Flight Delays and Cancellations:

What does the data tell us?

Assumptions about the data

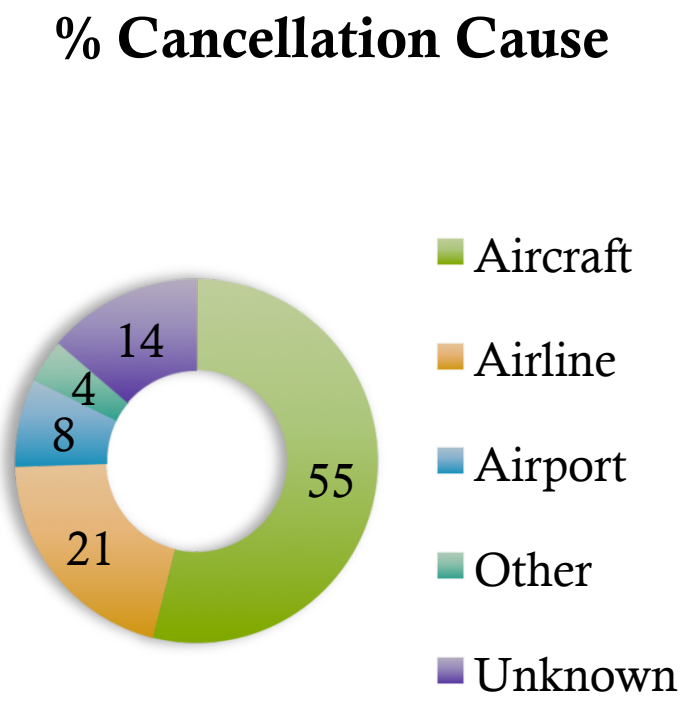
- Assume all numbers are reliable, accurate and timely
- Missing ‘actual departure time’ means flight was cancelled
- Missing ‘actual arrival time’ means arrived on schedule
- Other missing values replaced with either
  - ‘Missing’ if a non-numeric value
  - A default value e.g. average temperature

How data was used

- Used METAR weather data for origin and destination airports (within nearest hour)
- Calculated backlog of recent delays and cancellations
- Calculated average delays and cancellations for
  - Origin and Destination airports
  - Airline
  - Aircraft model

Predicting Cancellations

- Binary classification problem, with class imbalance
  - Samples from under-represented classes are given more weight
- Cancellation rate correlated with
  - Aircraft model 55%
  - Airline 21%
  - Airport 8%
  - Other 4%



Predicting Delays

- Regression problem (instead of classification) if we what to know how long will be the delay instead of if there will be a delay
  - Long delays are worse than short delays
  - Length of departure delay in is the single most important factor in length of arrival delay (62% explained variance)
  - Length of departure delay is hard to predict (31% predictable)
- More investigation needed
  - Seasonal factors (time of day, month of year, major holidays)
  - Classification of short (<10min), medium (< 2hrs) and long delays

Customer Experience Recommendations