CAPSTONE
NYC 311 DATA

Daniel Demoray
DSI-3

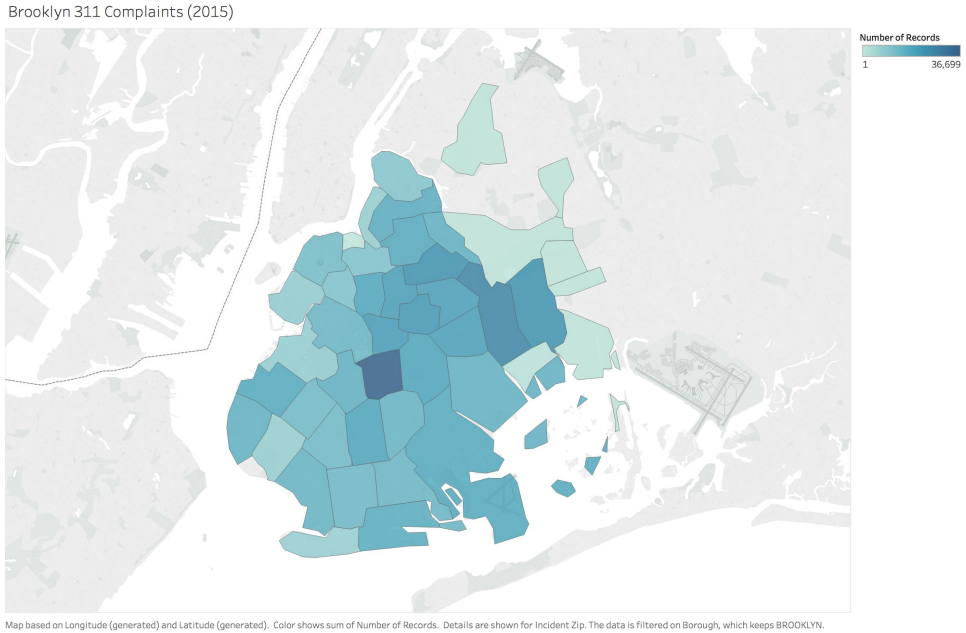
NYC 311

Overview

- Record of inquires placed by NY Residents concerning non-emergency city services.
- Evaluating 2015 Brooklyn specific inquires.

GOAL - Predict time to resolve inquiry

- NY Residents will have the proper expectation set about when their issue can be addressed.
- NY Agencies to have more transparency into what they should prioritize.



FINDINGS

We can make a highly accurate prediction!

- Accomplished promising results on early modeling attempts for Brooklyn, focusing on 2015 Data.

CHALLENGES

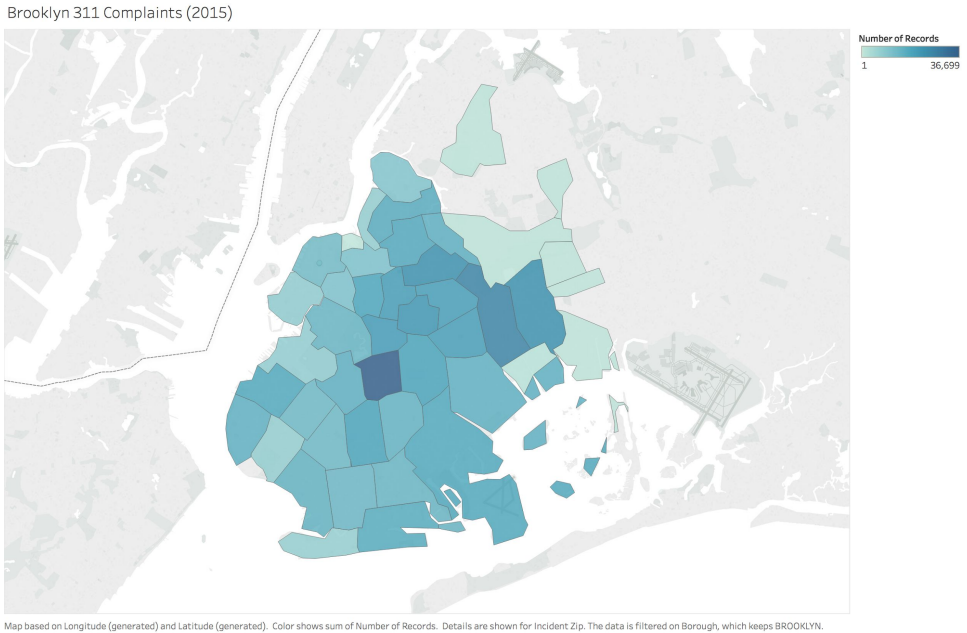
CATEGORICAL DATA IS A BEAST!

- Most insights of from the data are more descriptive, nothing too groundbreaking to find here without geospatial analysis and outside data sets.
- CAUTIOUSLY OPTIMISTIC: Data appeared to be fairly clean, difficult to gauge given data is categorical.

NYC 311

Overview

- Non-Emergency Services & Complaint Hotline
- NYC Open Data
- 8,255,038 Rows
- 52 Columns
  - 62 Distinct Government Agencies
  - 276 Distinct ComplaintTypes
  - 1698 Distinct Descriptors
  - 1764 Distinct Zip Codes
  - Latitude & Longitude
  - And more!



ETL, Cleaning & EDA

SQL

- Large CSV stored on NYC OpenData
- Created table in local MySQL Database
- Sample of larger data set
    - Filtered data by year (2015)
    - Filtered zip code to NYC Metro area
    - Filtered open calls
    - Specified columns of importance

Cleaning & EDA

- Imported filtered data into Python environment
- Sliced DataFrame to exclude values that do not have a specified borough.
- Feature engineering.

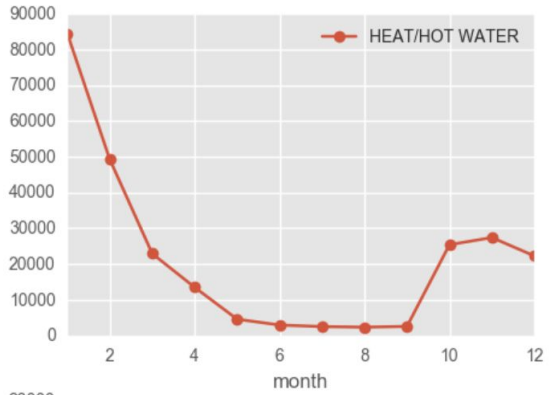
Feature Engineering

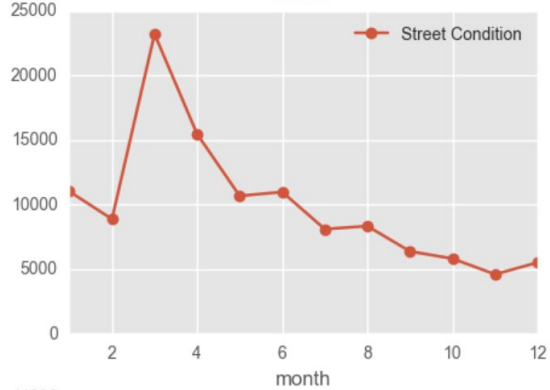
Incorporated the following features into the DataFrame before Analysis and processing for Model.

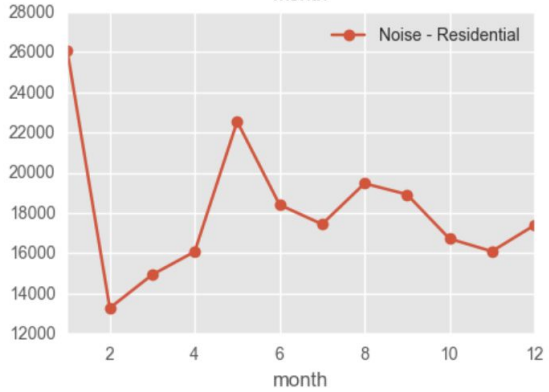
- Using CreatedDate Timestamp:
  - Month
  - Day of Week
  - Day of Month

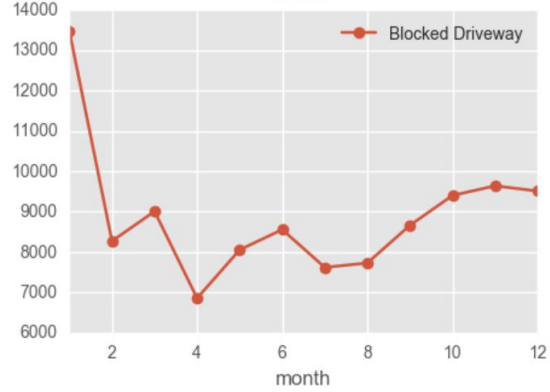
- Delta between Created Date and Closed Date:
  - Days

Monthly Volume of Calls by ComplaintType

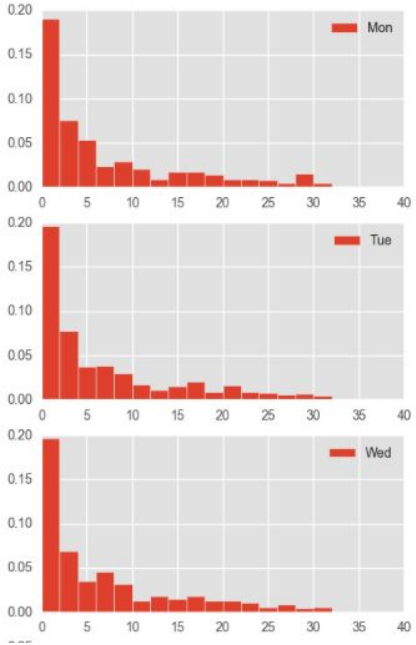


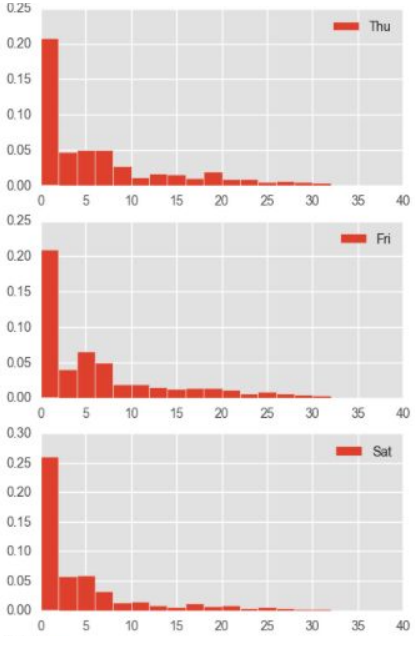


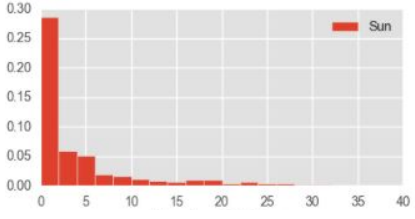




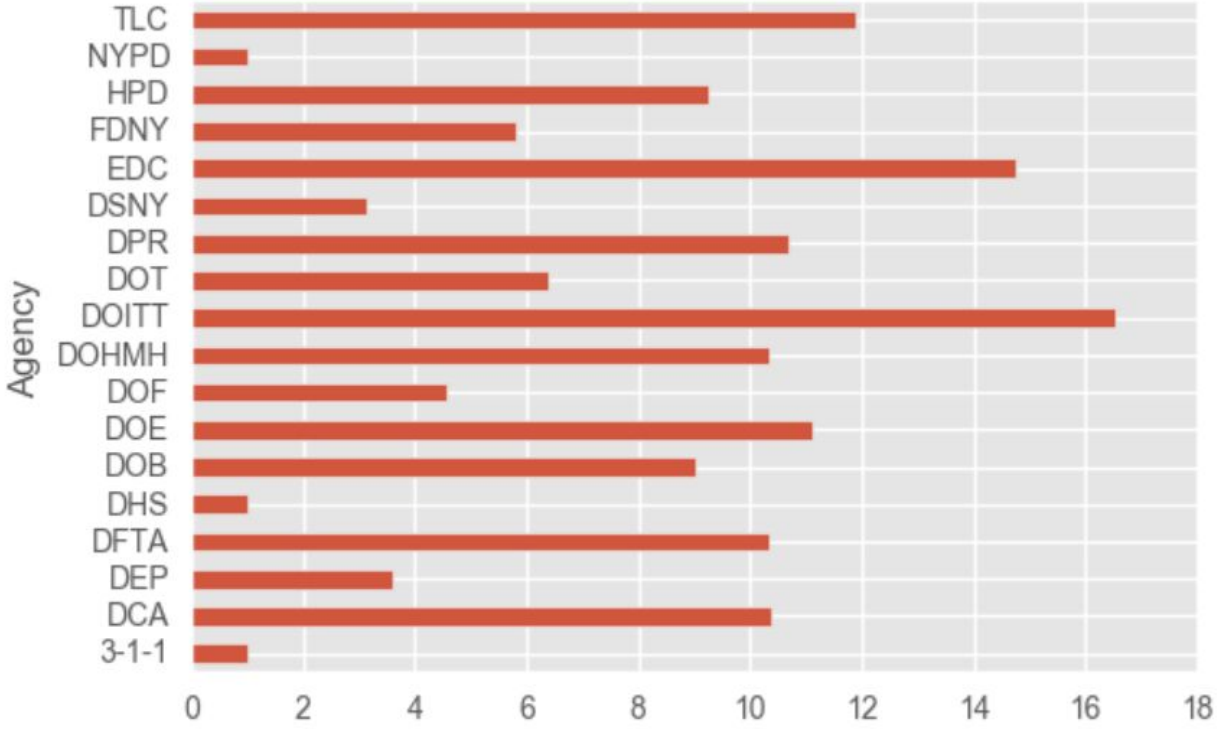
Day of Week



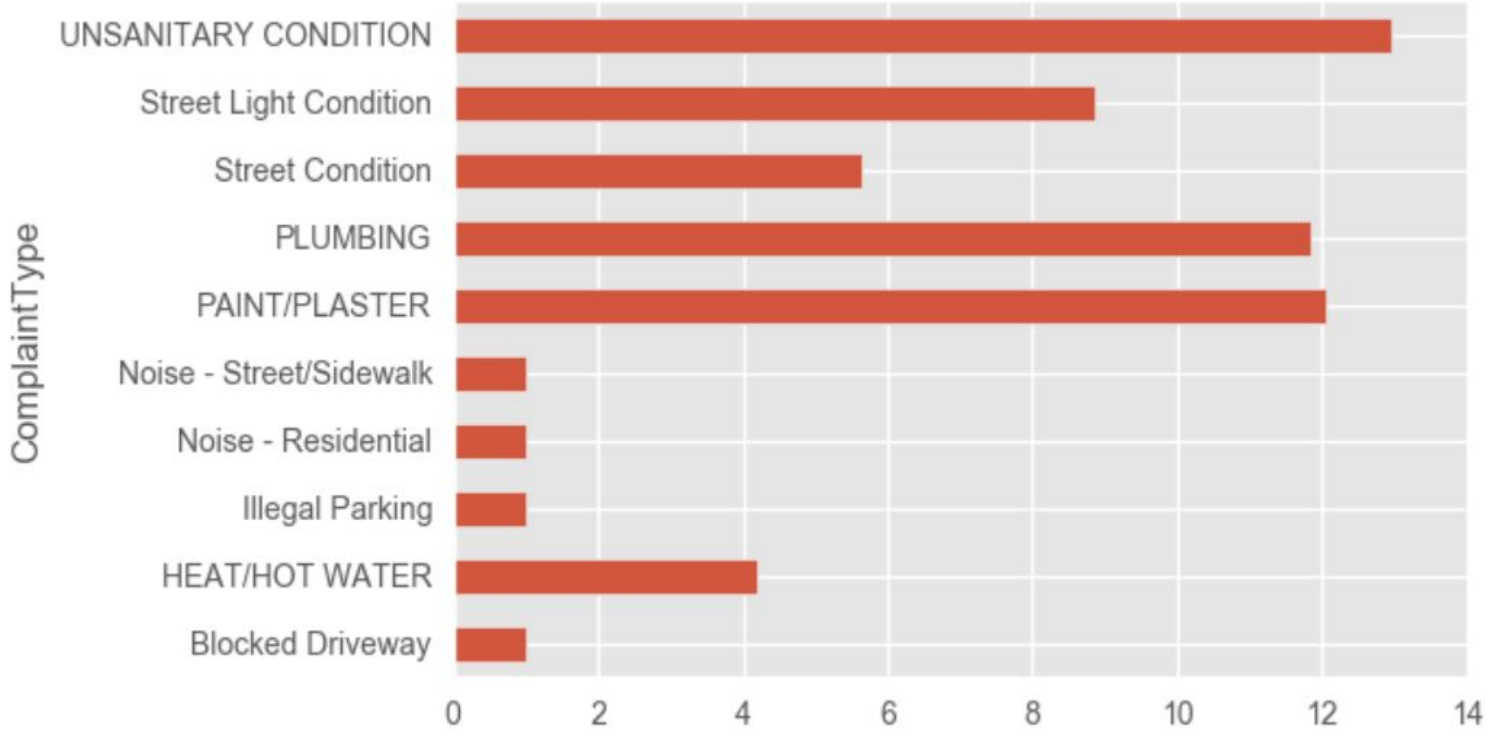




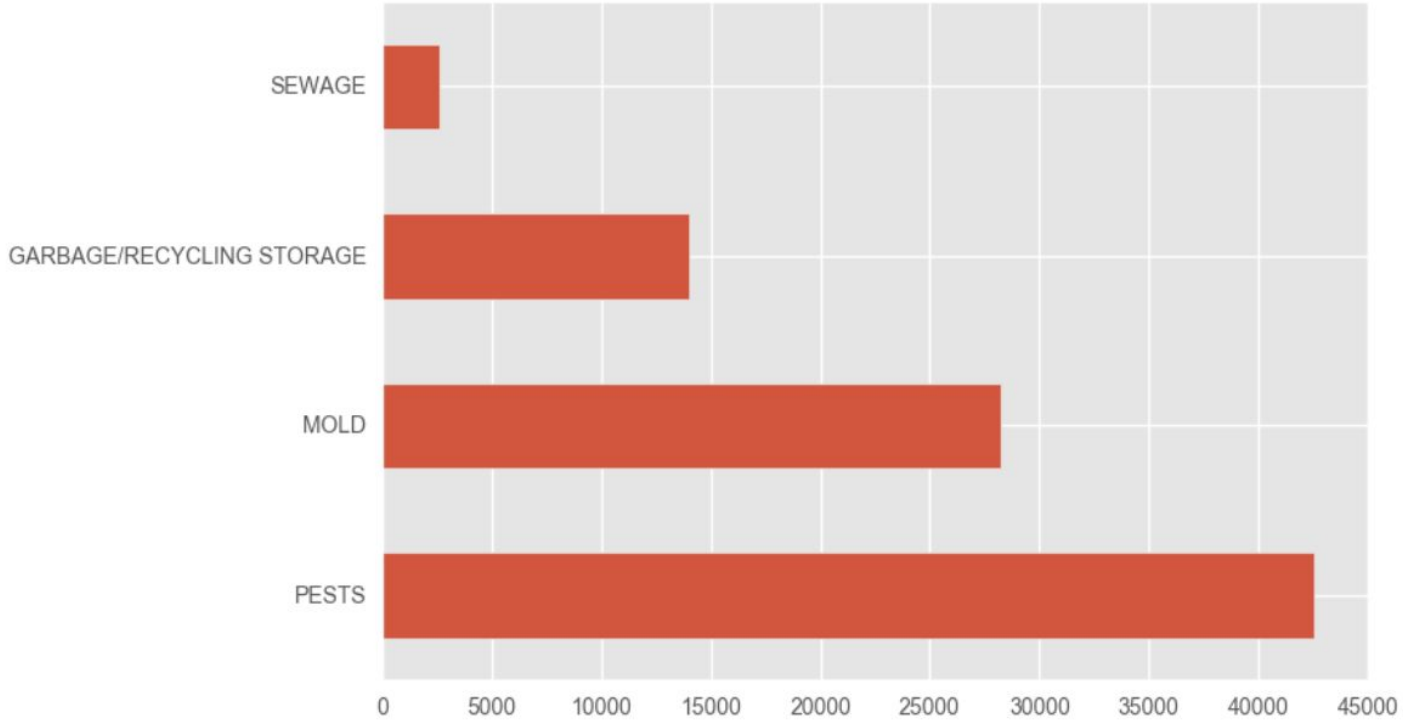
Average Response Time by Agency



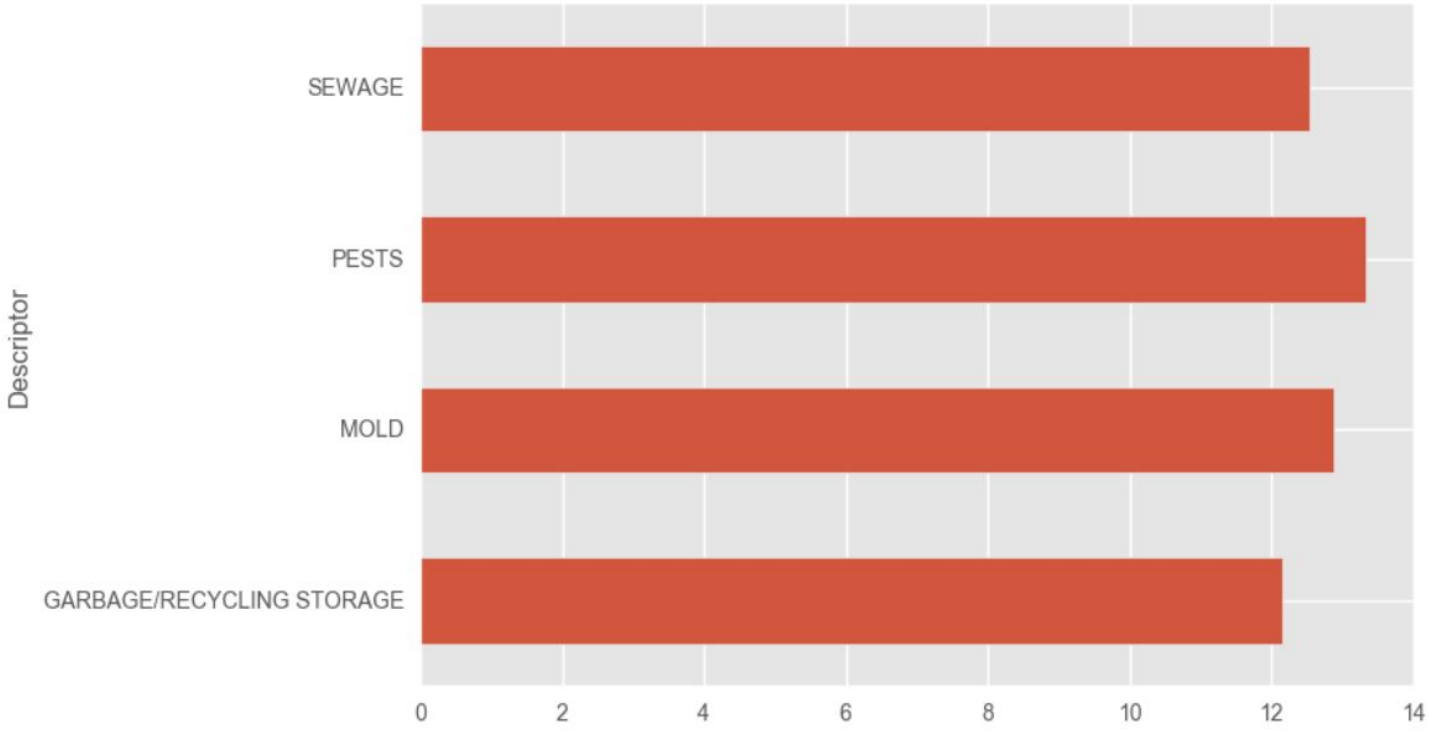
Response time (Most Common Complaints)



Unsanitary Conditions: Volume of Descriptors



Descriptors: Average Response Time (Days)



New Yorkers are all too familiar with Pests...

Given the data tells us what we already know...

Length of 311 Complaint!

MODELING - Feature Selection

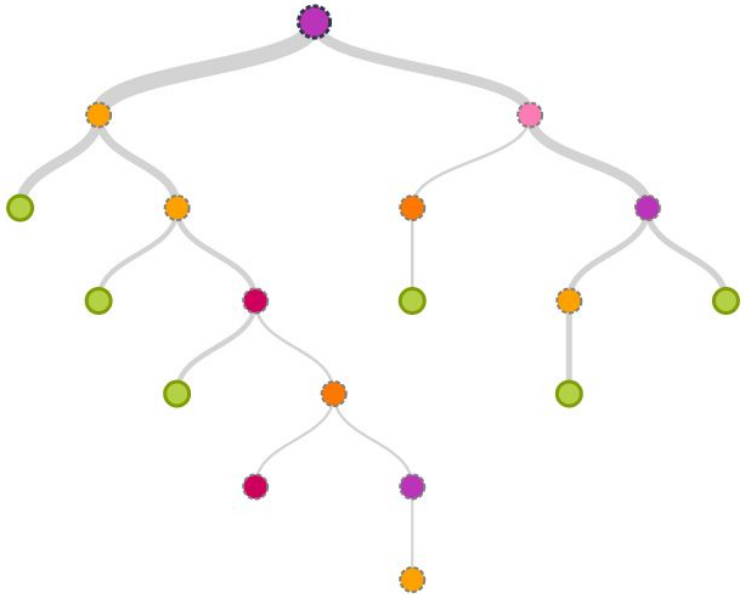
Features

- Descriptor
- IncidentZip
- DayofWeek
- Month
- DayofMonth

Target

- Days (1-31)

**Limited DataFrame to 2015, Brooklyn Data.
**Converted Features to Dummy Variables, corrected for collinearity.



Classification: Random Forest Classifier

MODEL

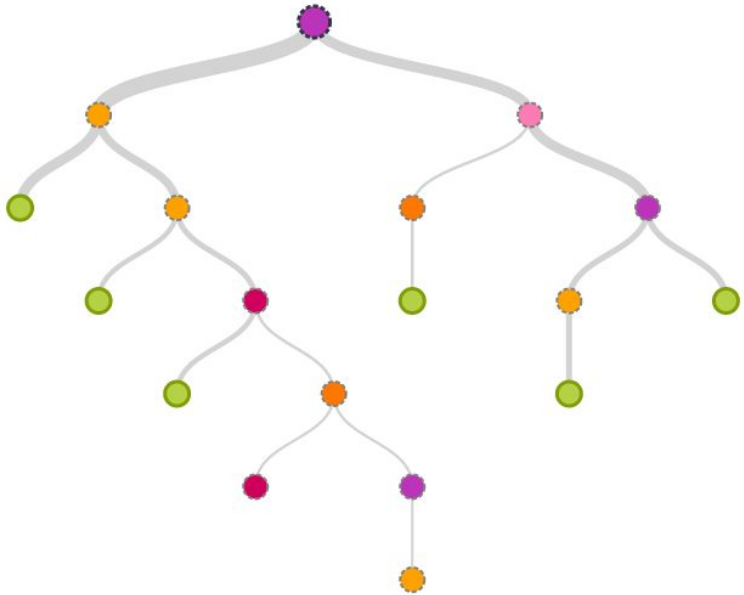
- SciKit Learn: Random Forest Classifier

CROSS VALIDATION

- Split out test and training data from DataFrame
- 30% Testing / 70% Training.
- Instantiated object with n = 100 Trees.

INITIAL RESULTS

- Average Accuracy of 53%.
- Accuracy for individual target classes (1-31) had a high variance.
- Decided to bucket target classes into three groups:



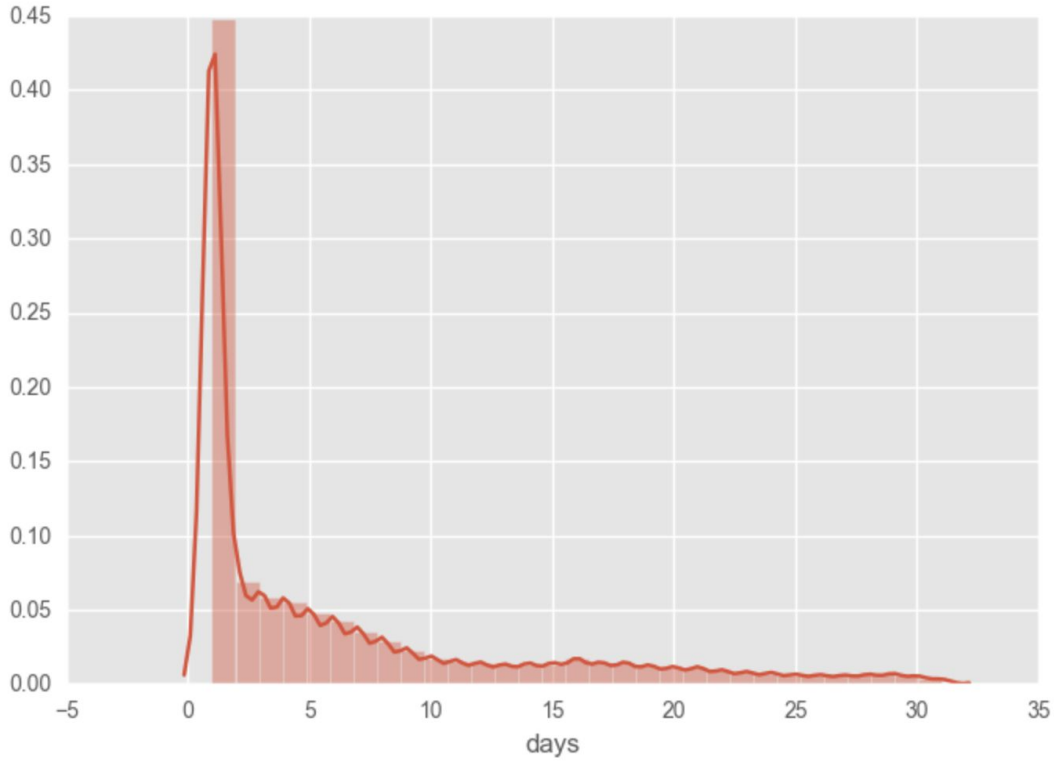
Grouping & Encoding Target Class

By Quantile

1 Day or less
- 43% of Data Points
- Baseline

2-6 Days
- 30% of Data Points

7 or More Days
- 26% of Data Points



Attempt 2: MODELING - Feature Selection

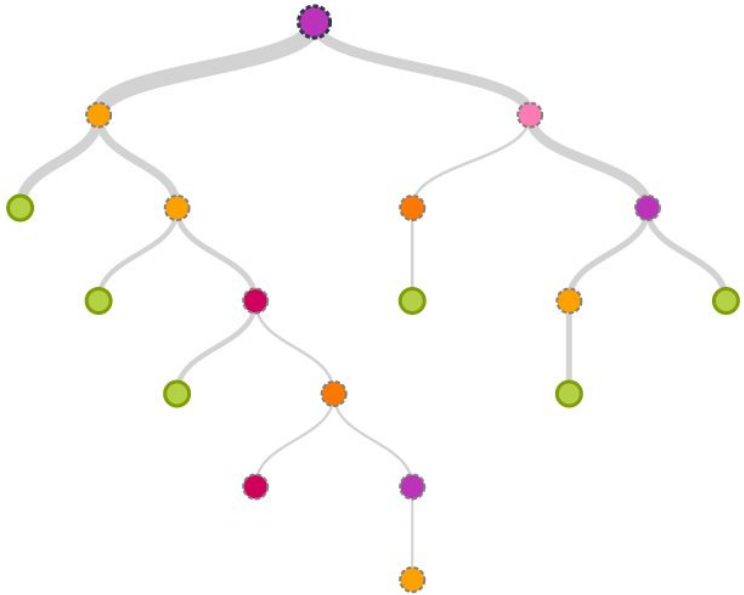
Features

- Descriptor
- IncidentZip
- DayofWeek
- Month
- DayofMonth

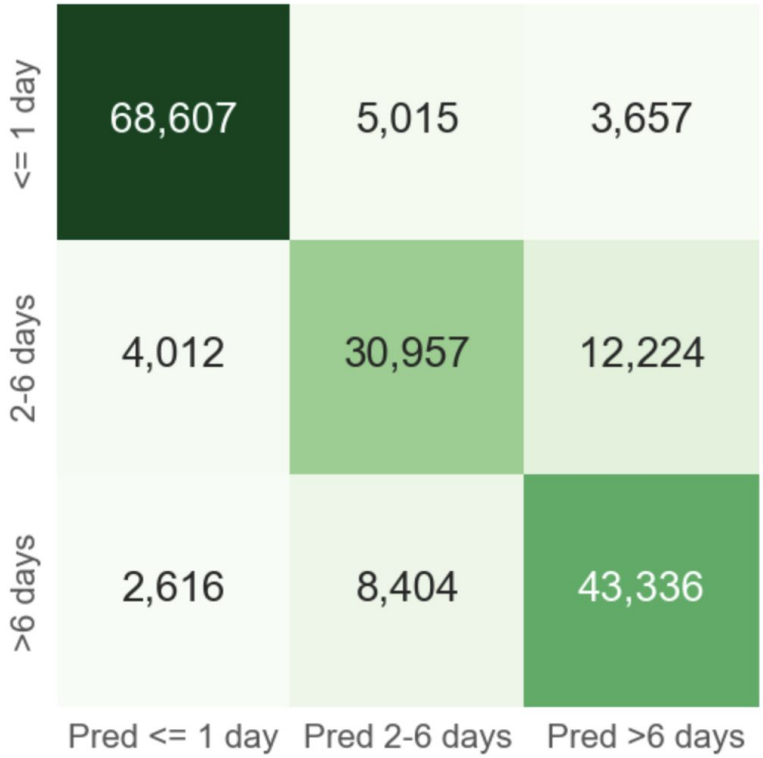
Target

- *Days (Bucketed & encoded into three groups)*
  - *0 : 43% (1 Day or Less)*
  - *1 : 26% (2-6 Days)*
  - *2 : 30% (7 or More Days)*

**Limited DataFrame to 2015, Brooklyn Data.
**Converted Features to Dummy Variables, corrected for collinearity.



CONFIDENCE



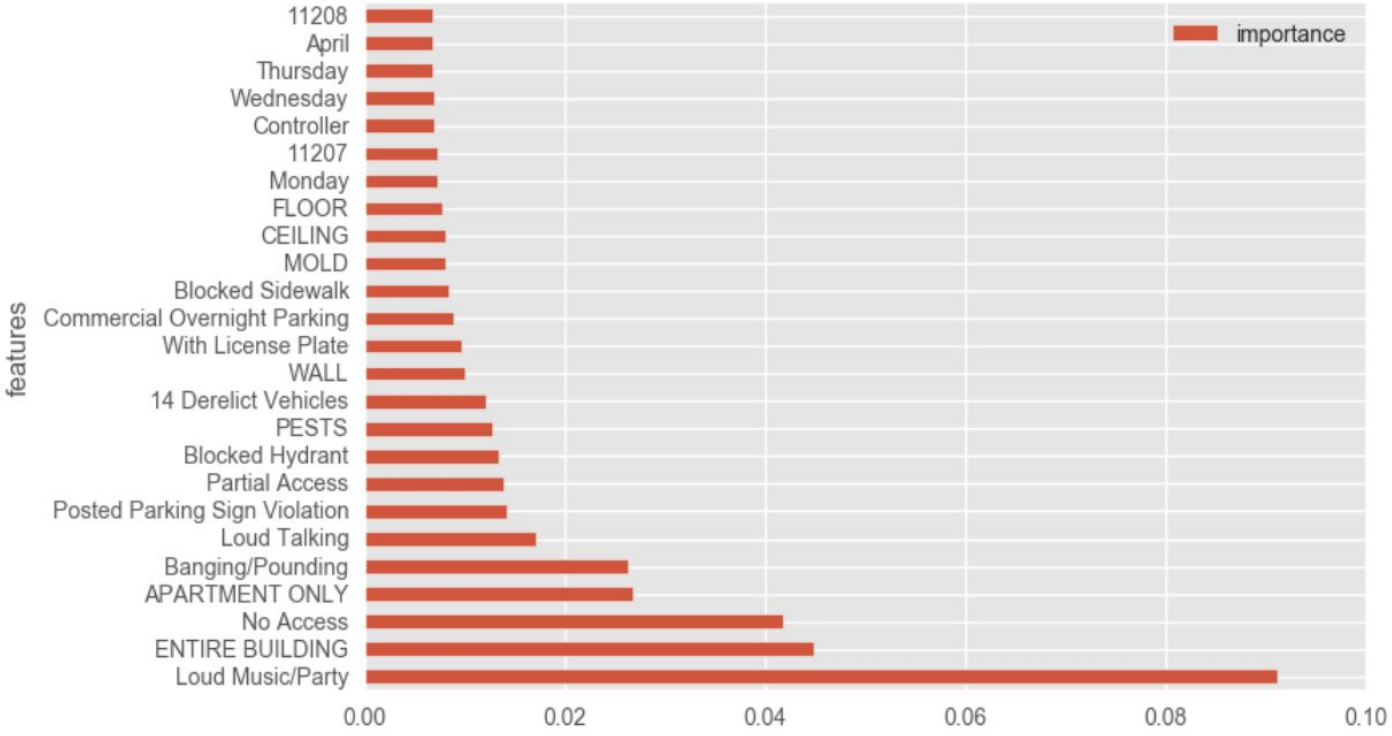
Accuracy: 80%

Baseline: 43%

Buckets:

- Call resolved in 1 day or less.
- Call resolved between 2-6 days.
- Call resolved in more than 6 days.

Feature Importance



CONCLUSION

- Confirms the general intuition of what it means to live in New York.
- From a resident lens: It quantifies the data in a way allows more visibility into short term reality of how long a 311 complaint will be resolved.
- From a public service perspective: Allows officials to get a zoomed out version of a large amount of data, and will allow those individuals to make important decisions about resource allocation.
- These results are only the beginning!

NEXT STEPS

Continue to refine model

- Expand to other Boroughs
- Make response time visible to the public.