Flight Durations and Delays: How Long Will Take My Flight?

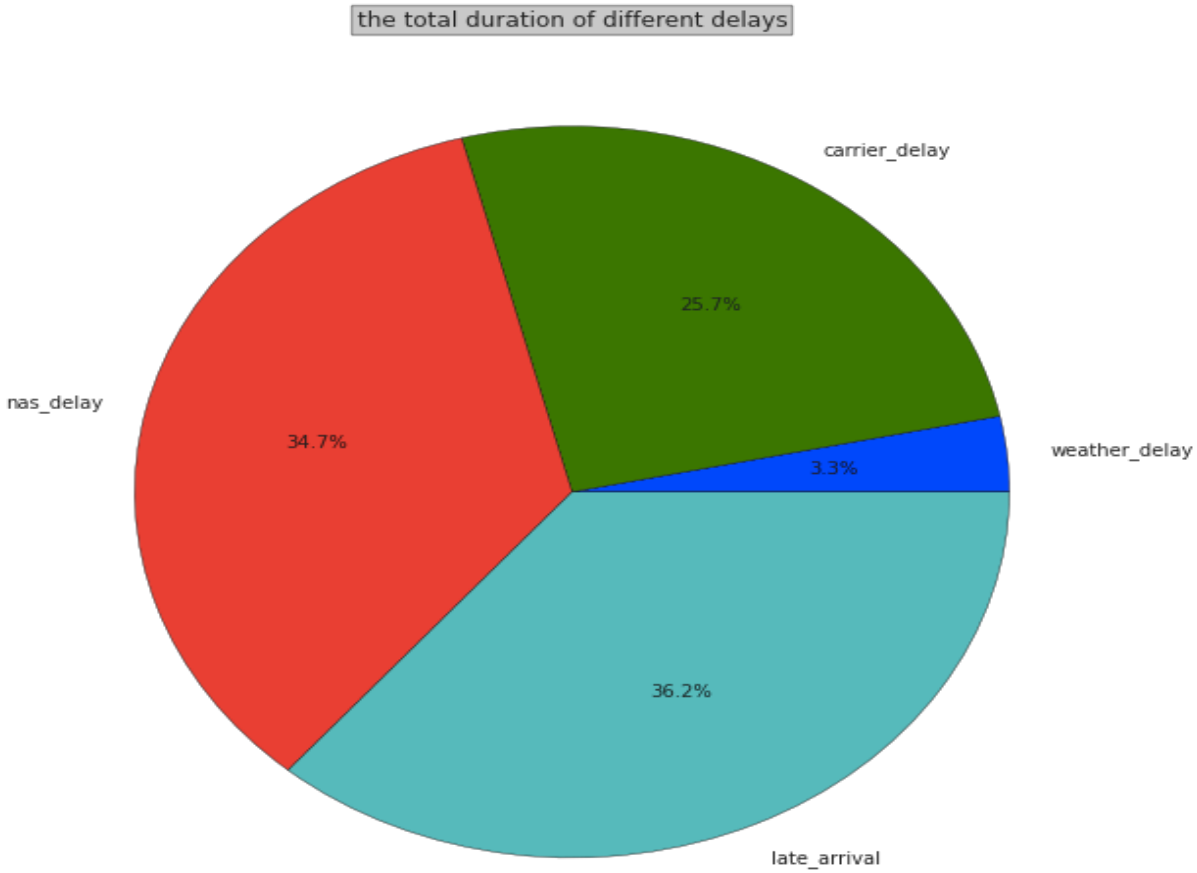
Anna Hughes

Model Architecture

- DATA
  - Most of data is downloaded from RITS: lots of files containing information about the flights from Aug 2011 to Sept 2014, planes specs from FAA
  - Messy, heavy (4GB)
- TOOLS
  - S3 for data storage
  - EC2 for data processing
  - SQL (sqlite) for merging tables

Flight delays

- 84% of the flights are on time, the rest are delayed

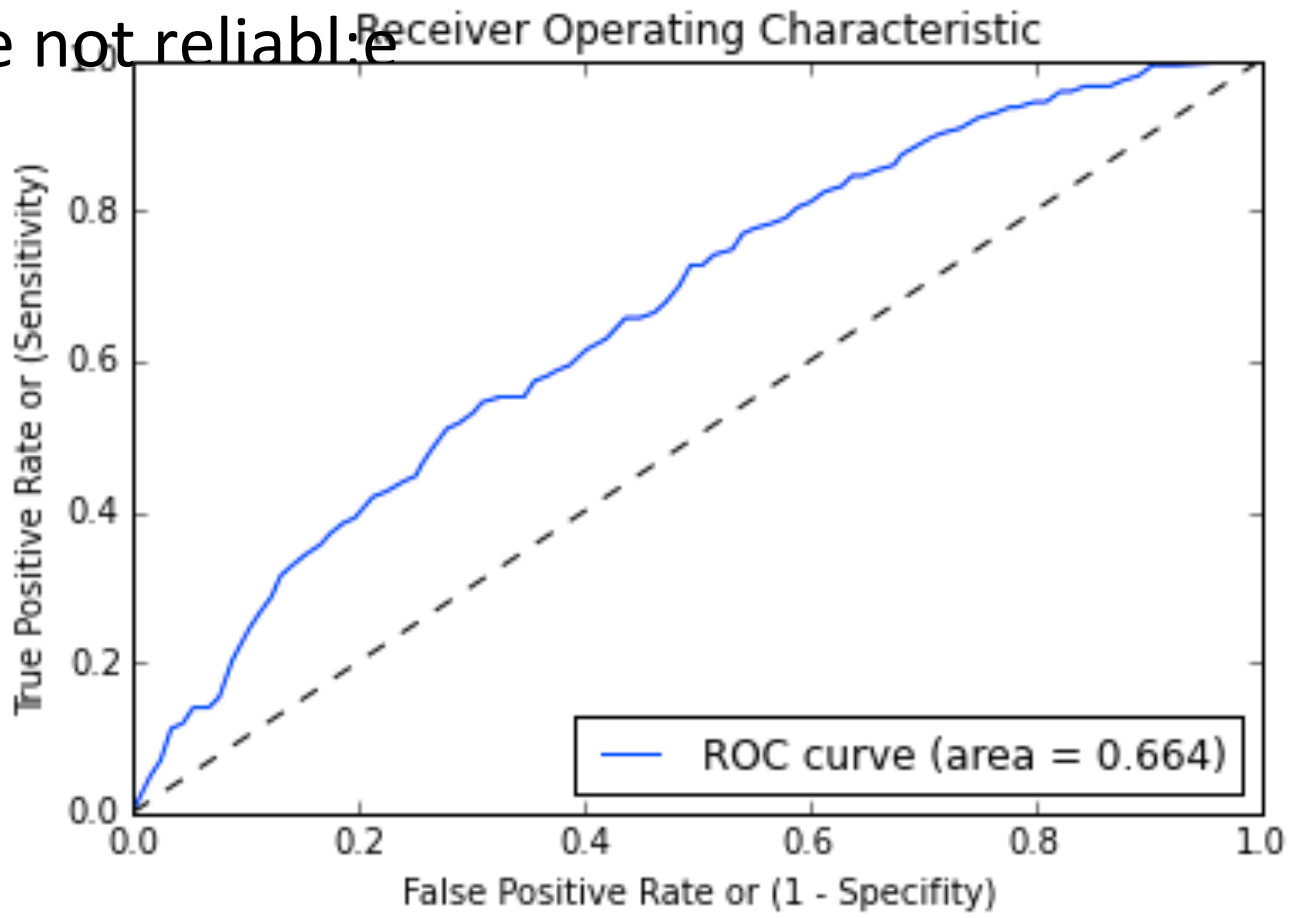


The Goal

- Initial goal
  - predict the probability of the flight being delayed
- Problem
  - Saying ‘every single flight is going to be on time’ is a better model than KNN, Naïve Bayes, Logistic Regression, Decision Trees and Random Forests
- Why?
  - Because the airlines include the ‘pad’ for each flight to make sure they are doing well with delay stats

The Goal: Problem

- When having a cross-validation score 0.84 the results are not reliable.

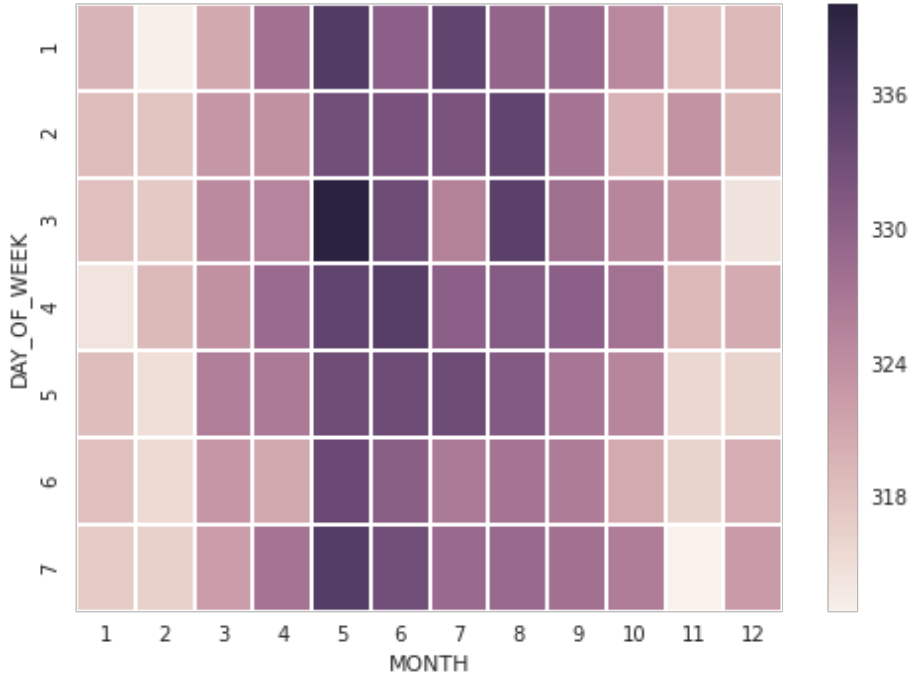


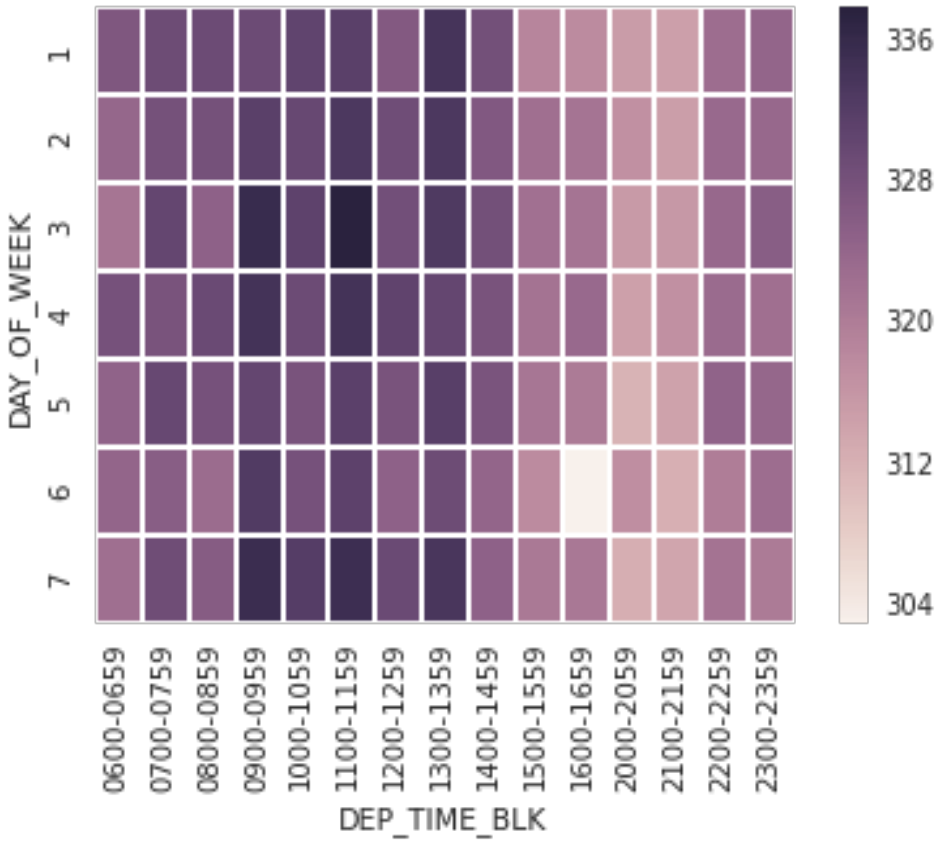
New Goal

- New goal
  - Predict the actual duration of the flight
- Inputs
  - Date (Month, Weekday, Time)
  - Aircraft specs (Year, Manufacturer, Model)
  - Flight number (includes the airline)
  - Scheduled duration of the flight

When to Fly

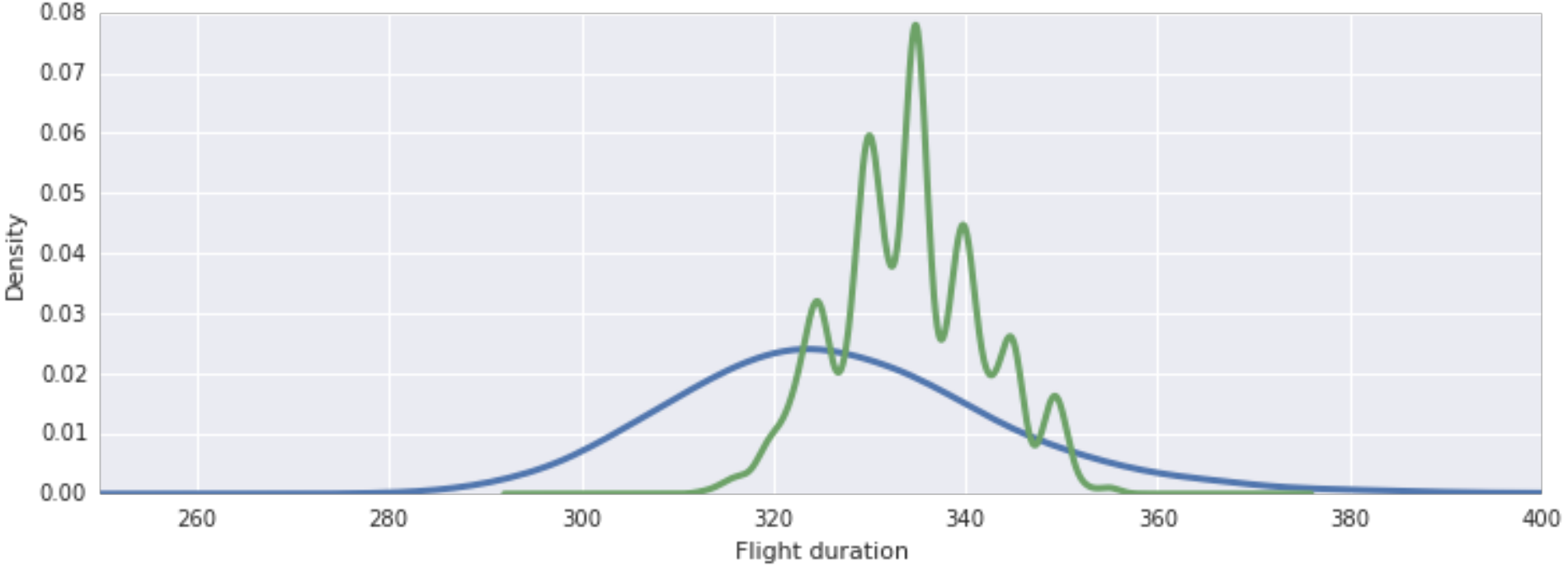
- Winters are better than summers
- On the evening is faster than on the morning





How Long Can Be a SFO JFK flight?

- Actual flight durations have a normal distribution (BLUE)
- Schedules flight durations have a normal distribution (GREEN)



Results

- My results (above) for prediction the duration of the flight are not that great, but they are more accurate than the actual scheduled duration

OLS Regression Results

Dep. Variable:	ACTUAL_ELAPSED_TIME	**R-squared:**	0.106
Model:	OLS	**Adj. R-squared:**	0.094
Method:	Least Squares	**F-statistic:**	8.821

OLS Regression Results

Dep. Variable:	ACTUAL_ELAPSED_TIME	**R-squared:**	0.059
Model:	OLS	**Adj. R-squared:**	0.059
Method:	Least Squares	**F-statistic:**	1126.
Date:	Thu, 18 Dec 2014	**Prob (F-statistic):**	1.98e-239