Predicting Flight Delays

Washington National Airport - Winter 2013

Christine Luongo

General Assembly Data Science 3

December 18 2014

Flight Delays

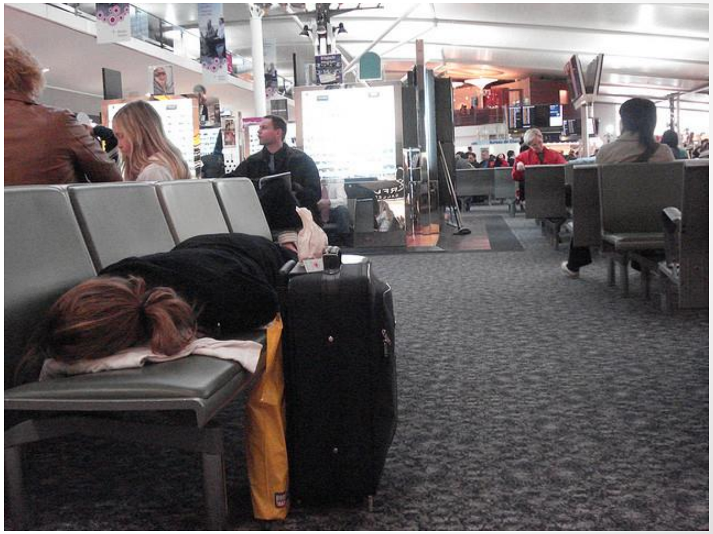
The **goal of this analysis** – Determine what factors can predict flight delay and how we as flyers can avoid delays.

An arriving flight is considered delayed if it arrives at its gate more than 15 minutes later than its scheduled arrival

Are the factors that determine a flight delay that the consumer can control?

On-Time Performance data is publically available on Department of Transportation's website.

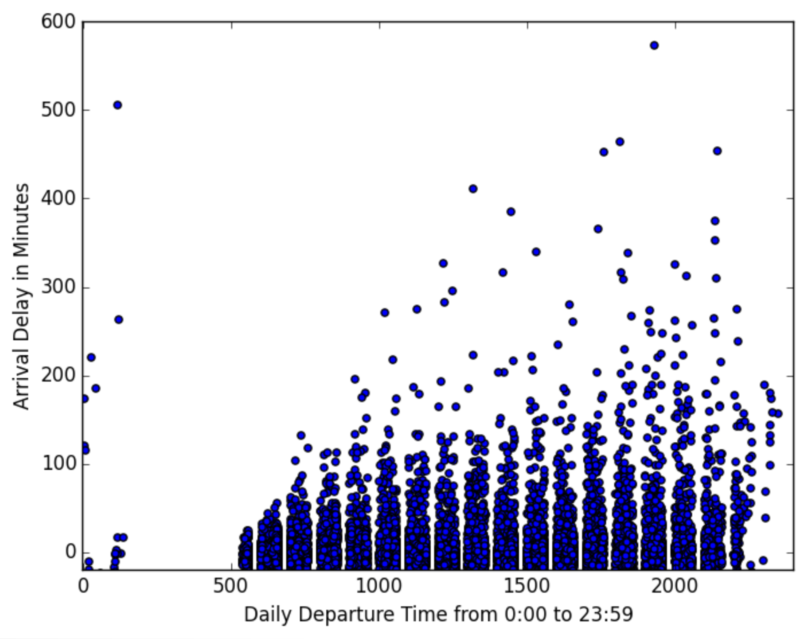
- Includes all domestic flights
- Flight information is updated monthly and goes back to 1987



Acquiring On-Time Performance Data

- **Pulling the Data:** Research and Innovative Technology Administration (RITA) provides publically available data that can be pulled monthly.
- **Make the Data Manageable:** There were 6 Million scheduled flights in 2013 – for the purposes of this analysis I only looked at flights in and out of Washington National Airport (DCA).
    - The dataset was even further subsetted down to the months of November 2013 through January 2014
- **What we were given:** Originating Airport, Destination Airport, Flight time, Distance, Scheduled Time of Flight, Actual Time of Flight, Delay, Carrier.

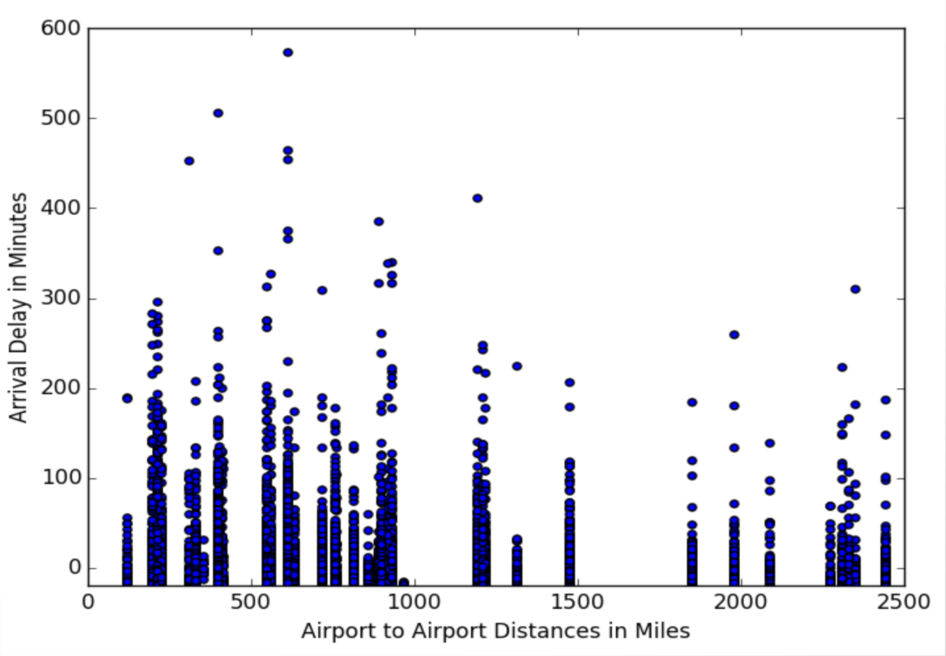
What can we infer from the On-Time Performance Data Alone?



The likelihood of delays increases as the day progresses. Before 7:00PM, the chance of delay is 20%; after 7:00PM the chance increases to 31%

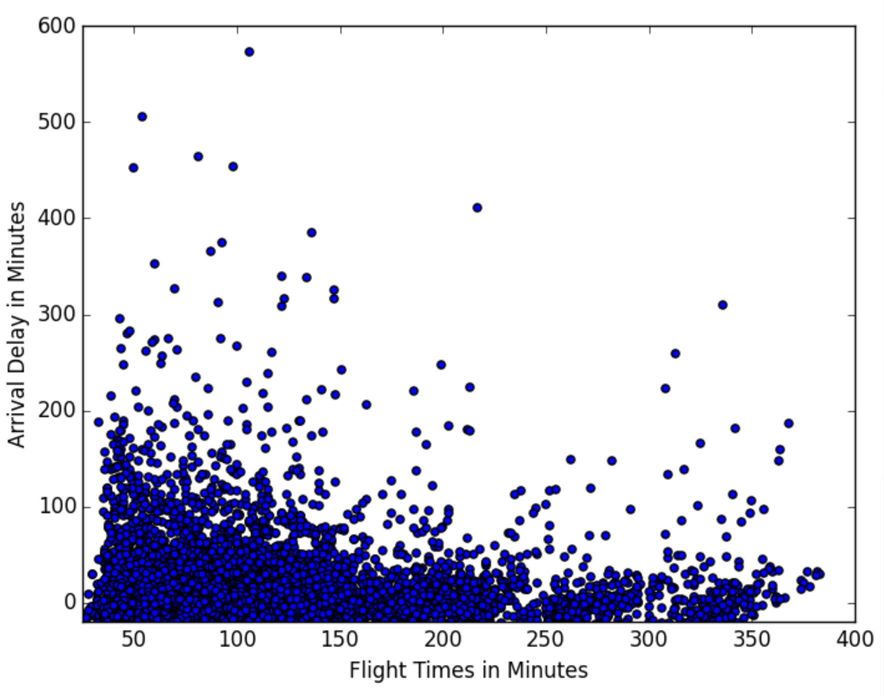
Albeit 77% of flights in/out of DCA fly before 7:00PM

The likelihood of delays increases the shorter the flight.



What can we infer from the On-Time Performance Data Alone?

Seeing a relationship between airport-to-airport distance and arrival delay, we can understand there's probably also a relationship between delays and flight time.



Which Major Carriers are the Worst Offenders?

Twenty-two percent of the flights in our subset are delayed.

envoy

Formerly American Eagle, envoy flights are delayed about 30% of the time.

28% of JetBlue’s flights are delayed.

jetBlue
AIRWAYS

Which Major Carriers have the best track record?

Delta has the best track-record of all the major airlines, with a 17.9% chance of delay.

US Airways and American Airlines (now merged) have 18.3% and 20% chance of delay, respectively.

Merging Other Datasets

What are we Missing? Upon looking at the data, there are other factors not included in the DOT's dataset that might help paint a better picture.

Iowa Environmental Mesonet (IEM) houses environmental data from Automated Surface Observing Systems (ASOS) that collect weather information at the minute level. IEM's website publishes data at the hour-level.

Most airports have ASOS – we can merge the datasets.

We can scrap IEM's website to pull ASOS hourly data for the airports we are interested in.

New Data Fields:

Precipitation

Visibility

Wind Speed

Wind Direction

Temperature

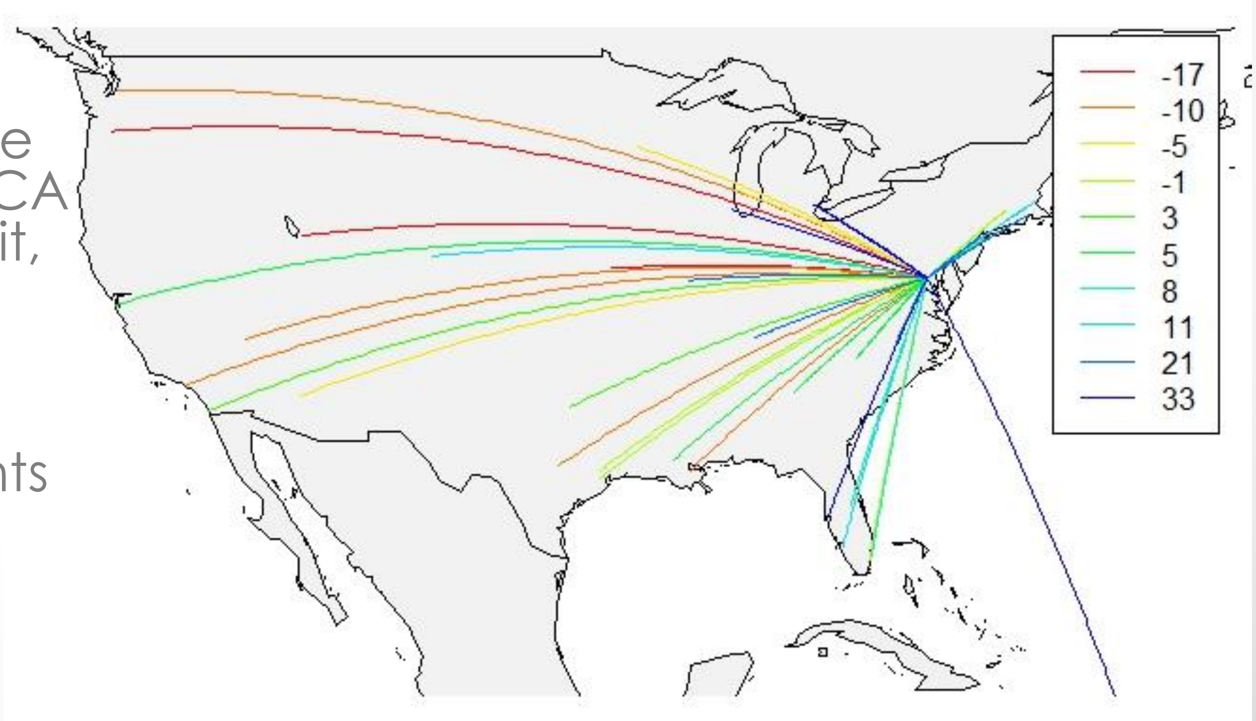
Airport Information

Openflights.org provides GPS coordinates along with other geospatial information.

With geospatial information, we can visualize delays based on Origin and Destination Airports.

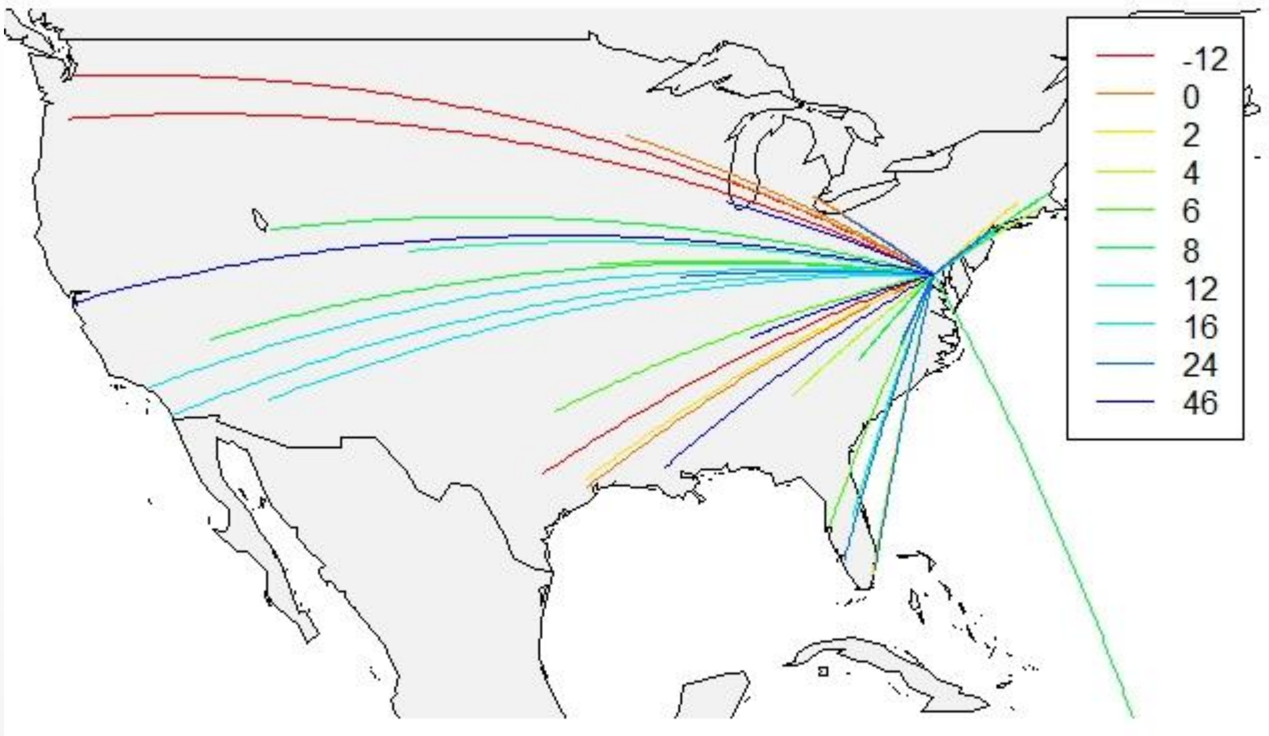
The worst delays are flights going into DCA from Tampa, Detroit, San Juan.

The best on-time performers are flights from Seattle, Portland, Salt Lake and St. Paul



Airport Information Cont.

The best on-time performers out of DCA are flights to Portland and Seattle. Flights headed toward the Southwest, historically are the worst.



Creating New Fields to Supplement our Datasets

- **Bearing:** Using GPS Coordinates, we can determine the direction of the flight.
- **Night Hours:** Flights between 7:00PM and 5:00AM are considered "night flights"
- **Weekend:** Saturday and Sunday are considered the weekend.
- Binary Variables for **Certain Carriers:** Jet Blue, Envoy, Delta, American Airlines, and US Airways.
- **Rush Hour**: a K-means cluster was created to determine whether or not a certain hour-of-the-week was a **commuter flight** or during **rush hour**

Logistic Regression was used to Predict the Chance of Delay

- **Dependent Variable:** DOT's definition of a delayed flight - a binary variable indicating if a flight was delayed or not.
- Of all the models created, the best model was a model that included Departure Delay as an independent variable.
- The significant variables:
    - Visibility for both Airports
    - Bearing
    - Weekend
    - Air Time
    - Airport-to-airport distance
    - Departure Delay

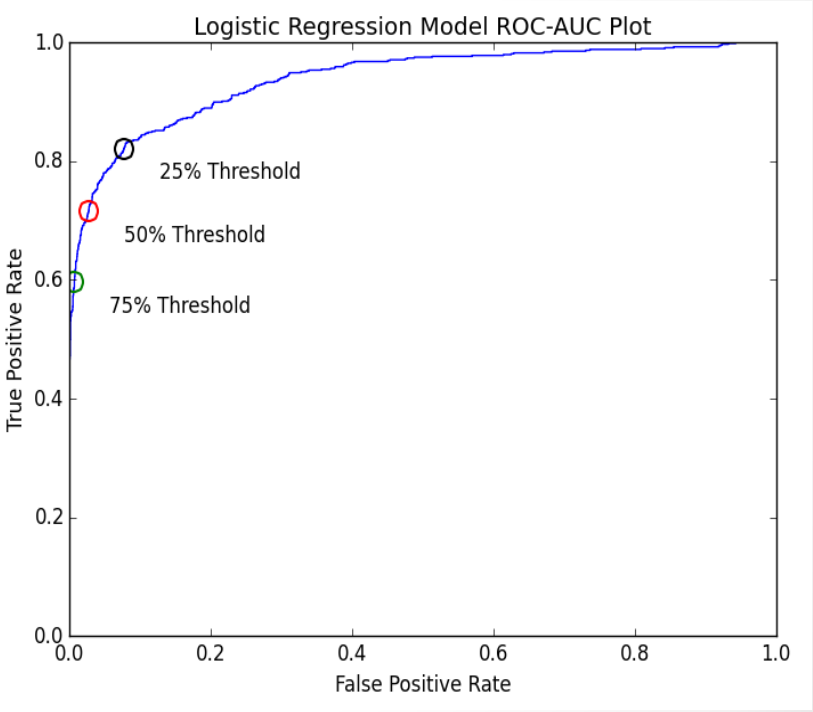
Predicting Flight Delays

A 50% threshold discrimination produced an accuracy of 92% and AUC score of 94%.

The null accuracy is 77.6%

As a flyer, incorrectly predicting an on-time flight will be delayed is less worse than incorrectly predicting a delayed flight will be on-time:

Lower the discrimination threshold to predict flight delays.



What can we do as flyers to avoid delays?

Some factors we can control when booking flights:

- Don’t fly after 7:00PM!
- Avoid Jet Blue and Envoy
- Fly on weekdays – Tuesdays and Wednesdays

Other factors that determine flight delays are not within the control of the flyer:

- Weather
- Departure Delay
- End Destination

Next Steps

- There may be other variables affecting the delay outcome of these flights.
  - Plane type and size may have been useful information
  - Low Pseudo R Squared
- Other types of models could have been analyzed – KNN and Decision Trees.
  - Included in the code