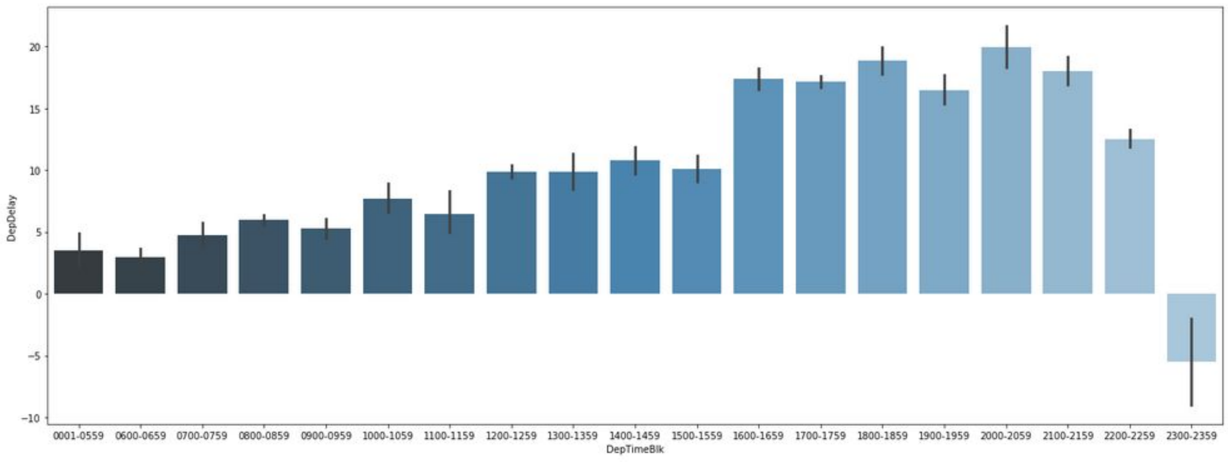
A decision was also made to use categorical data whenever possible as opposed to numerical data. Data points such as day of the week were assigned numerical values, for example Monday as 1 and Tuesday as 2. This can be advantageous since it provides a logical order to the days of the week, but it can unhinge the analysis at the end of the week where Monday (1) may not quite follow Sunday (7).

Daily weather data taken from Dulles Airport was merged with the flight data to try and enhance predictions. According to BTS weather delays cause about ⅓ to ½ of delay minutes.

Visualizing the data shows some clear trends. Day of the week, month of the year, and time of day all showed some clear cut differences.

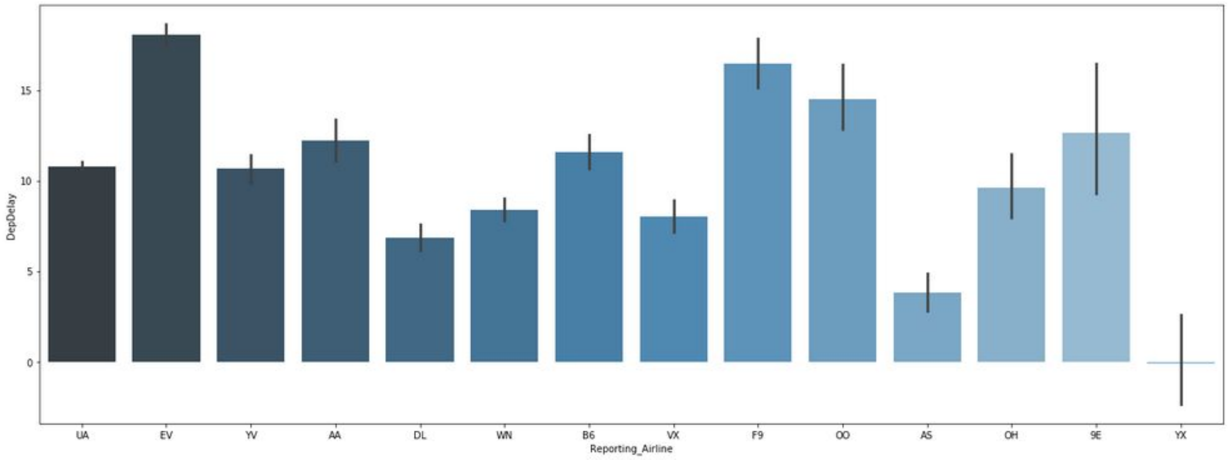






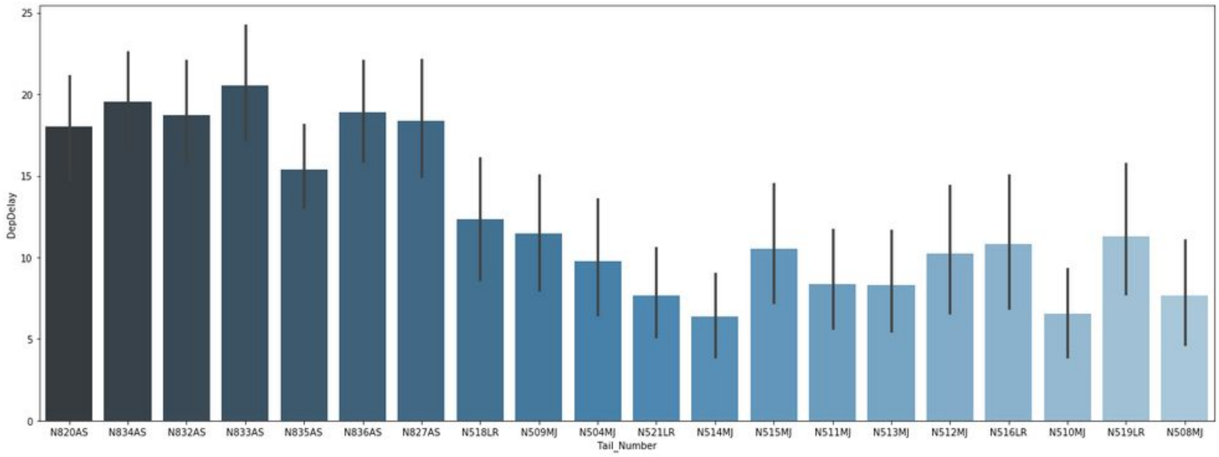
Based on these graphs a July fight at 8 pm on a Thursday will likely have a longer delay than an October flight at 11 pm on a Sunday.

Breaking down the departure delays by airline also show more clear cut trends.

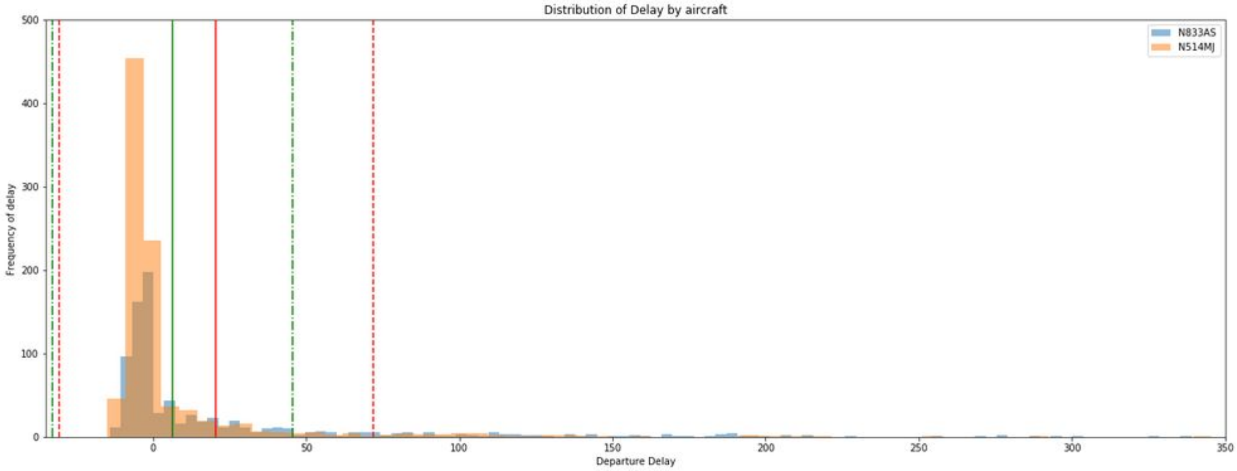


One interesting feature that is included in the BTS data is Tail Number. It does not make sense to include Tail Number in the final analysis since passengers aren't likely to have that data available in the days leading up the flight. However, it is interesting to see if there are "problem aircraft" that have longer average delays than the others.

Looking at the most frequently used aircraft over the past 5 years departing Dulles Airport shows some big differences in delays. These aircraft were used between 800 and 1,000 times in the past 5 years.



The aircraft on our above list with the longest delay and with the shortest delay were used for a deeper dive. N833AS has an average delay of about 20 minutes and N514MJ has an average delay around 6 minutes. The histogram below shows the distribution for delay for each aircraft. The long delay aircraft (N833AS) data is in blue with red lines and the short delay aircraft (N514MJ) data is in orange with green lines.



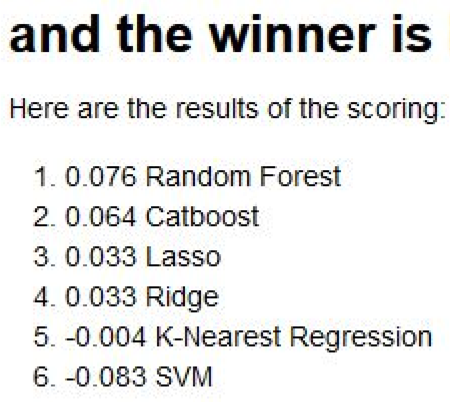
At the time the analysis was being done to show statistical significance of the delays for each aircraft. What this also happens to show is a trend for this data to not follow a clean or normal distribution. Data that looks like this can cause issues for predictions. The previous visualizations provided great hope for accurate predictions, this one does not inspire confidence.

Some final data cleaning needs to be done before the data can be analyzed by the models and predictions can be made. Many models are based on Linear Algebra and work best with specific data layouts. A heatmap was used to remove data with a high correlation. Object columns were shifted to numerical data with “one hot” encoding. Finally the data was scaled to ensure the size of the variables do not improperly influence predictions.

6 different models were used and then their results were compared to show what will work best on our dataset. Ridge, Lasso, K-Neighbors Regression, Support Vector Machine, and Random Forest were all used from the Python Scikit Learn (https://scikit-learn.org) package. Catboost (https://catboost.ai/) was also used as an “outsider” for comparison.

All of the Scikit Learn models used Cross Validation to some degree to improve parameter tuning. Due to some of the model’s being very computationally expensive the Cross Validation was, at times, done with only a sample of the full dataset. Catboost was run without parameter tuning. The R Squared statistic (coefficient of determination) was used for scoring.

R Squared scores can range from -1 to 1 with a perfect score being 1. Regrettably, none of the models scored well. The “winner” was Random Forest with an R Squared of 0.076. This would not be a good model to use for predicting the number of minutes that a flight will be delayed.



and the winner is Random Forest Regressor!

Here are the results of the scoring:

1. 0.076 Random Forest
2. 0.064 Catboost
3. 0.033 Lasso
4. 0.033 Ridge
5. -0.004 K-Nearest Regression
6. -0.083 SVM

Precisely predicting a delay for the flight is difficult. While ⅓ to ½ of all delay minutes are due to weather which we included in this analysis, about ⅓ are due to factors that are harder to predict, such as maintenance or crew problems, aircraft cleaning, baggage loading, fueling, etc. On top of that, a little less than ½ of delays are caused by aircraft arriving late. Since the focus here was only aircraft departing Dulles Airport and weather at Dulles Airport, that does not allow a good prediction of aircraft that will arrive late.

Additionally, the way the data does not follow a predictable shape also causes problems with precise predictions. Predicting delays is often done as a classification problem. Based on the data that we saw here it would be much more accurate to predict if a flight will be late by more than 15 minutes instead of trying to predict the precise number of minutes.

The next steps for this project is changing to a classification problem and, if possible, pulling in more data.