Delayed on the MTA

Seth Kaufman

Outline

- Problem/Idea
- Data Collection and Cleaning
- Model Building
- Results
- Next Steps/Goals

Problem

The MTA is constantly having delays!

- This disrupts the flow of traffic, upsetting New Yorkers daily and damaging the local economy.

- Per the NYT, delays just during the morning rush cost about $307 million in lost work time annually.

- Can we find some causes of delays, predict the worst locations, and determine which stations need the most investment from the MTA/NYS and Federal Government?

Idea

- Analyze data of delays from the past decade
- Analyze highest frequency station usage to determine where money should be spent
- Build a model that can determine stations that are most likely to have/cause issues with delays
- As proof of concept, this will be only done for IRT/Division A trains

IRT Lines

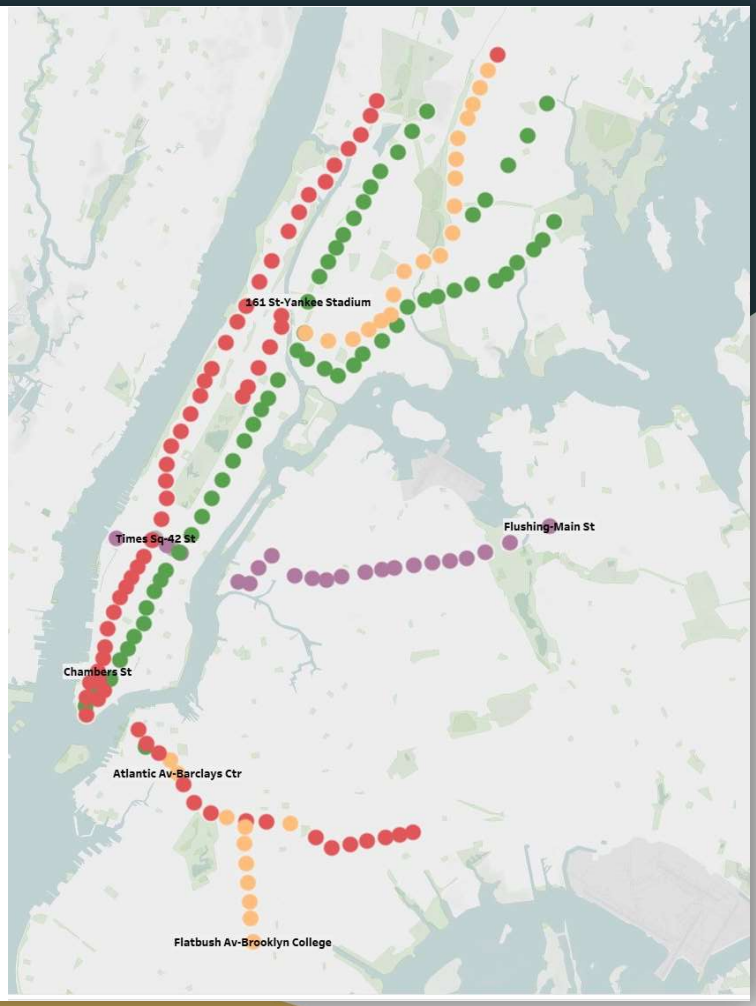
1/2/3 - Red

4/5/6 - Green

7 - Purple

S - 42 St Shuttle

2/3/4/5 combinations - Orange



Data

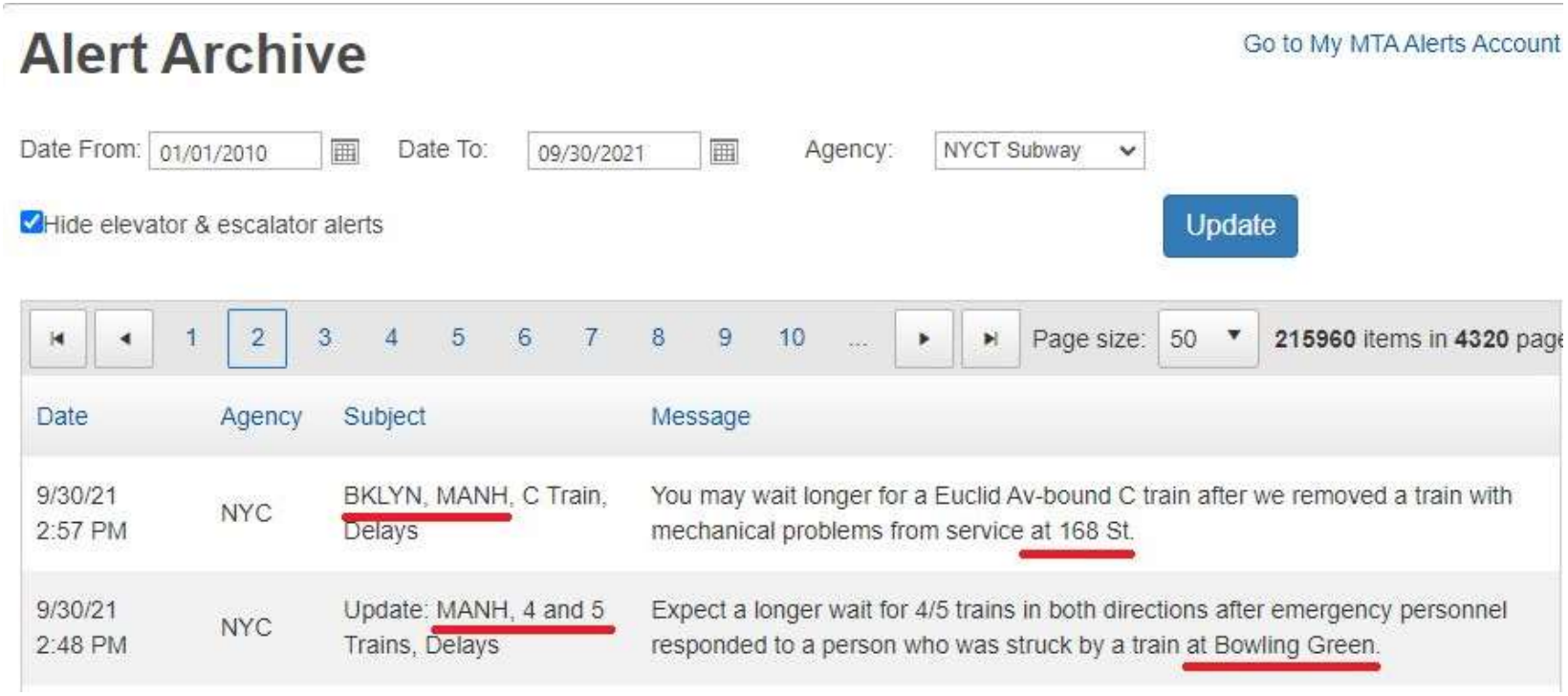
MTA Alerts Archive (2010-2020)

NOAA (2010-2020)

MTA Turnstile Data (2015-2020)

MTA Ridership Data (2015-2020)

What is an alert?



Alert Archive

Go to My MTA Alerts Account

Date From: 01/01/2010 Date To: 09/30/2021 Agency: NYCT Subway

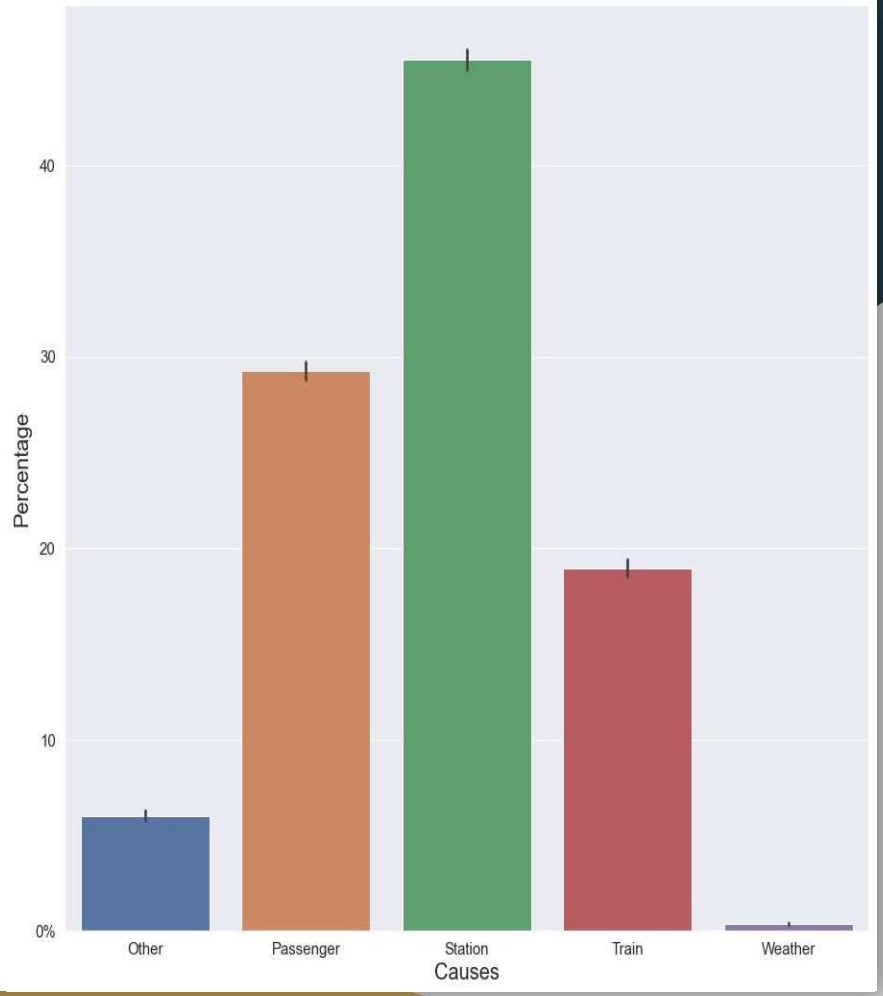
Hide elevator & escalator alerts Update

1 2 3 4 5 6 7 8 9 10 ... Page size: 50 **215960** items in **4320** page

Date	Agency	Subject	Message
9/30/21 2:57 PM	NYC	BKLYN, MANH, C Train, Delays	You may wait longer for a Euclid Av-bound C train after we removed a train with mechanical problems from service at 168 St.
9/30/21 2:48 PM	NYC	Update: MANH, 4 and 5 Trains, Delays	Expect a longer wait for 4/5 trains in both directions after emergency personnel responded to a person who was struck by a train at Bowling Green.

Features

- Time (rush hour, weekday or weekend)
- Weather conditions
- Cause of delay
- Location
- Direction of train



Current model/Results

Logistic Regression:

Overall Accuracy: 67.72%

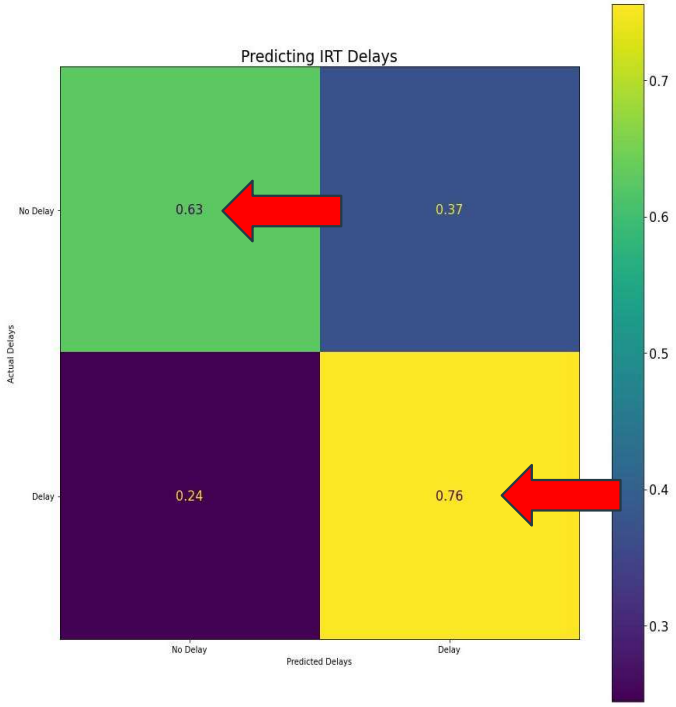
Predicting Minority Class (Delays): 65.63%

Bernoulli Naive Bayes:

Used on SMOTEd data

SMOTE Accuracy: 67.82%

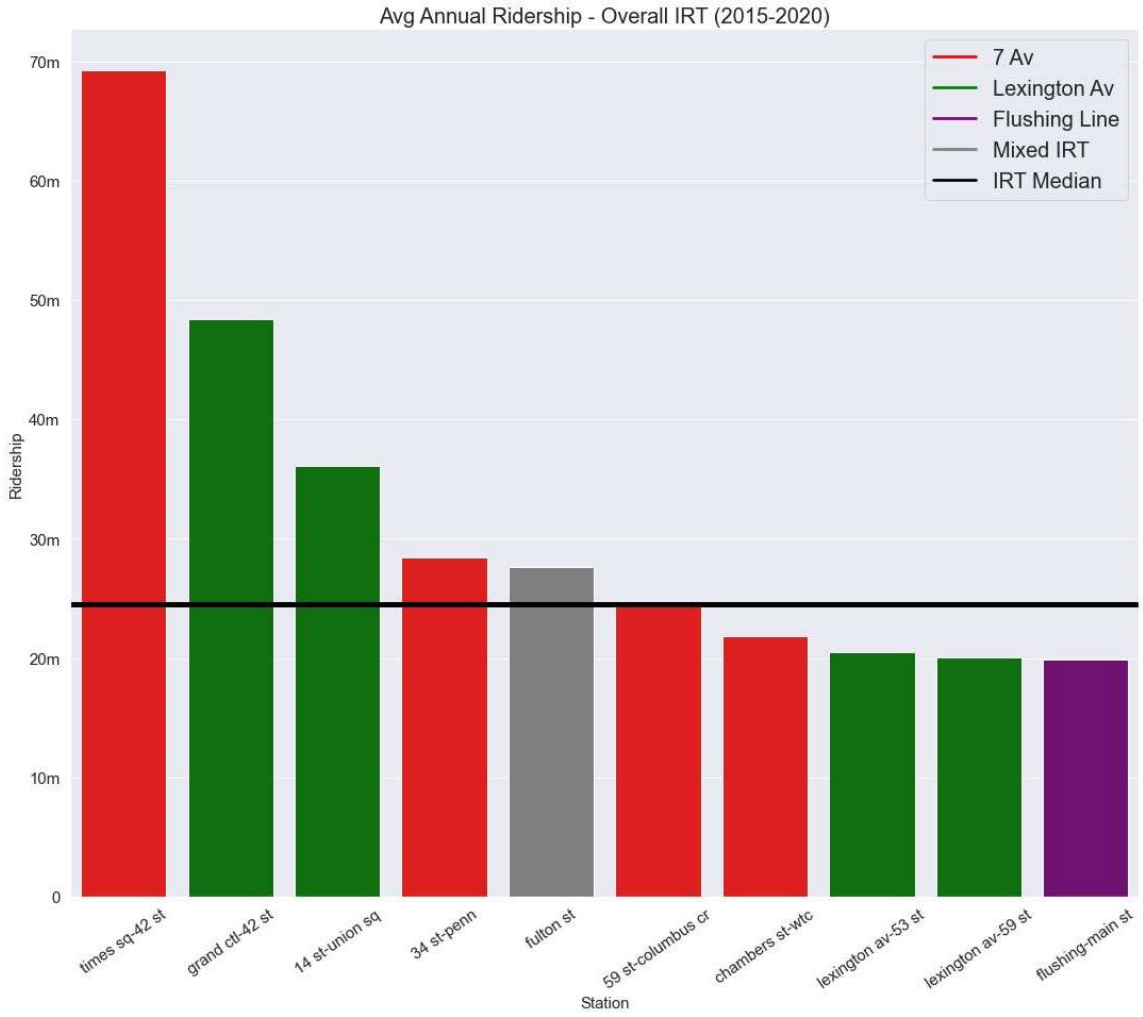
SMOTE F1: 65.07%




```
Pipeline(steps=[('bnb', BernoulliNB(alpha=0))])

Overall Accuracy: 69.05%,
F1 Score: 59.56%

Overall SMOTE Accuracy: 67.82000000000001%,
SMOTE F1 Score: 65.07%
```


Avg Annual Ridership - Overall IRT (2015-2020)
7 Av
Lexington Av
Flushing Line
Mixed IRT
IRT Median
70m
60m
50m
40m
30m
20m
10m
0
Ridership
times sq-42 st
grand ctl-42 st
14 st-union sq
34 st-penn
fulton st
59 st-columbus cr
chambers st-wtc
lexington av-53 st
lexington av-59 st
flushing-main st
Station

Deploying the model on the most used stations

- Fulton St, Grand Central, and 14 St-Union Sq are the stations most likely to cause delays

- Improving Signal system at these stations should improve general performance among IRT lines

Immediate Steps

- Create more features such as inputting outdoor vs non-outdoor station
- Deploy model
- Revisit turnstile data - create a commuter score
- Recreate main points of project for all stations in the Subway system

Thank you!

GitHub
SethKauf

LinkedIn
https://www.linkedin.com/in/seth-kauf/