Kalen Willits
Data Science Portfolio

Predicting Airline Flight Delays

Imagine yourself on the last day of work before the big family trip. You have coordinated with your manager to get extra time off, shopped for weeks to get the very best travel deals, and your phone gives you a notification that it's time to check in for your flight. You choose preferred seating so that the kids don't have to crawl on top of you to go to the bathroom and arrange for friends to pick you up. The next morning you pile everyone in the car, rush through security, and make it to your gate just in time to find out that the aircraft was rerouted due to a storm resulting in the airline scrambling for another aircraft. All the planning, preparation, and preferred seating is gone.

Wouldn't it have been nice to know how likely this could happen when you checked in?

In this analysis we look at ways to provide the technology to do just that. The data we will be using is historical and contains information from over 450,000 flights in the United States during January 2017. To view the detailed analysis and the code on how I arrived at these conclusions, check out the Jupyter Notebook on Git Hub.

The first thing to look at are what pieces contribute to the flight delays. We all know a severe storm halts travel plans, and our tools are pretty good at predicting them. What we are concerned with are the little ripples in airline traffic that create the smaller surprise delays. Factors recorded in our data set are departure time, taxi out, taxi in, arrival time, cancellations, diversions, distance, weather delays, and security delays just to name a few.

Once we isolated which pieces of data to use we could start identifying and visualizing correlations. Logically, we expect departure time and arrival time to have a strong correlation along with distance and air time as well. What was most interesting is the shape of a departure delay versus a late arriving aircraft. This shows that not all late departures result in a late arrival.

Scatter plot of late aircraft delays versus departure delays. What is interesting is that there an arc to the shape and late departures carry a linear ceiling on a late aircraft.

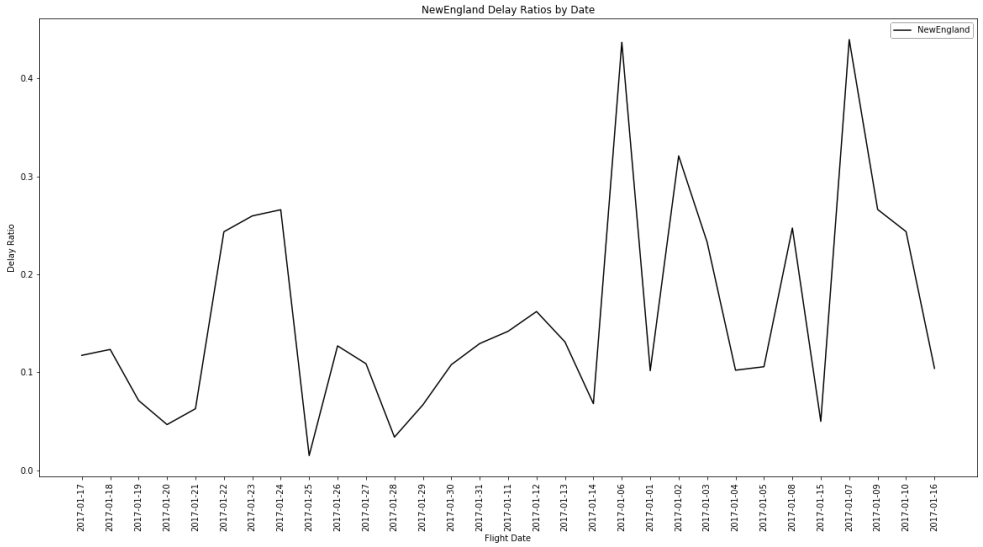
We now have an idea of how the features in the data work together. My hypothesis is that late flights arriving will cause a reverse ripple affect measured over time in late departures at the destination airport. Logically this makes sense because if a Boeing 737 is booked to leave Chicago at 8:00 AM Central and arrive in Miami at 11:32AM Eastern, and is delayed by security for 20 minutes, the subsequent flight that aircraft is has been booked for out of Miami will be affected by this delay. I suspect the arc in the scatter plot above is created due to various counter measures the airlines and ATC employ to reduce the reverse ripple effect. If an early aircraft takes the flight that the late one could not make due to a delay, our delays become reduced and hard to track. This is where measuring a local delay ratio comes into play.

The delay ratio is calculated by summing all the flights that have been delayed at the origin, and dividing by the total number of flights made at the origin. The trick is narrowing your scope by location and time. Doing so produces meaningful measurement that does not generalize too much.

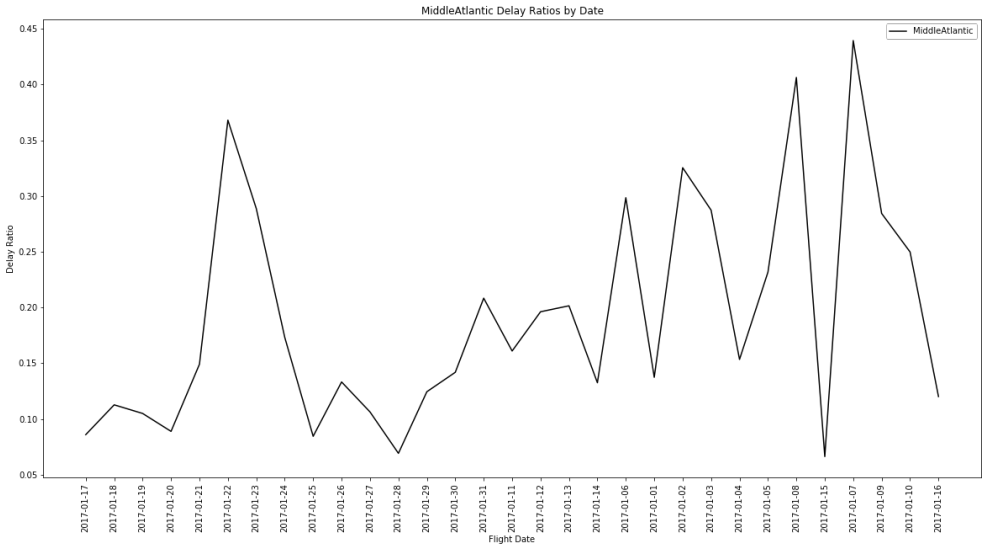
Reverse ripple shown from ORD, LAX, and DEN. Other delay ripples can be seen, however the blue line shows a ripple that has been clearly captured.

The above line plot shows the trends of delay ratios throughout three airports. There is a clear reverse ripple originating in O'Hare, sweeping through Los Angeles, then the smallest peak in Denver. This means we could have predicted the amount of flights that are delayed in Denver based on the amount of flights that are delayed from aircraft targeting Denver as their final destination. This is just one example of delay ratios leading to delays at other airports.

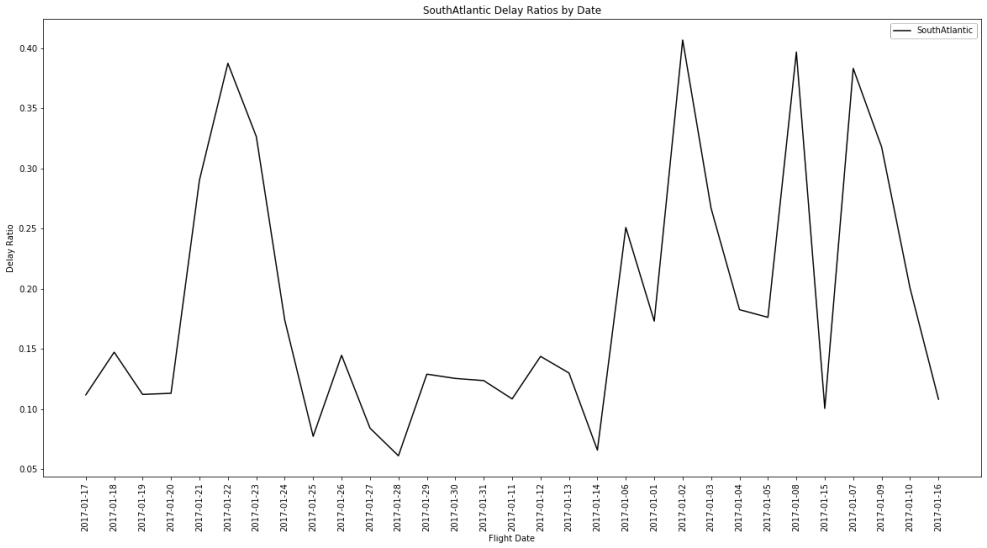
For the scope of this analysis, we will be looking at the top one hundred airports in categorized by the nine regions of the United States. Below is the trend found in the New England region.



New England delay ratios calculated by the top 100 airports in the US.



Middle Atlantic ratios calculated by the top 100 airports in the US.



South Atlantic delay ratios by top 100 airports in the US.

When comparing the Eastern delay ratio trends, we can see the delays are similar between the regions. One explanation for the spike in delays in near the beginning of the month was the severe winter storm originating in Philadelphia on January 7th. We can see that the trend shows delays splashed across to other regions.

Now we can use machine learning to predict the delay ratios by region. Several

machine learning models were attempted. Some are noted in my Jupyter Notebook. The algorithm that performed best on the test data was support vector regression.

SVR model predicting flight delays against the test data by region.

Using the US regions as test data, all regions performed with a mean square error and mean absolute error below 0.1. Data modeling parameters and metrics are recorded in the Git Hub repository.

Let's lower our scope and test our model against the Middle Atlantic flight delays.

SVR prediction on Middle Atlantic flights.

Now we can use our model to provide a delay ratio forecast to customers during check in. The scope of our data was daily delays over the course of a month, however the same model could be used over hours as well. By watching the ripples that are created when delays in one area occurs, we could follow those aircraft and provide passengers with an estimation of what to expect when they arrive at their gate. This will give passengers the time to prepare for possible delays. ATC and airlines could use these forecasts by month, like the one performed here, to provide insight on what aircraft should be prepared and where in case of a delay. The flight could be taken on the held aircraft and reduce the amount of ripples through the day. Airlines can also use the hourly delay forecast to incentive passengers downloading their respective mobile apps, and only provide a static forecast to at check in.

The current data used to train the model was limited to one month. During this

month, there was a winter storm causing significant delays originating in the South East. This storm could be throwing the model off when training for non-weather related flight delays. Also, the single month of data is an unacceptable constraint for production. For more accurate predictions, we would want to use years of data to incorporate how seasons affect flight delays.

2017-Jan-OnTimeFlightData-USA

Data.world

Top 100 airports of 2017

worldairports.com

Top 100 airports in US

fi-aeroweb.com

Census Region Division of the US

US Census

Storm Prediciton Center

spc.noaa.gov

 Kalen Willits / August 15, 2020 /

Kalen Willits / WordPress.com.