Project McNulty Summary
Predicting Flight Delays at Seatac Airport

Dana Lindquist
November 2, 2018

Project Summary

The goal of this project was to predict if a domestic flight arriving at Seatac Airport in Seattle, WA would be on-time or delayed. Delayed was defined as 15 minutes after the scheduled arrival time which is the standard in the airline industry.

Data

Data was collected from

- Bureau of Transportation web site (transtats.bts.gov) which has all the flight arrival and departures up to July 2018
- NOAA web site (www.ncdc.noaa.gov/cdo-web/datasets) which has historical weather data by location

The flight data included

- Arrival and Departure Date, planned times and actual times
- Departure and arrival city
- Flight distance
- Airline and flight number
- Reason for delay (if delayed)

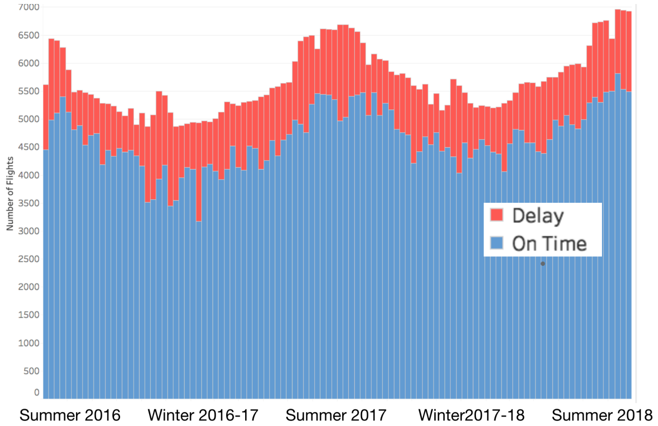
The weather data used in this project contained daily summaries. The cities of Seattle, Chicago and New York City were included with the following information.

- Date and location
- Wind speed
- Snow
- Temperature

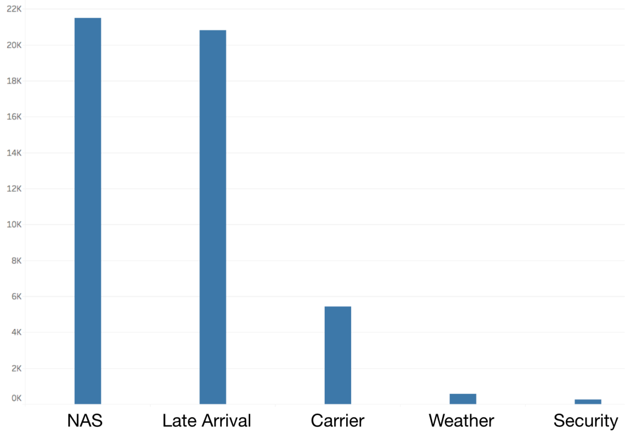
The dataset for flights arriving at Seatac contained 274,303 flights, 19% of which were delayed.

Data Visualization

Before getting too involved with modeling, it was helpful to get a better understanding of the data. Most of the data visualization was done in Tableau which made it easy to 'look' at what information was contained in the dataset.

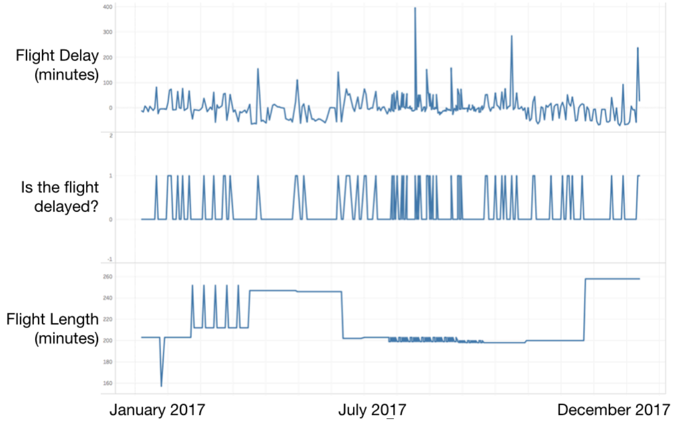


By looking at the trend of on-time and delayed flights arriving at Seatac for the 2 year period in the dataset, it was interesting to note that the percentage of delayed flights remained relatively constant throughout the year.



While weather had some impact on flight delays, it was not a major contributor. In fact the major contributor to flight delays was the National Air Service (NAS) which is out of control of the airlines. Late arrival of the aircraft before the flight in to Seatac was a major contributor to delays as well as general issues with the air carriers. To account for Late Arrival aircraft while modeling, weather from Chicago and New York City were included in an attempt to capture the source of Late Arrival delays.

By looking at the data it was possible to see airlines changing the length of flights by adjusting the take-off and landing times of flights. Above is the change in flight time



and delays for Alaska Airline Flight 259 from San Jose, CA to Seatac for calendar year 2017. You can see that the flight length has been changed and it has the expected impact on flight delays. I believe that the airlines have identified a tolerance for flight delays and are adjusting to meet these delays.

Models

The dataset was biassed 81% toward the majority class (on-time) and 19% to the minority class (delayed). My early models wanted to predict 100% on-time. To force the models to use the minority class, the majority class was downsampled so the classes were the same size. This was only done for training. The cross-validation and test sets contained the original distribution.

Several models were used for this project and results are listed in the table below. The Random Forest model performed the best overall, but also had the best score for Recall, the statistical measure most relevant to this project. Recall is important because people won’t get too upset if you predict that a flight will be delayed when it’s actually on-time (too many false positives) and they will be upset if you predict that a flight will be on-time and it’s actually delayed (too many false negatives).
The Random Forest model had the highest Recall at 0.64 and actually ranked highest in all statistical measurements. The Decision Tree model came in second on Recall but the KNN model came in second for all measurements.

Model	Accuracy	Precision	Recall	F1	AUC	Notes
KNN	0.63	0.28	0.56	0.37	0.65	n_neighbors = 10
SVM	0.54	0.21	0.52	0.30	0.54	kernel=‘linear' C=1
Logistic Regression	0.56	0.23	0.56	0.33	0.58	C=1 penalty = L2
Decision Tree	0.58	0.25	0.59	0.35	0.61	max_depth = 14 min_samples_leaf = 6
Random Forest	0.65	0.31	0.64	0.41	0.70	n_estimators = 100 max_depth = 10 min_samples_leaf = 4

Conclusions and Future Extensions

This project only looked at arrivals at Seatac airport. It would be very interesting to look at a different city and compare the results. Seattle is at the west side of the continental US. What would data from a city in the center or east side look like?

It would also be interesting to make this a multi-class project by adding another class called ‘Very Delayed’ for flights over an hour delayed. Would this be easier to predict or harder to predict.

Extending this project to the extreme, combining all the cities in the USA into one large dataset would allow for the addition of information about the multiple flight paths of an aircraft. This project, however, would require a very large dataset and some extreme computing power.

On a much smaller scale, adding this analysis to a flask app to be included in a blog would be a fun extension of this project. From the app, anyone could enter their flight information and the weather forecast and get a prediction of the on time/delay of their flight.