Flight Delay Predictions

...

The Problem

Flight delays take an average of 11 hours to resolve and $1,154 in missed work and out-of-pocket expenses.

The financial impact of delays was $245 million in lost work time for 2014.

Proposed Solution

Using available flight and weather data, create a model that can predict the probability of an individual flight being delayed.

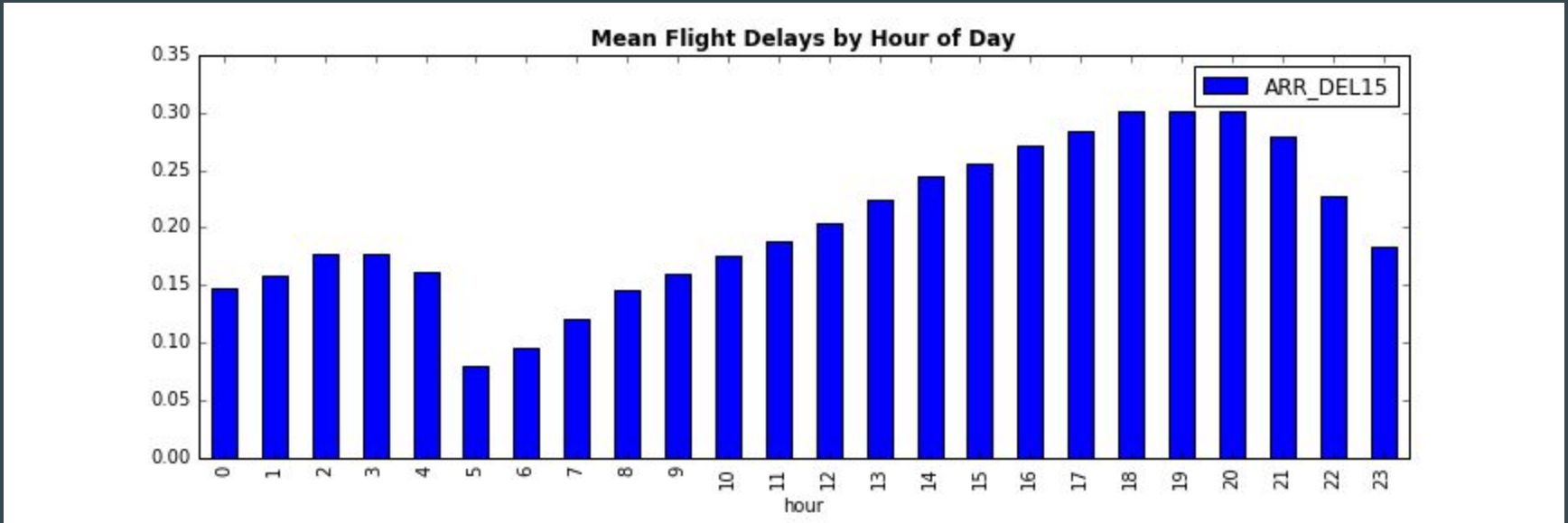
This new metric can be added to flight purchasing sites to help companies/individuals decide which flights would produce less loss in working hours.

Data / Methods Used

- Data from CrowdANALYTIX predict US airline delays contest was used.
  - Flight performance data for ~9 million flights in the US between January 2014 and June 2015
  - Weather information for the US covering the same time period as the flight performance data

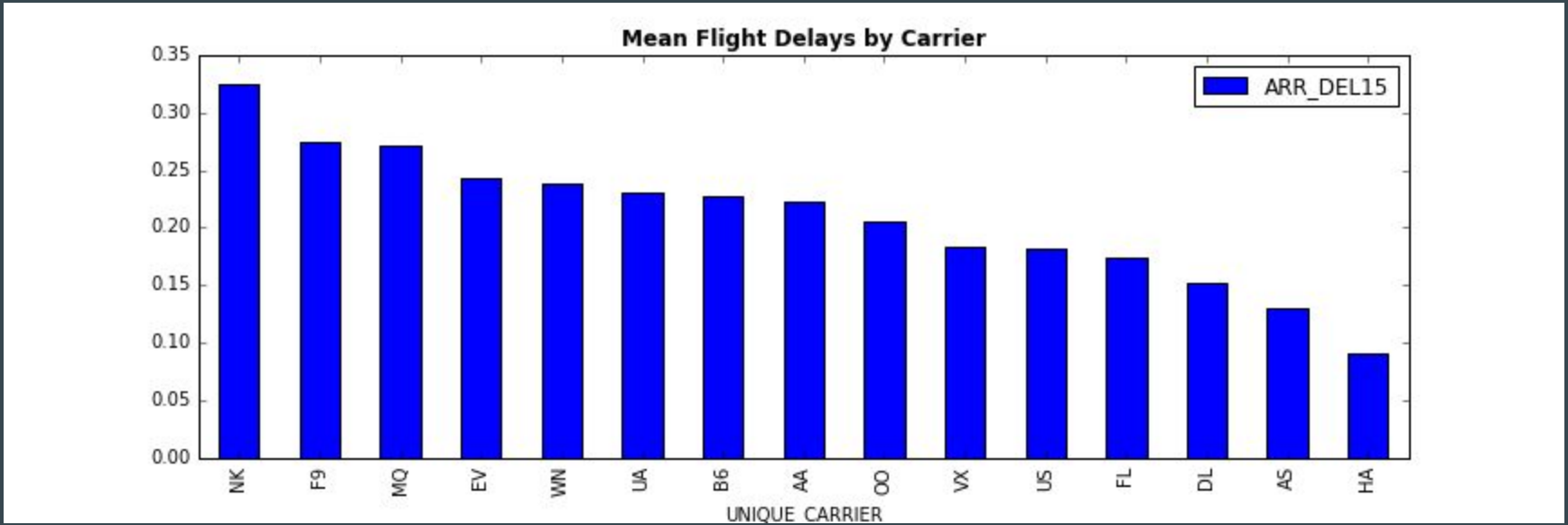
- The data was imported into a Python environment for processing, analysis, and modeling.

Delays by Hour



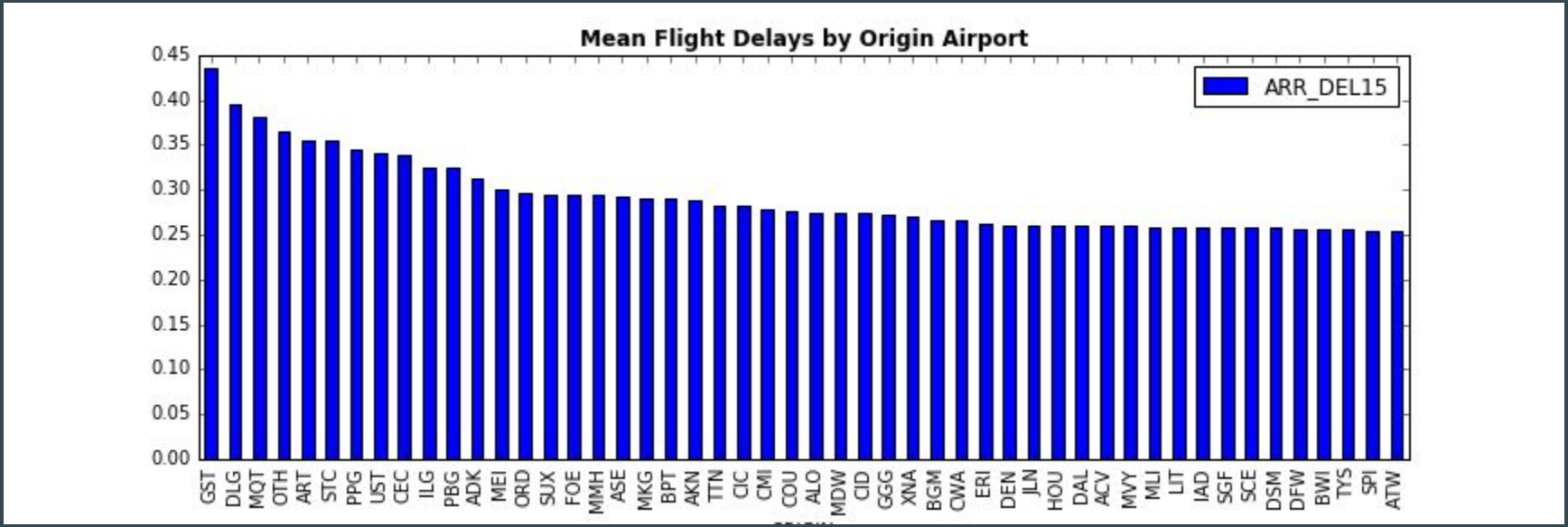
Delays appear to compound on each other through the course of the day indicating that an early morning flight would probably be best.

Delays by Carrier



Some carriers are more prone to delays. Spirit Airlines (NK) show a higher average flight delay while major carriers like Delta (DL) are on the lower end.

Delays by Origin Airport



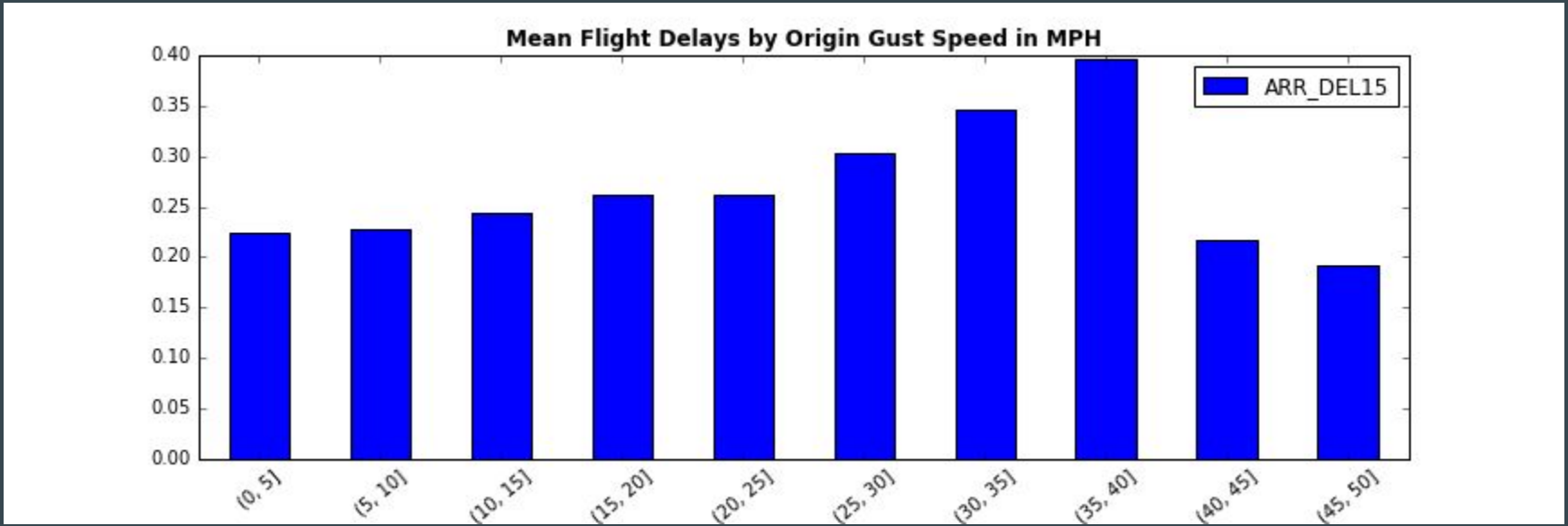
Airports also show differences in average flight delays; however, the choice of origin airport is limiting. Note that Gustavus has the most flight delays, but is a small airport in Alaska which wouldn't offer a choice for travelers.

Weather's Impact

Various variables from the weather set were combined with the flight performance data in order to explore its impact on flight delays.

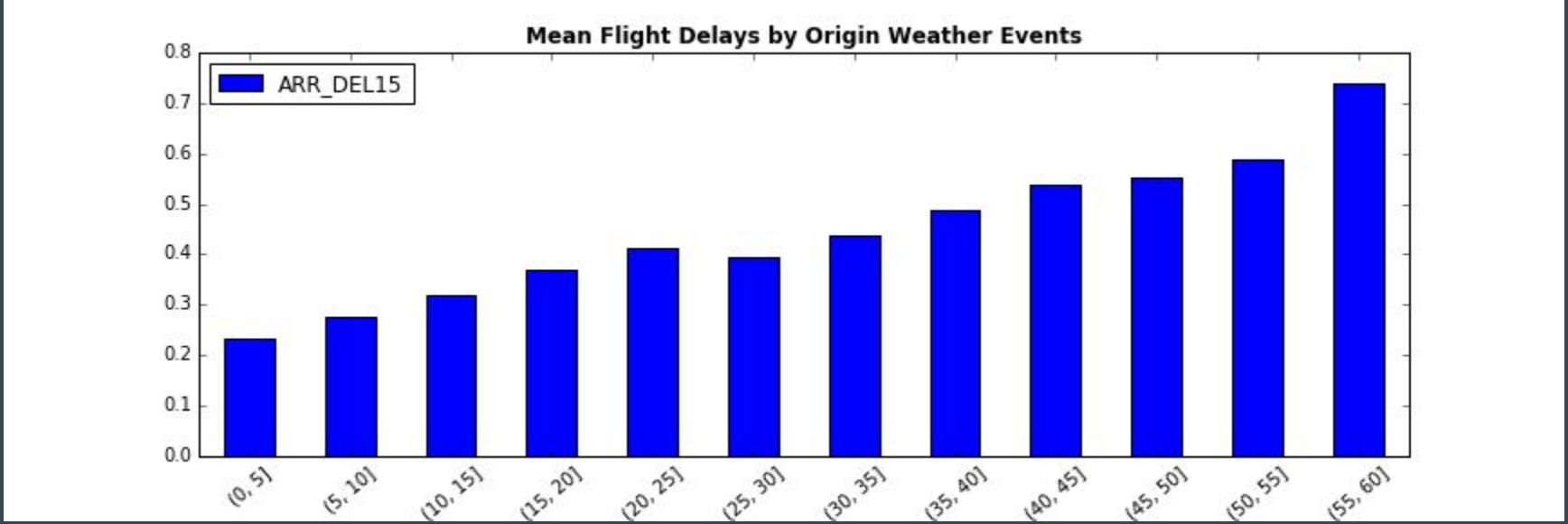
The impact of certain types of weather conditions are shown in the following charts.

Wind Speeds



Gust wind speeds at the origin airport impact the delays only up to a certain amount and then drop off dramatically. Crosswinds as opposed to overall wind speed would be more of a factor for the ability to takeoff.

Weather Events



The weather set contained a variable called events. These events ranged from “Clear” to “Funnel Cloud.” These events were manually given ratings based on the likelihood of causing a delay (e.g. all events involving ice were given larger numbers). The above chart shows that these types of events can be used a predictor of delays.

Model Creation

The records were randomly sampled down in order to make a more manageable data set on which to work.

A model using decision trees was created to classify the flight data into either delayed or on-time based on 70% of the smallest dataset. This model was then tested on the remaining 30% of the dataset showing an accuracy of 73% in correctly predicting if flights were delayed or on-time.

Proposed Next Steps

1. Run the model using much larger sets of the flight and weather data.

2. Refine predictions of weather for future flights based on geography, time of year, and past weather.

3. Incorporate the model into travel sites or on a separate site for consumers/business users to lookup delay probablities.