5. Airline Delays

Data description

The dataset consists of all flights in the US from 1987 until 2012. The original dataset comes from Revolution Analytics website (see AirlineData87to08.tar.gz). There are 143 million rows in the dataset and 44 columns in the dataset. The columns in the dataset are described in full in this text file.

Problem statement

To understand the factors that lead to airline delays, and the impact. The dataset only spans 1987 to 2008, so instead of using it to perform prediction, the challenge is to *model* delays.

The front-end should allow for interactive visualisation of the data, the factors that lead to delays, and their interpretation. The entire dataset is 120GB (when compressed), so one challenge is how to query the data, or even store it so that the front-end will not be laggy.

From the modeling perspective, it is unlikely that a single model works well over the entire period. The impact of factors change over time. Another challenge is to augment this data with other internal and external factors if necessary. Examples of external factors are weather, holiday periods, airport capacity, plane types, etc. An example of internal factors would be *cascading delays*: perhaps a delay at a hub airport leads to delays at other airports - is it possible to build a model that takes this factor into account?

Personas

Terry is an air-traffic controller at Federal Aviation Authority. He monitors air-traffic delays across the nation. The abovementioned dashboard above is one of the tools he uses to understand how delays begin (indications that a delay is beginning, and how it is propagating). The dashboard allows him to revisit major delays in history and "replay" them to gain intuition about the expensive phenomenon of airline delays.

Final presentation

you have pre-computed several models
- show what is good about this model
- show interpretations of this model, what happens when i use this model for different time and space

Both teams should present their work to the end-user.

Additional information

The dataset was used in a visualisation competition at a conference 2009 (JSM data expo). Here are results related to the posters that were created from this dataset. Please look through them to gain an idea of the types of visualisations you can make, and the types of analysis you can do:

- http://stat-computing.org/dataexpo/2009/posters/dey-phillips-steele.pdf
- https://www.usna.edu/Users/math/dphillip/jsm.pdf
- http://stat-computing.org/dataexpo/2009/posters/sun.pdf
- https://blog.revolutionanalytics.com/2009/09/analysis-of-airline-performance.html
- Search results page:
    - https://duckduckgo.com/?q=stat-computing.org%3A+2009+airline+delay+posters&t=ffab&atb=v339-1&ia=web
- The consulting company SAS has a white paper on modeling airline delays that can give some ideas.