An Analysis of Crime in NYC

Ali Abbas Causer
Sprint 2 - November 2023

The Goal

To identify root causes and predict crime in New York (5 boroughs)

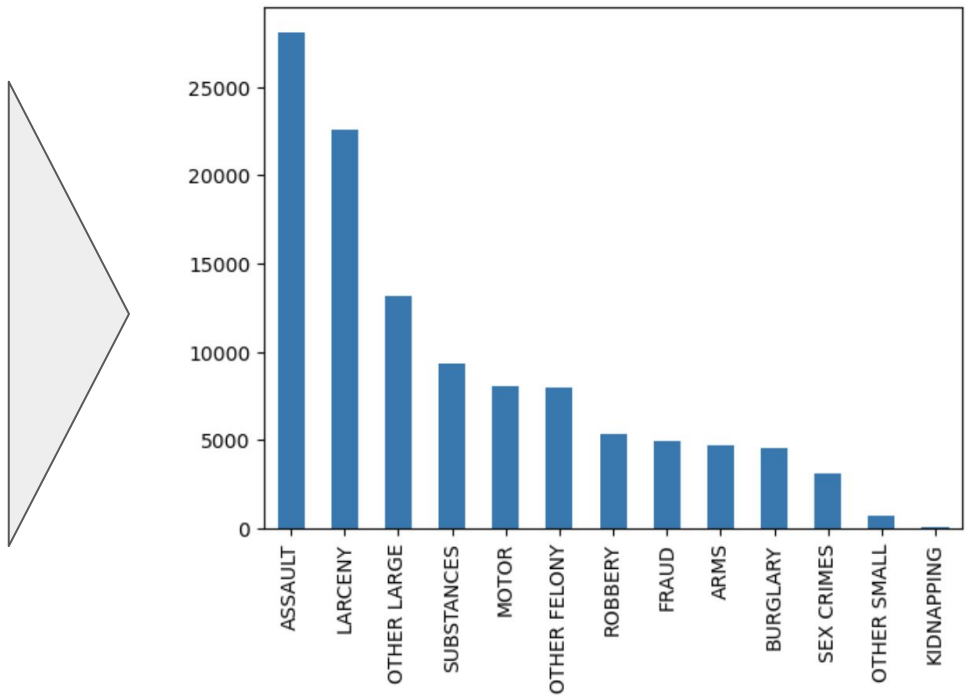
- Can you predict a categorical type and severity of crime based on day of week, weather, location (demographics)? (Log Regression, KNN, Decision Trees)
    - Why is this useful: Can help provide information to residents about safety
- What underlying demographic features (excluding race/ethnicity) are most heavily correlated with crime (type and frequency)? (Log regression)
    - Why is this useful: Can help provide local public organizations areas to improve (e.g. education, shelter, healthcare, jobs)
- Relationships of type/severity of crime with age groups and why? (Linear regression)
    - Why is this useful: “Customer” segmentation to develop a targeted approach for action

Since last time…

- Feature engineering
- Run a model
- Identified a better data source for relevant demographic features to incorporate in the model
- Join arrests data with demographic and economics data using geopandas

Feature engineering

```
ASSAULT 3 & RELATED OFFENSES      16619
PETIT LARCENY                     12070
FELONY ASSAULT                    10474
DANGEROUS DRUGS                    7530
MISCELLANEOUS PENAL LAW            7327
                                   ...
PARKING OFFENSES                      3
FELONY SEX CRIMES                     2
DISRUPTION OF A RELIGIOUS SERV        2
UNLAWFUL POSS. WEAP. ON SCHOOL        1
ADMINISTRATIVE CODES                  1
Name: OFNS_DESC, Length: 63, dtype: int64
```

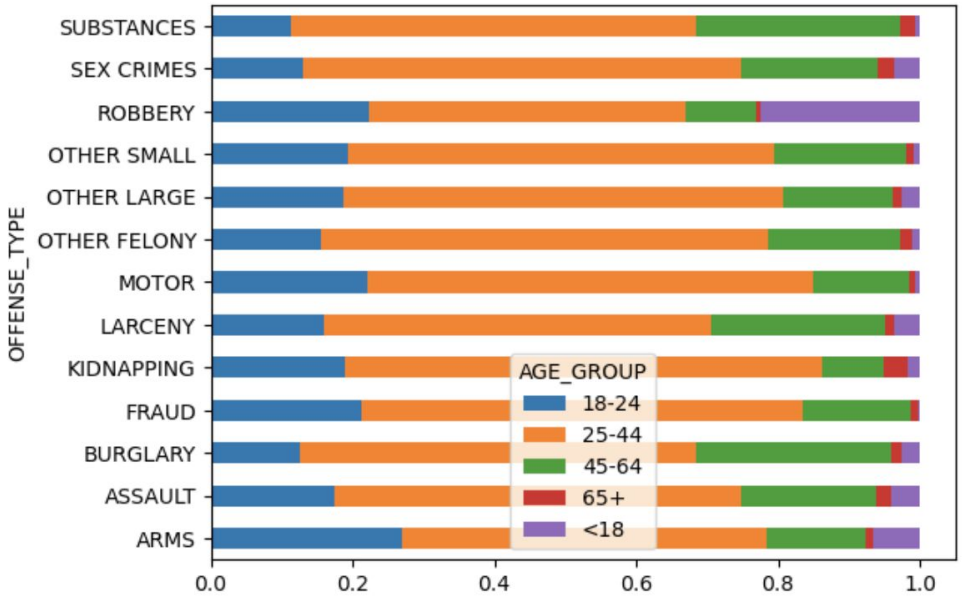


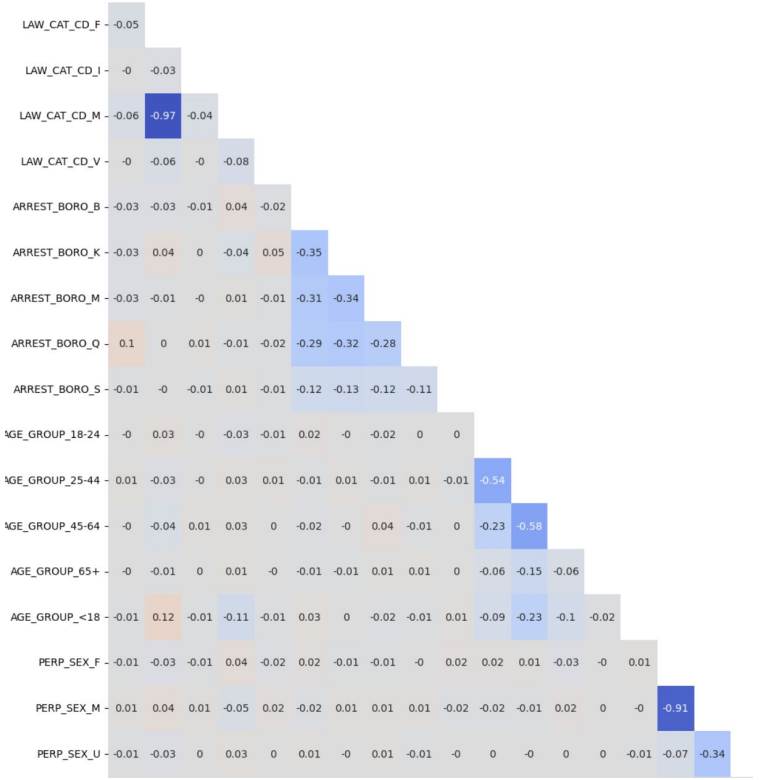
Imputations

```
larceny = arrests[arrests['OFFENSE_TYPE'].str.contains('|'.join(['larceny','thef','stolen','steal']),case=False,na=False)==True]
burglary = arrests[arrests['OFFENSE_TYPE'].str.contains('|'.join(['burglar','trespass']),case=False,na=False)==True]
substance = arrests[arrests['OFFENSE_TYPE'].str.contains('|'.join(['drug','cannabis','alcohol']),case=False,na=False)==True]
```

Felony > 80% of category = Other Felony, Category > 100 instances = Other Large

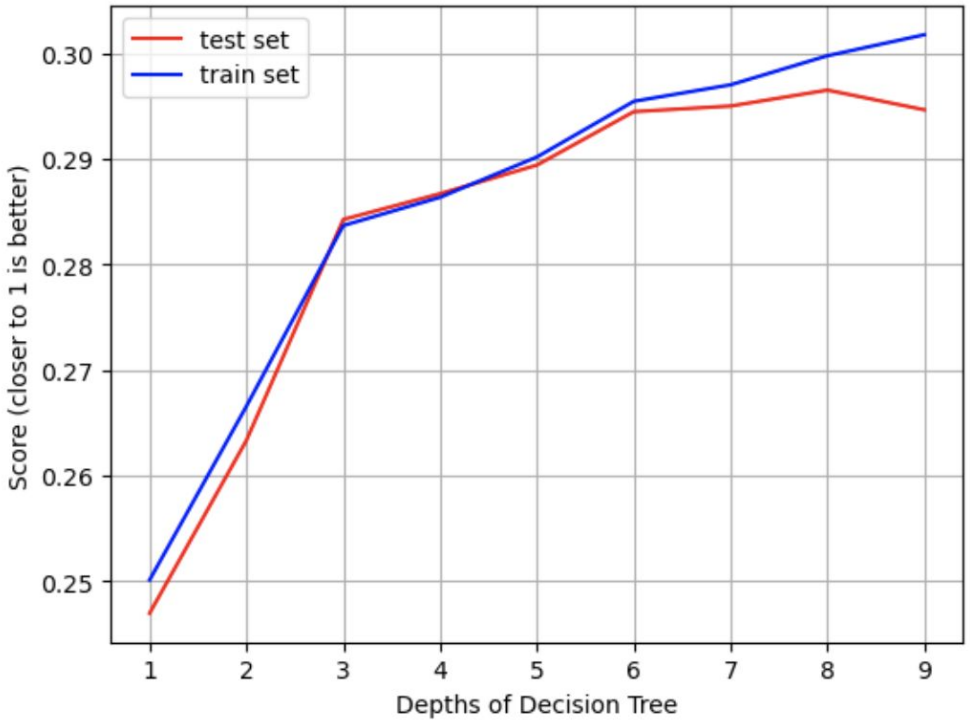
Pre-model assessment



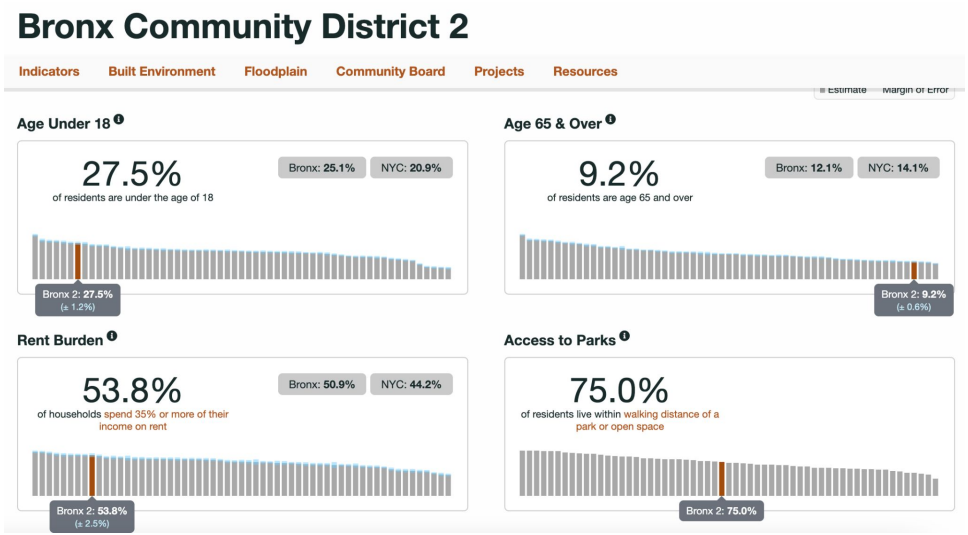


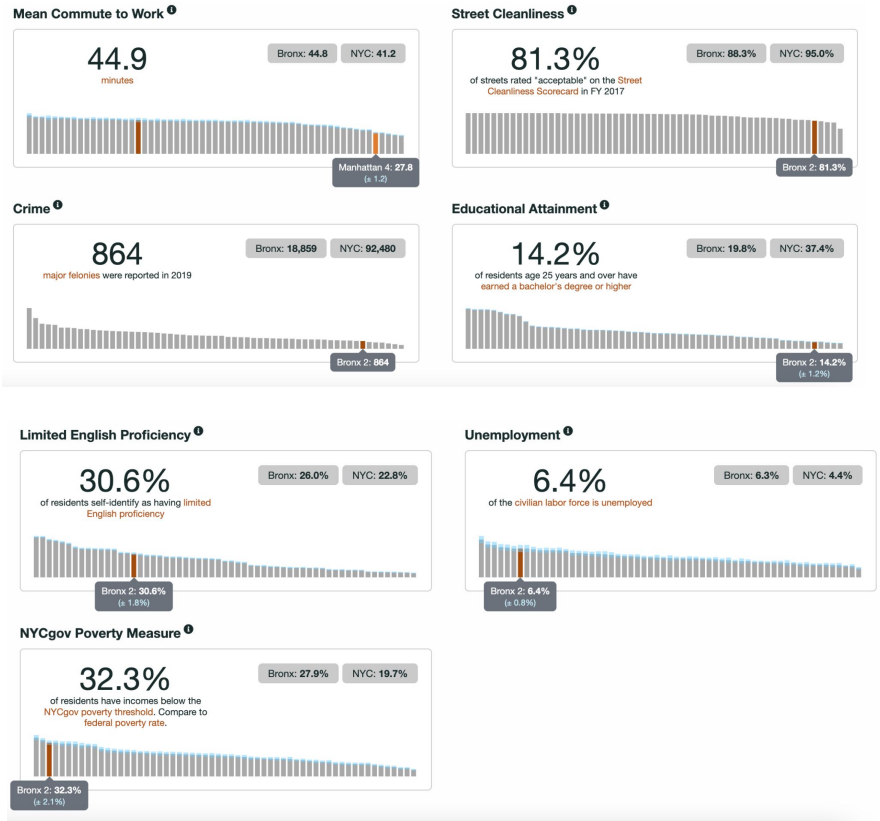
Logistic Regression / Decision Tree

31% Train
30% Test



Additional location based features





Next steps for model improvement…

- Further tighten target categories
- Re-run analysis using expanded feature set with demographic data
- Get daily weather data
- Break down the analysis into small studies
- GridSearch to assess performance across multiple models and params