Predicting flight delays

Kihoon Sohn | Data Scientist | July 17, 2018

Table of Contents

Project objective:

Predicting flight delays in Washington metro area with aircrafts models & makers

Data Science Problem Statement

Can aircraft models or makers predict flights delays in the local market?

- If so, compared them with the airlines fleets in possession.
- Which would lower the risk of the delay and its cost? → Diversified or simplified company’s fleet inventory

Research

Already existed many prediction models in the flight delays

Mr. Fabien Daniel / Kaggle Tutorial

State of the art tutorial on predicting delays with strong visualization and codes, which I referred and brought in my work a lot.

Models:

- Linear/Polynomial regression
  - Univariate / Multivariate
  - MSE - train 89.55 / test 74.8

Source: Kaggle tutorial of Fabien Daniel

Mr. Scott Cole & Tom Donoghue/ Ph.D Students, UC San Diego

Precise analysis on flight delays and prediction model built.

Models:

- Classification for delay or no delay
- Logistic Regression for each airports
- AUC = 0.689

Source: Scott Cole’s webpage

Data Sources

Data Sources

On-Time Performance dataset

Pulled monthly data on domestic passenger flights between 2013-2017; 29M data points

Features:

- Date, Carrier, Airport, Delay(mins)
- Cause of delays(Carrier, Weather, National Air System, Security, Late Aircrafts)

Flightradar24.com Aircrafts information

Webscraped with Selenium on aircraft details with tail # provided from OTP data. Around 6,000 unique tail # to scrap

Features:

- Aircraft type, manufacturer, age

Computational reasons, subset data into smaller piece

29M

Total domestic flights

2013 - 2017

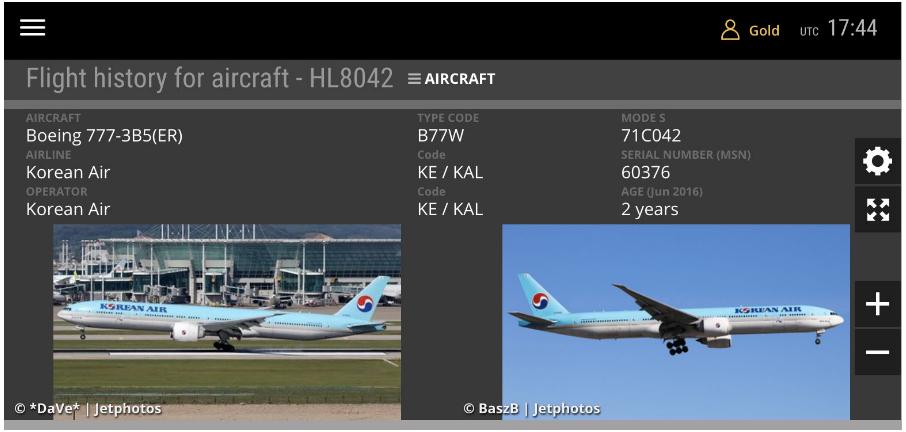
2.2M

Washington Metro

DCA, BWI, IAD

With tail no, aircrafts model/maker/age can be obtained

Tail #, the unique identifier of each airplane, provided in the on-time performance dataset



Webscrapped from flightradar24.com for the new variables

Tail #	Maker	Airlines	Type	Age	Delivered
HL8042	Boeing	Korean Air	B77W	2 years	June 2016

Selenium-ed 6K unique aircrafts' tail number

and then, left-joined them with the dataframe by tail number

Majority model/builder

- Boeing
- 737 series(B737,B738,B733,B739)

Inference / Limitation

- Will aircraft builders, types be a good predictors
- However, too many values are unknown.

	value	counts	(%)
0	Boeing	120992	0.5188
1	Airbus	41009	0.1759
2	Unknown	26109	0.1120
3	Embraer	25002	0.1072
4	Other	12642	0.0542
5	Bombardier	7443	0.0319

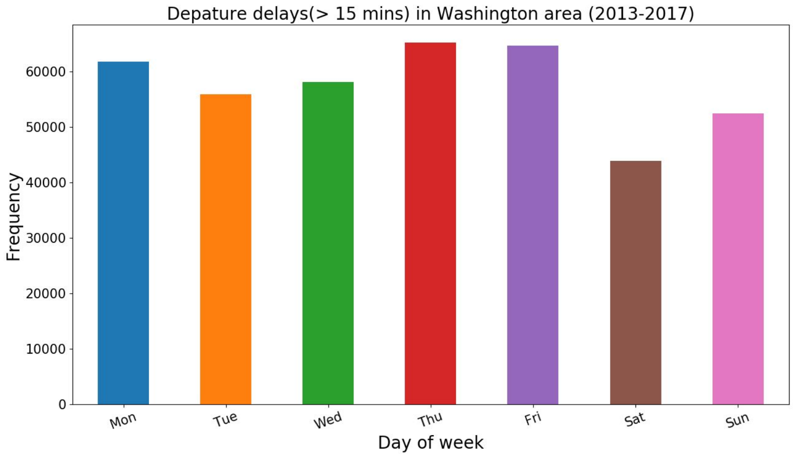
	value	counts	(%)
0	B737	63941	0.2464
1	Unknown	52705	0.2031
2	Other	32980	0.1271
3	B738	22535	0.0868
4	A320	20904	0.0806
5	A319	16970	0.0654
6	E190	14185	0.0547
7	B733	13858	0.0534
8	B739	9313	0.0359
9	E145	6833	0.0263

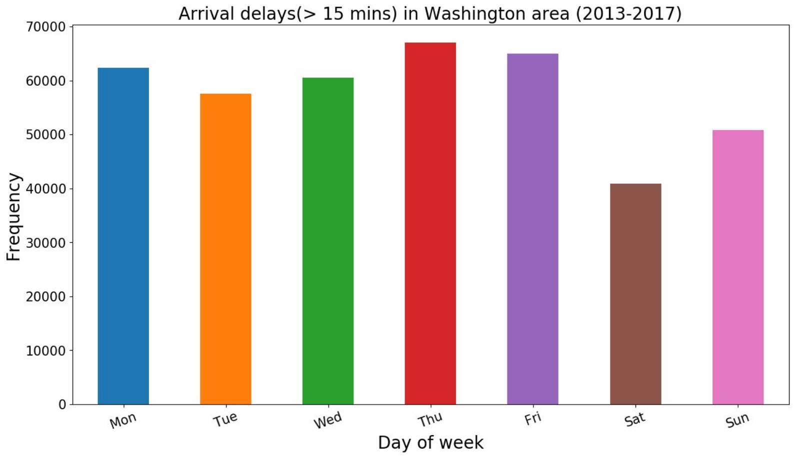
Findings

-Basic Exploratory Data Analysis

Flight delays in Washington Metro area (2013-2017)

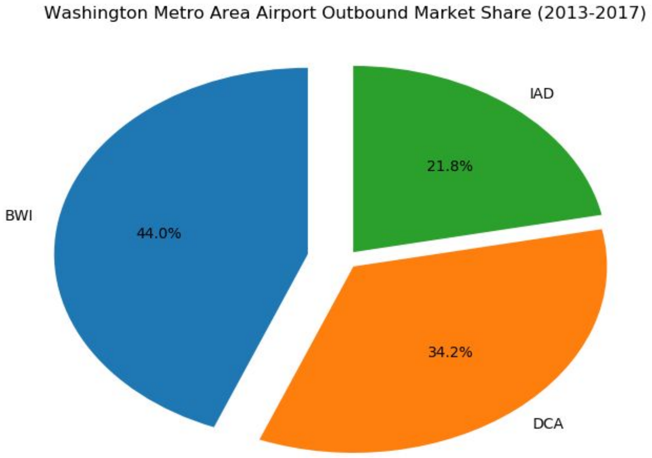
- In Thursday and Friday expected to have higher delays in both departure & arrival
- Saturday and Sunday in lower expected delays

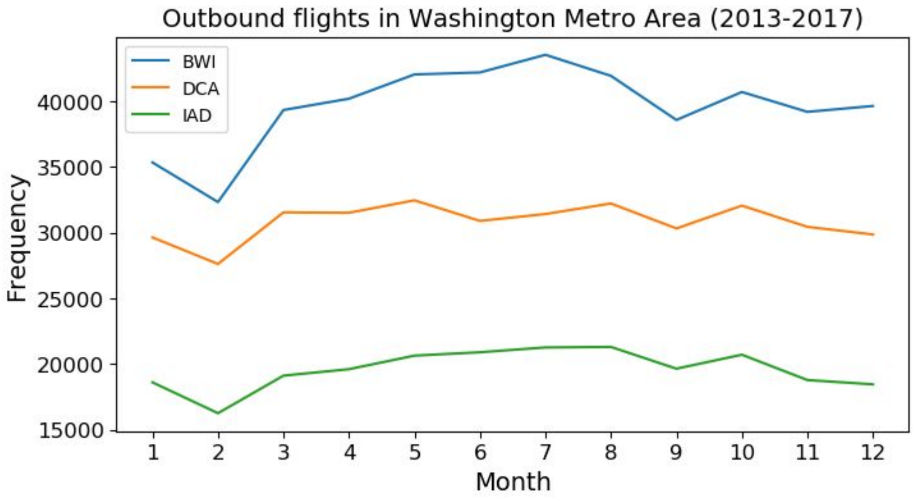




Overview: Washington Metro Area

Marketshare - Flight counts by the airport

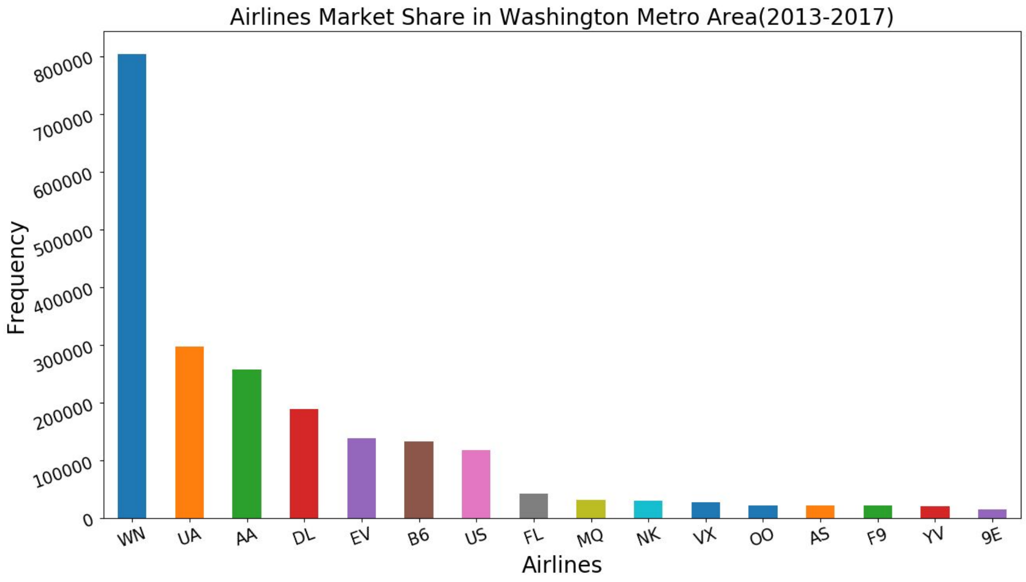


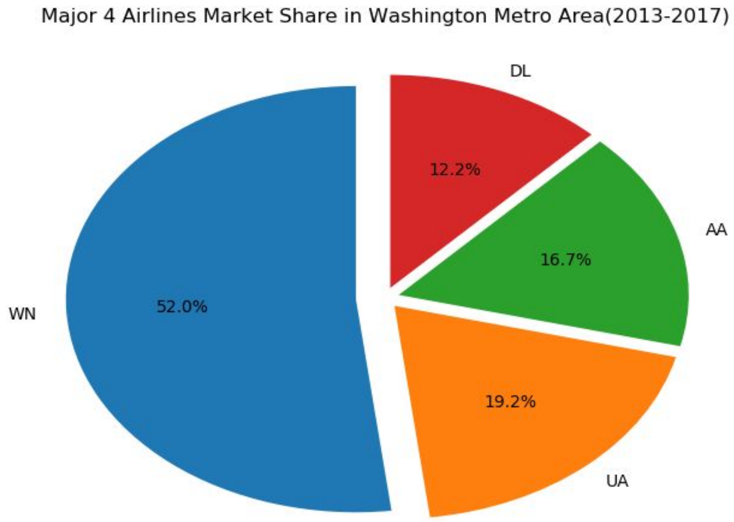


- Baltimore/Washington International(BWI)
- Ronald Reagan Washington National (DCA)
- Washington Dulles International(IAD)

Overview: Washington Metro Area

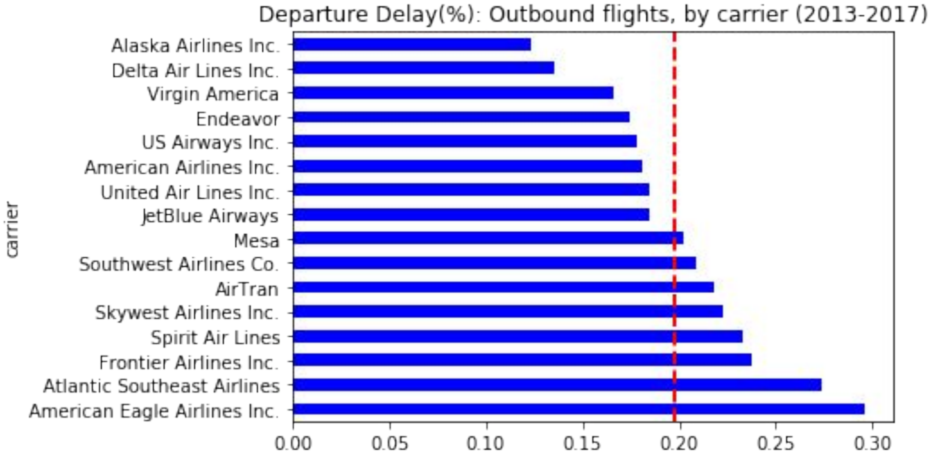
Marketshare - Flight counts by the airline

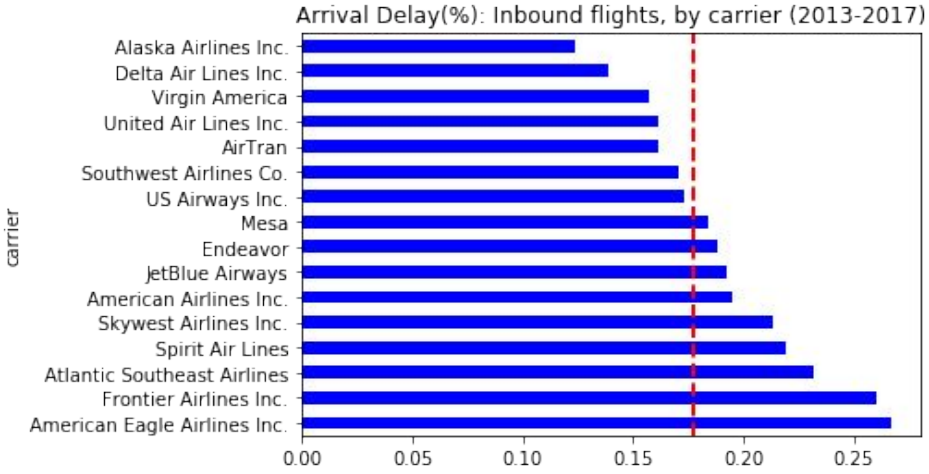




- Southwest(WN) > United(UA) > American(AA) > Delta(DL)

On-time performance in Washington (2013-2017)





- Best: Alaska, Delta
- Worst: American Eagle, Atlantic Southeast(Skywest subsidiary for DL, UA, AA), Frontier, Spirit

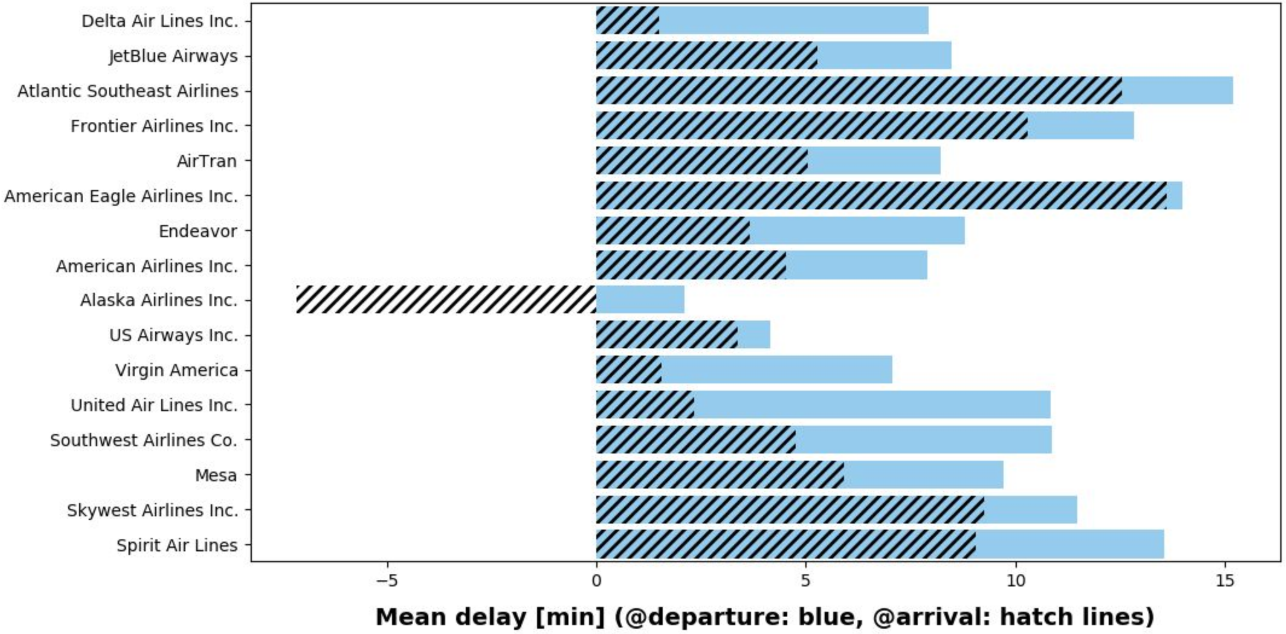
Mean outbound/inbound delay by airlines

Findings

- Alaska arrives earlier than its scheduled arrival time
- No airlines exceeded the length of arrival delays to departure delays.

Inference

- Fly fast to make up time
- Less traffic in arrival than departure

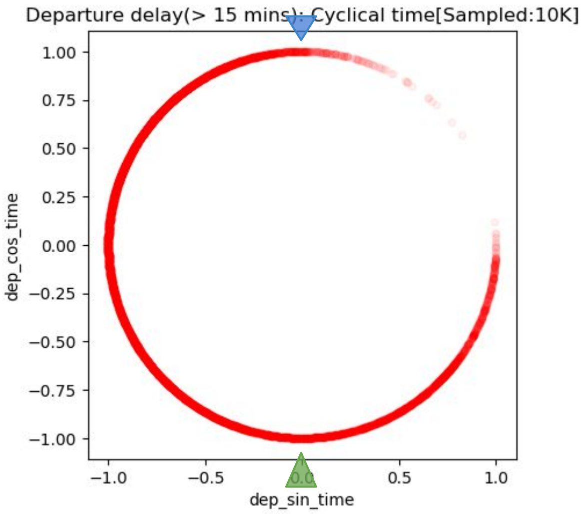


Plotting code source: Kaggle tutorial of Fabien Daniel

Vectorize time for variables

Cyclical - sin and cos time

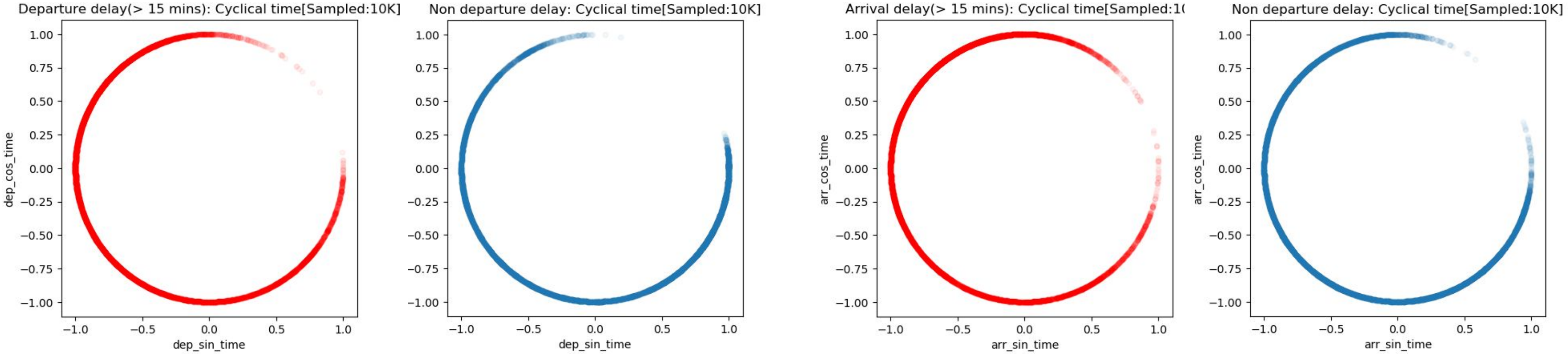
- Imagine the plotted circle as 24-hours-clock.
  - Green marker - Noon
  - Blue marker - Midnight
- Each dot(transparency 30%) represent single flight record, darker the more flights in that time period of the day, brighter the less.
- How to read: there is nearly no departure flights in between 3 am - 5 am



Cyclical time on delays/non-delays

Vectorized time

- Randomly sampled 10,000 with sin/cos time on departure/arrival delays in Washington area.
- Red represented each time for delayed-flights made, as blue displayed scheduled time
- Findings: delayed flights made late operations in the airport between midnight-3am



Modeling

Modelings

Features

- Cyclical Cos/Sin time
- Dummified
  - Day of week
  - Carrier
  - Departure Airports
  - Arrival Airports
  - Aircraft type
  - Aircraft maker

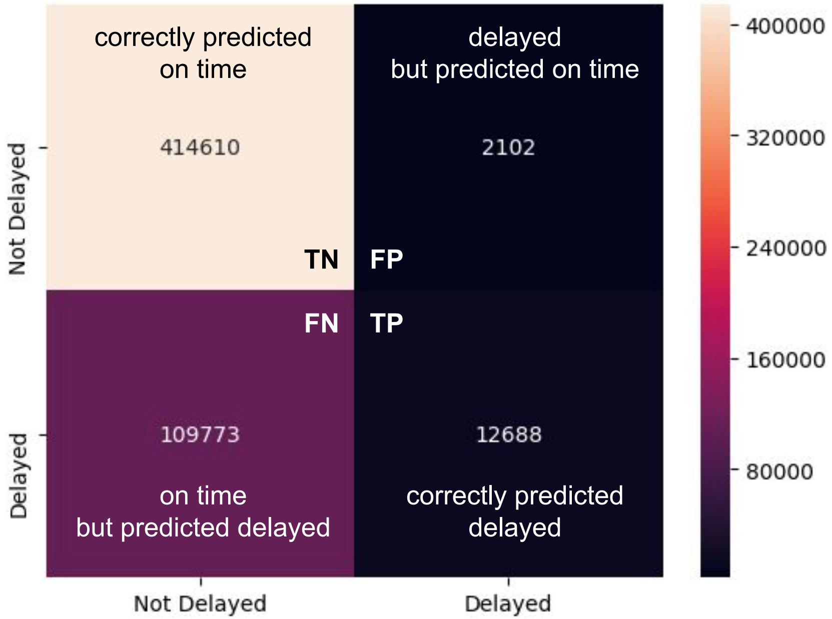
Models

- Models
  - Logistic Regression
  - Random Forest
- Tools
  - GridSearchCV
  - SMOTE

Results

- Logistic Regression
  - Train 0.7734
  - Test 0.7731
- Random Forest
  - Train 0.7916
  - Test 0.7901

Confusion Matrix (Delay=1, No Delay=0)



- Accuracy - 79.25%
- Precision - 85.78%
- Percent that was truly delayed out of all predicted to be delayed(Recall): 10.36%
- Percent that was truly on time out of all predicted to be on time(Specificity): 99.49%

Next steps

- Find more accurate data source or clean it in tail number
- Build up for Neural Network
- Bring time series analysis
- Add more variables: weather
- Apply to bigger/different angle
    - Hub airport by airlines
    - Top 20 most frequent route
    - Top 20 busiest airport

Thank you!

GitHub for the project: http://bit.ly/2LHG01P