Applications of Big Data & Machine Learning for Crime Analytics in New York City

Master of Quantitative Economics, UCLA
402A: Macroeconomic Theory
Author: Carlos Hernandez
Faculty Advisor: Professor Patrick Convery
Date Completed: 03/11/2022

Table of Contents

1. Abstract

This paper aims to leverage crime and economic data in New York City by implementing a variety of big data and machine learning models for two topics of interest in crime analytics. The first focuses on the crime rate regression, where we implement regularization techniques that focus on dimensionality reduction and helping to identify which variables have the largest influence on the crime rate. We then shift our focus to using classification models to predict the location of a crime within each borough of the city. The primary goal of this analysis is to utilize the significant amount of data on crime and leverage it to understand crime in the city. This has benefits not only for law enforcement, but also for policy makers who may not consider some of the modern statistical techniques that can be very useful in combatting crime. Another important benefit is that this analysis could also serve as a template for other American cities that face similar crime issues as New York City, further underscoring the usefulness that such models may have in dealing with all types of crime.

2. Introduction

Understanding and predicting crime is a difficult task due to its random and varied nature. Not only is it challenging to determine when and why a crime might occur, it is often not reported to the authorities. According to a 2012 press release from the Bureau of Justice, around 3.4 million violent crimes were unreported between 2006 and 2010. Some reasons included fear of retaliation and victims feeling they could not go to the police, further underscoring how difficult it is to fight all crime, but also highlighting the importance of what may cause crime.

While, unfortunately, not all crimes are reported to police, available crime data is still an incredibly useful as a tool to combat crime because the characteristics and frequency of certain crimes can help law enforcement to better and more efficiently utilize their resources. For example, if assault is common in a certain neighborhood during the night, both police and possible victims would benefit from knowing this by having more officers patrolling this area. This could help to both prevent crimes from ever taking place and lead to quicker law enforcement response. This data can also be useful in determining what factors that may be driving crime. A stronger understanding of these factors can lead to improved policies or programs to help combat this.

The concepts mentioned above serve as the foundation for this paper, where we utilize crime and census data in New York City (NYC) to implement big data and machine learning algorithms for crime analytics. Specifically, we implement these models for two purposes. The first is using regularization techniques to understand the factors that influence the crime rate in NYC. We will then leverage the data to predict the location of crimes for each borough using classification models. With this, we can not only gain a strong idea of the factors that influence crime the most in NYC, but also determine the accuracy and effectiveness of these models in predicting where a crime might occur within each of the city's five boroughs.

3. Methodology

As mentioned, the goal of this paper is to understand what may affect the crime rate in NYC and to predict the location of crimes in each borough. As such, we run models for each of the two topics.

For crime rate, we implement regularization techniques to identify which variables contribute most heavily to the crime rate. This ultimately will give insight on the indicators that

contribute the most to the level of the crime rate. I will also evaluate the predictive power of the model to ensure that it is an appropriate model for the data. This section will focus on all of NYC and the models used are LASSO, Ridge, and Elastic Net.

In the Crime location prediction portion of the analysis, the focus is narrowed to each borough and uses classification models. This includes Logistic Regression, K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA). The dataset provided an easy opportunity to classify where crimes happen because GPS coordinates are provided, but also because two of the boroughs were already broken into categories by northern and southern zones. If these categories were not provided, we used K-means to create clusters then implemented classification algorithms based on this.

3.1. Overview of Model Techniques & Evaluation Methods

Below, we give a brief theoretical overview of each of the models, as well as the evaluation methods that we intend to use to test the performance of the models.

3.1.1. LASSO, Ridge, & Elastic Net for NYC Crime Rate Regression

The three models that we use for Crime Rate regression are very useful for dealing with high dimensional data. They essentially act as an extension of linear regression but introduce a penalty parameter that "shrinks" the coefficients of interest. The concept of shrinkage essentially acts as a method of variable selection. This is because the values of coefficients shrink to be very close to zero to reduce the variance. In this process, many coefficients may become equal to zero, which means these models can eliminate certain variables and allow us to determine which variables have the most influence on the dependent variable of interest. In this case, it provides insight on the variables that have the most influence on the crime rate in NYC.

When we consider the classic Ordinary Least Squares model, we seek to minimize Residual Sum of Squares (RSS), defined as the following:

$$RSS = \sum (y_i - \beta_0 - \sum \beta_j x_{i,j})^2$$

Ridge and Lasso have a very similar equation but add a penalty parameter (lambda) in addition to the RSS equation. For Ridge, this is:

$$Ridge = RSS + \lambda \sum \beta_j^2$$

For LASSO, the absolute value of the coefficient is taken, rather than the square of the coefficient:

$$LASSO = RSS + \lambda \sum |\beta_j|$$

In the case of Ridge Regression, the penalty term does not force the coefficients to become zero. Thus, many variables will remain in the model and, even if the coefficients are very close to zero, the model is difficult to interpret. With the LASSO penalty, some coefficients do become exactly zero and thus eliminates some variable which helps address the problem seen in Ridge, making interpretability easier, while more explicitly selecting features that are important. Despite this, it is important to note that just because a coefficient is equal to zero, it does not render the variable completely irrelevant. Rather, it allows us to compare within the model which variables contributed the most to changes in the value of the output. This brings us to the regularization technique, Elastic Net, which is a combination of both Ridge and LASSO:

$$RSS + \lambda_1 \sum |\beta_j| + \lambda \sum \beta_j^2$$

By combining the two penalty parameters to create one, Elastic Net looks to address the drawbacks of running LASSO or Ridge regression on their own.

3.1.2. Techniques used for Borough Location Prediction

The first model that we will introduce is K-means. We use this to create location clusters for each of NYC's five boroughs based on the GPS coordinates provided in the data. The K in this method represents the number of clusters we wish to obtain from the data. Every observation belongs to a single cluster, with no overlap, with the goal being for all the observations in a cluster to be as similar as possible.

The purpose of creating intra-borough clusters is to then implement classification models, the first being logistic regression. Logistic regression allows us to implement a model where our target variable is a binary categorical variable with an outcome of 1 if "yes" or 0 if "no". This method looks to determine the probability that Y equals 1, given a set of predictors. That is:

$$p(X) = Pr(Y = 1|X = x)$$

Calculating the probabilities ensures that predicted values remain between 0 and 1, whereas in linear regression, the predicted value can be any value. This calculation remains between these values because logistic regression utilizes the sigmoid function, which takes the form:

$$P(X) = e^{\beta_0+\beta_1 X}/1 + e^{\beta_0+\beta_1 X}$$

This ensures that the predicted value will not exceed 1. Predicted values are most likely not going to be exactly 1 or 0. To overcome this, we set a threshold, usually .5. To elaborate, if a predicted value is .6, we classify it as 1 because it is greater than .5. Any value below .5 will be classified as zero.

With logistic regression serving as a baseline classification model for location prediction, we then introduce k-NN. To understand this method, we consider the name of the technique itself, Nearest Neighbors. When we are looking to classify an observation, we look at the "nearest" observation with similar predictor variables and then do this "K" times while classifying a record by the predominant class. To calculate the distance, the classic Euclidean Distance formula is used. We then choose K to be the number of nearest neighbors considered when looking to classify an observation and is chosen based on the lowest testing set error. In general, lower K values are good at capturing local structure and noise since they are only considering a few nearby records. A larger K may not be able to do this but does provide less noise and more smoothing at the expense of capturing local structure. K-NN is incredibly useful due to the lack of assumptions it makes on the data.

The final classification models that we consider is LDA. This method looks to perform the same tasks as Logistic Regression in that the variable being predicted is categorical. There are key differences between the two, however. The first is that LDA is more effective when the classes are easily separable. In addition to this, this algorithm has two key assumptions: the observations are a random sample, and each predictor is normally distributed. Another key feature of LDA is that it can be used to classify more than two classes, though in this paper we only have a binary output.

3.2. Methods of Evaluation

We utilize several evaluation methods that will test the robustness and performance of our models, while also helping us to choose hyperparameters for the models where this is applicable. This includes using training and testing sets and cross validation. We will also test the accuracy of the models and error rates using R^2 and Root Mean Square Error (RMSE) for regression models. For our classification models, we use confusion matrices while also calculating the accuracy, precision, and recall.

4. Data

Most of the data used for the purposes of this paper comes from the NYC Open Data. Specifically, the dataset used was the "NYPD Complaint Data Historic", with the data filtered to focus from the start of 2017 to the end of 2020. Complaint data is defined as being all the crimes that were reported to the police. This data was chosen since this is what is used to calculate the crime rate. This is defined as the number of crimes reported divided by the population of an area for a given period per 100,000 people (State of California, DoJ 2014). In this paper, the crime rate is monthly.

Since the crime data we used contained only information related to each of the reported crimes, we also wanted to consider and use other potential variables that may or may not influence crime. As a result, we also pulled and calculated several economic and demographic indicators from the USA Census data. This included unemployment rate, educational attainment, and share of population by race or ethnicity, among other variables. Some of these indicators were pulled from the 1-year ACS surveys for the respective years considered in the analysis and the rest of the variables came from the 2020 census.

4.1. Variables used in both Crime Rate & Location Prediction Models

1. **Type of Crime:** Dummy variable. It has three options with a crime that was a "violation" the reference. This meant that the other two crime types were included in the model. There was significant overlap with this variable and the type of crime committed. Given the overlap between the two, type of crime was used as it already captures the severity of a crime.
2. **Status of Crime:** Dummy variable. It has two options: "completed" or "attempted", with the latter being the reference.
3. **Race of suspect:** Dummy variable with 7 options. If the race of the suspect was not known, this was the reference dummy. We keep unknown suspects because this is very common in crimes and dropping this would have been a significant number of observations.
4. **Age of suspect:** Dummy variable with 6 options. Reference was also unknown.
5. **Sex of suspect:** Dummy variable with 3 options. Reference was also unknown.
6. **Race of victim:** Dummy variable with 7 options. Reference again was unknown.
7. **Age of victim:** Dummy variable with 6 options. Reference was again unknown.

8. **Sex of Victim:** Dummy variable with 3 options. Reference was other. This mostly meant people that are non-binary in the data.
9. **Month:** Dummy variable for month crime occurred. January was the reference month.
10. **Premises of crime:** Dummy variable with 75 options. This indicator describes the premises of where the crime occurred and is very specific. The reference is if a crime occurred on a street.
11. **Continuous variables:**
    a. **Crime Rate**: Continuous, monthly. Author's calculation using definitions of crime rate and is per 100k people. This was calculated for each of NYC's borough from 2017 to 2020 using the population of each borough in each year. This was the dependent variable for the crime rate section but an independent variable in crime location prediction. For a given month, the formula is as follows:

       Crime Rate = (Total number of crimes/population) x 100,000

    b. Median age by borough. Pulled for each year between 2017 and 2020 from the 1-year ACS.
    c. Unemployment rate by borough, data from 1-year ACS.
    d. Poverty rate by borough. Data comes from the 1-year ACS. Poverty rate definition is that of the US census.
    e. Educational attainment by borough. This is defined as the percentage of the population over 25yrs old w at least a bachelor's degree. Data is from 1-year ACS.
    f. Labor force participation rate by borough. Data is from 1-year ACS.
    g. Percentage of each borough that is Hispanic. Calculated manually using population estimates for each borough and data from 2020 Census.
    h. Percentage of each borough that is black. Same calculation as the rate of Hispanic people but using the number of black people in each borough.
    i. Percentage of each borough that is Asian. Same calculation as the rate of Hispanic people but using the number of Asian people in each borough.
    j. Difference in days between start and end date of a crime. This was calculated manually based on the crime dates provided by NYPD data. If the end date was labeled "not applicable", then it meant that the crime ended on the same day as it started, and the data was adjusted to reflect this and make the calculations.

4.2. Variables used only for Crime Rate Models

1. **Borough:** Dummy variable with Manhattan being the reference. This was only used for the crime rate models because zones within each borough were the dependent variable in the crime location prediction models.

4.3. Variables used only for Crime Location Prediction

1. **GPS Coordinates:** The crime data provided the exact latitude and longitude of a crime. This data was not directly used in the models but was used to create intra-borough clusters using K-means clustering.
2. **Intra-borough clusters:** Three (Brooklyn, Manhattan, Queens) provided a categorical variable that stated if a crime occurred in the north or south. This will serve as our output variable for these boroughs. For Staten Island and the Bronx, we will leverage the GPS coordinates of each crime. We begin by creating two clusters and implementing the classification models. We will then test additional clusters, including doing this for the boroughs that already provided zones.

4.4. Data Preparation & Descriptive Statistics

4.4.1. Data Preparation

The process of preparing the data to implement the models of interest was extensive due to multiple data sources and the need to prepare the variables for use. As previously mentioned, the crime data from NYC open data was largely made up of categories that were labeled and needed to be converted to dummy variables. Some of the variables had many observations that were labeled as "unknown" or "NA". However, these cases did not necessarily reduce the importance of these variables. Rather, in the case of crime these observations could prove very insightful. For example, in our list of variables we set the reference dummy for a suspect's age, race, and gender as "Unknown". This makes sense in practice because it is possible that a suspect has not yet been identified or that they managed to escape the crime scene. It also could mean poor reporting to police, which further highlights the importance that our models capture this factor rather than modify or drop these observations. As a result, some of the dummy variables have "unknown" as the reference.

Other variables that were prepared using the NYC crime data included crime rate and the difference in the start and end date of a crime. The former was calculated by simply summing all the crimes that took place in a given month and year, then applying the formula that was described in

the variable description. The latter was a simple subtraction. However, what were likely errors in data entry meant that a few observations were dropped.

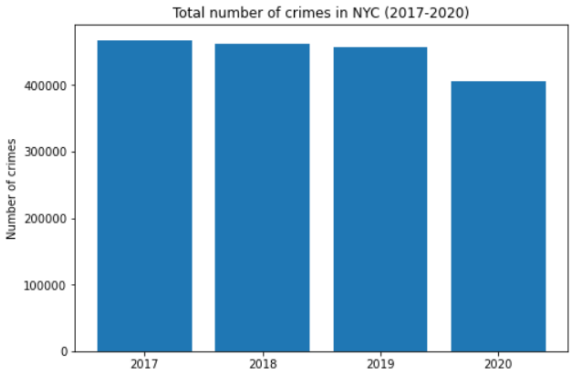
NYC crime data provided an extensive number of variables related to crime. However, there are many demographic and economic factors that likely contribute to the level and severity of crimes. With this, we pulled a variety of economic and demographic variables from the census bureau database. The data was focused on each borough and came from the 1-year ACS for the years between 2017 and 2020. If the data was not available in the 1-year ACS, then the 2020 census data was used. Once all the data was compiled, it was integrated into the NYC crime data to implement the models.

The final variable that required preparation was creating clusters within each borough by GPS coordinates. The exact location of each crime was reported, and this allowed an excellent opportunity to create clusters by location and implement classification models. The process will be described more in the next section. With all the data collected, the final step was to standardize all continuous variables to bring to a comparable scale. This was necessary because of the models we are implementing, which are sensitive to large values. Had the variables been maintained in their regular format, some variables would likely end up dominating the models and would lead to results that may not reflect the actual situation.

The raw crime data had 1,792,148, but with the need to merge economic indicators and the presence of missing observations, this was reduced to 1,790,832 for our analysis with regularization models. When doing location prediction, the range of number of observations was from 77,474 for the Bronx to 521,622 for Brooklyn.

4.4.2. Descriptive Statistics

For this section, we introduce the data more, focusing on some of the characteristics of the crime data, beginning with the total number of crimes in the period considered:



The total volume of crimes in the city appeared to decline slightly, with the most notable drop in 2020. We then look at average monthly crime rate by borough, as well as descriptive statistics on crime rates:

Table: Average Crime Rate by Borough (LHS) & Crime Rate Statistics (RHS)

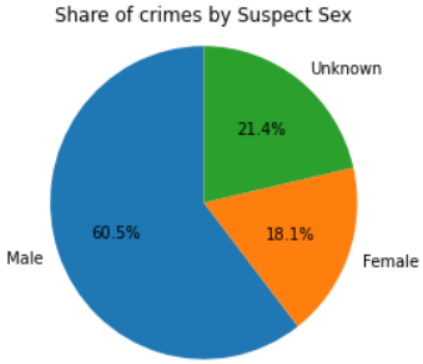
Borough	Avg Crime Rate
bronx	566.192
brooklyn	413.065
manhattan	553.111
queens	323.449
staten	335.458

	crime rate statistics
mean	461.837
std	110.469
min	221.040
25%	358.720
50%	454.390
75%	567.340
max	666.590

The Bronx had the highest average monthly crime rate followed closely by Manhattan. When looking at the overall crime rate statistics, these two boroughs far exceeded the mean average crime rate of 461.837 seen on the table to the right.

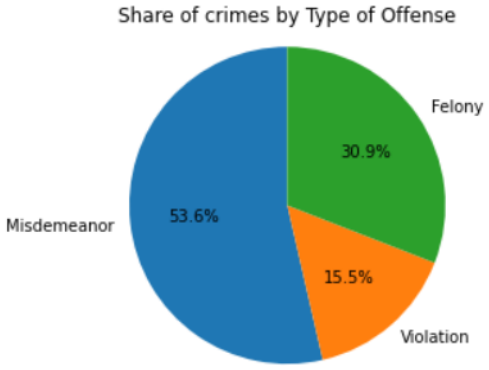
We also look at some characteristics of suspects, considering that they are the ones that almost certainly committed the crime. Below, a table for the number of suspects by age group and a pie chart of the share of suspects by gender is shown:

Age Group	Number of Suspects
UNKNOWN	522,639
25-44	448,296
45-64	162,995
18-24	156,850
<18	45,528
65+	14,710



Most of the suspects aged below 45 years old, when looking at the table. In addition, it appears many of these suspects are Male, as might be expected. However, the large number of unknown suspects is noteworthy and forms a key part of our analysis because we wish to capture this.

Our final plot for this section shows another pie chart, but for the type of offense:



Misdemeanors, which are more severe than violations but less severe than felonies, made up just over half of all the crimes. Felonies had the second largest share. The severity of the crime may likely be a key indicator since it also relates to the type of crime committed.

5. Models & Results of Analysis

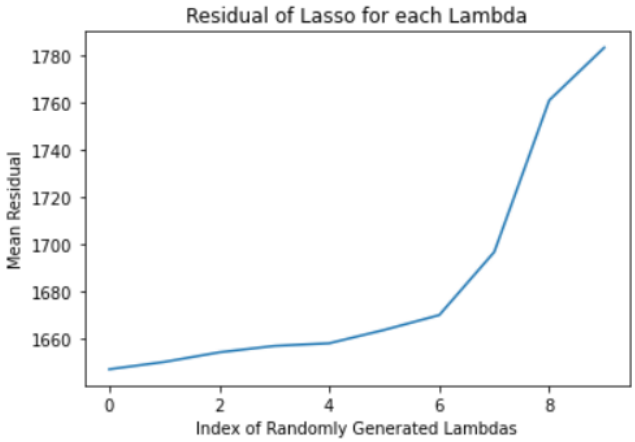
Now that the models and variables that will be used to analyze crime in NYC have been discussed, we now begin describing the process of developing and testing different models before assessing the results. We begin with the Crime Rate models before transitioning to our classification models for Crime Location prediction.

5.1. Crime Rate Models & Results

For our crime rate regression models, we will describe the process of developing the model for LASSO, as well as the results to provide clarity on how the models and hyperparameters were chosen. However, we will not repeat this process with Ridge and Elastic Net as the process was the same. Instead, we will simply discuss the results. Each model had 126 variables, with most being dummy variables as highlighted in the variable description. The Monthly Crime rate per 100,000 people is the dependent variable and the dataset had a total 1,792,148 observations.

5.1.1. LASSO Models

A key part of implementing LASSO is choosing an optimal penalty parameter or Lambda. To do this, we use a 5-fold Cross Validation for the model. This is a feature of the Python library that tests a range of possible values of Lambda and chooses the best performing parameter. In our case, it returned a Lambda equal to .08. This led us to conclude that the optimal Lambda was close to this value, but we wished to test this by using a list of ten numbers between 0 and 1. We then randomly split the data into training and testing sets and re-implemented LASSO with the random Lambdas. In each iteration of Lambdas, the mean value of the residuals was calculated, and we plot it below:



Index	Lambda
0	0.0124
1	0.0325
2	0.0593
3	0.0842
4	0.0943
5	0.1560
6	0.2350
7	0.4710
8	0.8530
9	0.9540

The plot is useful in telling us which Lambda had the smallest mean residual, which in our analysis is the desired Lambda since it represents the smallest mean error among the predicted values. The plot was consistent with our Cross Validation with a Lambda of .0124.

With the Lambdas from the LASSO Cross Validation (CV model) and "grid search" now selected, we once more implement Lasso using training and testing sets, with 30 percent of the data making up the latter. Here, we not only compare model performance, but also observe which

variables are not reduced to zero. Below we see the top 20 coefficients by largest absolute values for our two selected Lambdas:

Table: Top 20 LASSO Coefficients for Grid Lambda (LHS) & CV Lambda (RHS)

Variable Description	Coefficients	|coefficients|
Bronx	126.3073	126.3073
Queens	-90.5754	90.5754
Educational attainment	54.8433	54.8433
February	-48.2677	48.2677
July	34.7064	34.7064
August	33.9600	33.9600
Hispanic share of population	32.2455	32.2455
April	-31.1925	31.1925
December	-27.1575	27.1575
October	23.1400	23.1400
March	-22.9379	22.9379
Staten Island	-22.7288	22.7288
Labor Force Participation Rate	22.3705	22.3705
Brooklyn	-22.2644	22.2644
Homeless Shelter	-21.4177	21.4177
Population Median Age	-18.7757	18.7757
November	-17.9293	17.9293
September	15.9425	15.9425
May	15.2539	15.2539
Bar/Nightclub	11.4519	11.4519

Variable Description	Coefficients	|coefficients|
Educational attainment	78.5106	78.5106
Hispanic share of population	52.8878	52.8878
Poverty Rate	48.4975	48.4975
February	-48.1643	48.1643
July	32.9494	32.9494
August	32.1918	32.1918
April	-31.1767	31.1767
December	-27.2281	27.2281
Asian share of population	-24.5061	24.5061
March	-22.8820	22.8820
October	21.4445	21.4445
November	-18.0216	18.0216
September	14.2069	14.2069
May	13.6080	13.6080
Labor force participation rate	9.8395	9.8395
June	8.2310	8.2310
Population Median Age	-3.9185	3.9185
Bar/Nightclub	3.5401	3.5401
Suspect under 18 yrs old	3.1802	3.1802
Transit crimes: Subway	2.5811	2.5811

The first aspect of the results that stands out is the number of times a month appeared in the list of coefficients, regardless of model. We notice that some months had positive coefficients and others negative, which may suggest that the level of crime is higher in certain months, such as July and August, while lower in others. Interestingly, the summer months of July and August had large positive coefficients, while November through February months had negative coefficients, suggesting a possible seasonal aspect to crime levels in NYC. Since the Winter can be very cold in this region, it is quite possible that possible that this leads to less crime since people are more likely to stay indoors.

We also note that several demographic indicators contributed to the crime rate, particularly in the CV model. In both models, a higher share of Hispanics in the population appeared to mean higher levels of crime. This is likely closely tied to a higher level of poverty and more precarious

financial situations, possibly leading this group to commit more crimes. We also see positive coefficients for the Labor Force Participation Rate and Educational attainment, a surprising result that appeared in both models. One might think that someone with more education or areas that have greater LFPR would have lower crime rates. However, the dataset does include a significant portion of crimes associated with businesses and corporations, particularly fraud. Since a person with more education might be more likely to have such jobs, this might be what the model is capturing. It is nevertheless a surprising result.

In terms of location, we see that the grid search model included all the boroughs that were not the reference. The Bronx simply appeared to contribute to a higher level of crime while the opposite was true for Queens and Staten Island when compared to Manhattan. The models also had large positive coefficients for crimes that occurred at Bars or Nightclubs and subways. These results make sense, but also point to a city that is appears to be more dangerous at night.

Table: LASSO Model Parameters & Evaluation

Model	Lambda	R-squared	RMSE	Intercept
Lasso CV Lambda	0.0802	0.8649	40.699	465.429
Grid Search Lambda	0.0124	0.8656	40.587	461.148

With the discussion of the coefficients, we turn to the performance of the models as a key sign of the effectiveness of the models. As seen in the table above, both models had an R-squared slightly above 86%, showing a strong fit to the data. In addition, the RMSE in both cases was at around 40 and the values of the intercept were similar. This points to a similar performance overall and suggest that LASSO was a strong fit for the data.

5.1.2. Ridge Models

For our Ridge models, we followed the same procedure as our Lasso models beginning with Cross validation for an optimal Lambda, returning a value of 10.0. This was much larger than the LASSO models. We than tested several possible Lambda values and plotted the means residuals:

Table: Results of Ridge "Grid Search"

Index	lambdas	Mean Residual
0	8.1	1645.56490
1	10.5	1645.56458
2	11.7	1645.56447
3	13.0	1645.56439
4	14.7	1645.56433
5	15.3	1645.56432
6	16.0	1645.56431
7	20.0	1645.56436
8	23.3	1645.56450
9	25.1	1645.56460

The values of the mean residuals were near identical, suggesting very little difference between the Lambda of choice. Nevertheless, we choose the Lambda with the smallest mean residuals, 16.0.

With our Lambda's selected, we re-run the models with our two selected Lambdas to compare. We will show the top 20 coefficients for Ridge "grid search" lambda. For the Ridge CV lambda, very few indicators were not equal to zero, suggesting the hyperparameter was not optimal and that this model was not as well suited to the data.

Table: Top 20 Coefficients for Ridge Regression with Grid Search Lambda

Variable Description	Coefficients	|coefficients|
February	-48.2831	48.2831
Educational attainment	41.3394	41.3394
Asian share of population	-37.5861	37.5861
July	35.0441	35.0441
August	34.2943	34.2943
April	-31.1800	31.1800
Labor Force Participation Rate	31.0924	31.0924
Poverty Rate	30.0481	30.0481
Hispanice share of population	29.0718	29.0718
December	-27.1336	27.1336
Homeless Shelter	-26.2357	26.2357
October	23.4733	23.4733
March	-22.9410	22.9410
Black share of population	-20.0872	20.0872
Brooklyn	-19.1231	19.1231
Population Median Age	-18.3010	18.3010
November	-17.8914	17.8914
Unemployment Rate	16.3032	16.3032
September	16.2820	16.2820
May	15.5703	15.5703

A key similarity with the grid search model for Ridge and the LASSO models is how frequently month dummy variables appear. In addition, we once more see a similar seasonality component with summer months having positive coefficients, strengthening this argument.

The demographic indicators also exhibited similar behavior as the LASSO models, with LFPR, Unemployment Rate, and Poverty rate all showing positive coefficients. In the case of the latter two, this suggests that areas with higher poverty and jobless rates are likely contributors to increasing the level of crime.

Location dummy variables for each borough also produced similar results, with Brooklyn and Queens having a lower crime level than the reference Manhattan. This model also pointed once more to homeless shelters having a negative coefficient, simply meaning that there is a lower level of crime here than on the reference dummy, "street". This possibly shows the benefit of these shelters, as many homeless people are far more vulnerable to crime on streets.

There were clear similarities with the Ridge Grid Search Model and the LASSO models, and this also proved to be the case with the overall model performance, as shown below.

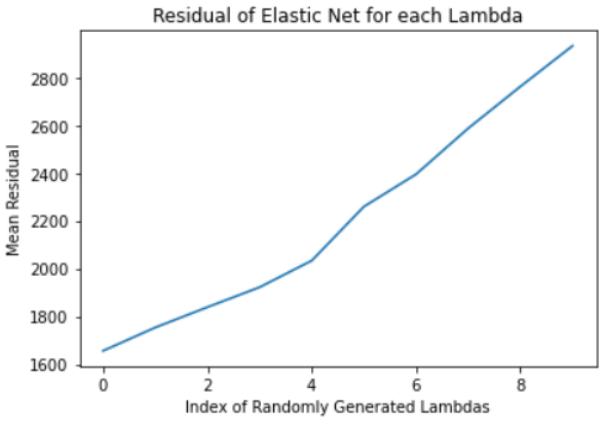
Table: Ridge Model Parameters & Evaluation

Model	Lambda	R-squared	RMSE	Intercept
Ridge CV Lambda	10.0	0.7564	54.653	464.715
Ridge Grid Search Lambda	16.0	0.8658	40.566	468.551

We can see that the grid search Lambda had an R-squared of just over 86 percent, but this dropped to 75.64 percent in the CV model. The RMSE was also much higher for the CV model, confirming it performed poorly in comparison to the Grid search lambda model.

5.1.3. Elastic Net Models

We repeated the same process one final time but for elastic net, which combines the penalty parameters from LASSO and Ridge to create a single parameter. With this, the goal is to observe if such a combination will boost model performance and perhaps confirm which variables are most significant. The initial cross validation returned a lambda of .1604. After repeating a grid search for values close to this, we obtained a lambda of .0153 based on the smallest mean residual of the lambdas tested:



Index	Lambda
0	0.0153
1	0.0841
2	0.1348
3	0.1858
4	0.2574
5	0.4131
6	0.5107
7	0.6522
8	0.7835
9	0.9133

With the selected lambdas, we re-implement Elastic Net and look to identify the top variables based on the size of the coefficient as shown in the table:

Table: Top 20 Elastic Net Coefficients for Grid Lambda (LHS) & CV Lambda (RHS)

Variable Description	Coefficients	|coefficients|
February	-43.2989	43.2989
Educational Attainment	41.3599	41.3599
Asian Share of Population	-37.6850	37.6850
July	32.5056	32.5056
August	31.8221	31.8221
Labor Force Participation Rate	30.7234	30.7234
Poverty rate	30.1157	30.1157
Hispanic Share of Population	29.1128	29.1128
April	-27.8762	27.8762
December	-24.2792	24.2792
October	21.8480	21.8480
March	-20.4177	20.4177
Black share of population	-19.6187	19.6187
Brooklyn	-18.7579	18.7579
Population Median Age	-17.9451	17.9451
Unemployment Rate	15.9357	15.9357
November	-15.9211	15.9211
September	15.1963	15.1963
May	14.5924	14.5924
Queens	-12.5452	12.5452

Variable Description	Coefficients	|coefficients|
Educational attainment	38.4415	38.4415
Asian share of population	-36.7188	36.7188
Poverty Rate	28.2675	28.2675
Labor Force Participation Rate	27.2998	27.2998
Hispanic Share of Population	26.5528	26.5528
February	-22.6558	22.6558
Black share of population	-18.0391	18.0391
July	17.9947	17.9947
August	17.6393	17.6393
Brooklyn	-17.0195	17.0195
Population Median Age	-16.9031	16.9031
Unemployment Rate	14.7215	14.7215
April	-14.6095	14.6095
December	-12.8537	12.8537
Queens	-12.4551	12.4551
October	11.8558	11.8558
March	-10.7712	10.7712
Bronx	9.1218	9.1218
November	-8.4437	8.4437
September	8.0092	8.0092

The results for the elastic net models proved to be very similar to the previous models, with many of the same variables selected. In addition, the value and direction of the coefficient is close to the LASSO models, with particularly similarities for the month dummy variables. The similarities

may be further evidence of the validity of the results and may also serve as a confirmation of some of the more surprising results. We not only saw similarities in the coefficients, but also in the model performance:

Table: Elastic Net Model Parameters & Evaluation

Model	Lambda	R-squared	RMSE	Intercept
Elastic Net CV Lambda	0.1605	0.8466	43.364	470.476
Elastic Net Grid Search Lambda	0.0153	0.8649	40.694	469.305

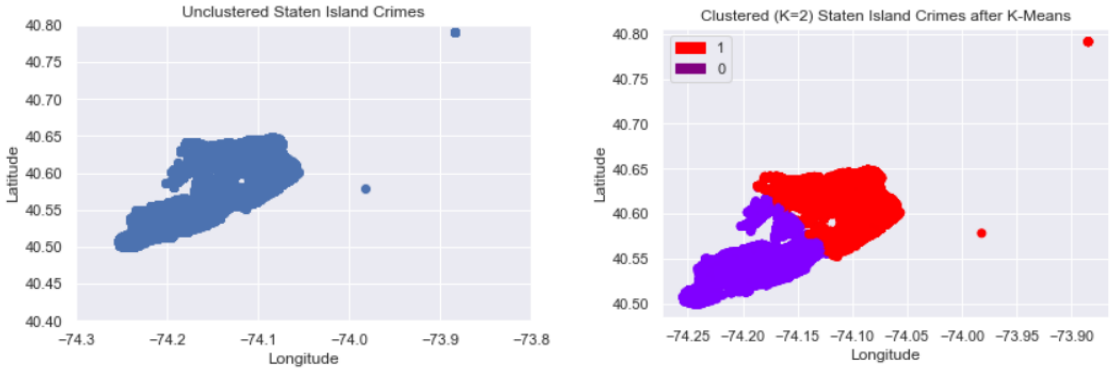
Both models had an R-squared of at least 84%, with the CV model performing slightly worse in this aspect as well as in RMSE. This confirms that the grid search model appears to be a stronger performing model.

5.2. Crime Location Prediction Classification Models

Like our regularization models, we will first discuss the process of preparing and running the models for Staten Island. We start by describing the process of creating clusters using K-means, before then applying the classification models previously discussed. With all this, we evaluate the model by using confusion matrices and evaluating the general accuracy of the models, as well as the per-class precision and recall rates. The number of observations varied by borough and these models have the same number of independent variables as the regularization techniques at 126. However, crime rate is an predictor, while binary location classes are the output.

5.2.1. Staten Island

To create our clusters, we implement k-means using only longitude and latitude while specifying that we wish to have two clusters outputted. We compare the un-clustered data to the clustered data after K-means:



We can now see the clear division in the clusters, with "0" being purple and mainly representing the western part of the borough. The red or eastern portion is labeled as "1". The Staten Island portion of the data consisted of 77,474 observations, with 58,638 labeled as "1" and the rest as "0".

K-means makes the use of classification models quite easy, which is our next step, as we now test the data Logistic Regression, k-NN, and LDA. The implementation of these techniques was simpler than in the crime rate section, but some hyperparameters needed to be specified regardless. For Logistic Regression, we added a parameter that essentially acts as Lasso by setting alpha equal to ten. Although variable selection is not the focus of this section, given that these models do have a fairly large number of indicators, the addition of this parameter may help to focus on the most important variables. The other key hyperparameter that we specified was the numbest of nearest neighbors. We selected this using the equation $K = sqrt(n)$, where k is the number of nearest neighbors and n is the number of observations. Given the size of the data, K will be large at 277.

To assess the model, we utilize a popular evaluation metrics for classification methods. These are confusion matrices, precision and recall rates, and accuracy, including a 10-fold CV to test the model's robustness. Confusion matrices directly allow us to see the predicted values. Based on this, the precision and recall rates are calculated. These are critical metrics in classification models, because simply calculating the accuracy does not reveal the performance of the predictions. This is revealed with the precision and recall, with the former showing the number of relevant results per class and the latter showing the number of relevant results correctly classified. In the case of crime, both are relevant in assessing how well the model can predict a location, with recall showing if a model can tell where a crime occurred. Below, we show these metrics for Staten Island:

Tables: Model Results for Staten Island

Logit Staten Island	Predicted:0	Predicted: 1	Performance
Actual: 0	370	5,242	
Actual: 1	329	17,302	
Accuracy			75.9%
10-fold CV accuracy			76.0%

	Precision	Recall
"0": Purple	79%	23%
"1": Red	52%	93%

k-NN Staten Island	Predicted:0	Predicted: 1	Performance
Actual: 0	370	5,242	
Actual: 1	329	17,302	
Accuracy			75.9%
10-fold CV accuracy			76.0%

	Precision	Recall
"0": Purple	53%	7%
"1": Red	77%	98%

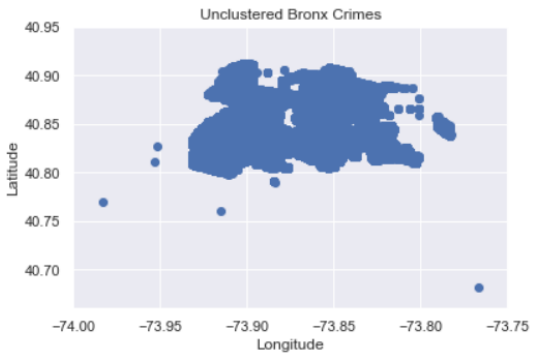
LDA Staten Island	Predicted:0	Predicted: 1	Performance
Actual: 0	111	5,510	
Actual: 1	57	17,565	
Accuracy			76.0%
10-fold CV accuracy			75.9%

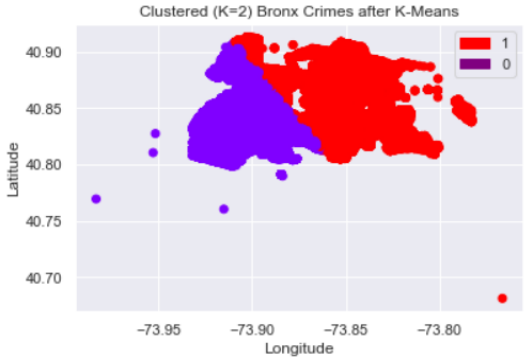
	Precision	Recall
"0": Purple	66%	2%
"1": Red	76%	100%

In terms of model accuracy, the Cross validation and overall accuracy was consistently above 75 percent in each model. This suggests a strong performing model that captured the data well and could generally predict where a crime would occur within Staten Island. There was greater variation in the precision and recall rates, however. The k-NN and LDA model had had relatively high precision rates and very high recall for the "Red" class, which represented the northeastern part of the borough. In this case, the precision tells us the number of results that are relevant while the recall shows the percentage that were correctly classified. When looking at the confusion matrix, we can quickly note why the recall rates were so high, as very few observations were predicted as zero. This does hurt the model's effectiveness and is due to the number of observations labeled as "1" in the Staten Island data. In this regard, the logistic regression model performed better for the "0" class, suggesting this might be the best performing model for Staten Island.

5.2.2. Bronx

We follow the same steps as Staten Island, using K-means to generate two clusters for out classification models:





Eastern Bronx was labeled "1" while the west was labeled "0", with 131,790 observations making up the former class and 260,843 the latter. With this, we once more use our three classification models (k=625 for k-NN) and show the results below.

Tables: Model Results for the Bronx

Logit Bronx	Predicted:0	Predicted: 1	Performance
Actual: 0	74,671	3,653	
Actual: 1	32,689	6,777	
Accuracy			69.1%
10-fold CV accuracy			69.2%

	Precision	Recall
"0": Purple	70%	95%
"1": Red	65%	17%

k-NN Bronx	Predicted:0	Predicted: 1	Performance
Actual: 0	77,879	445	
Actual: 1	37,711	1,755	
Accuracy			67.6%
10-fold CV accuracy			67.6%

	Precision	Recall
"0": Purple	67%	99%
"1": Red	80%	4%

LDA Bronx	Predicted:0	Predicted: 1	Performance
Actual: 0	76,860	1,464	
Actual: 1	35,404	4,062	
Accuracy			68.7%
10-fold CV accuracy			68.6%

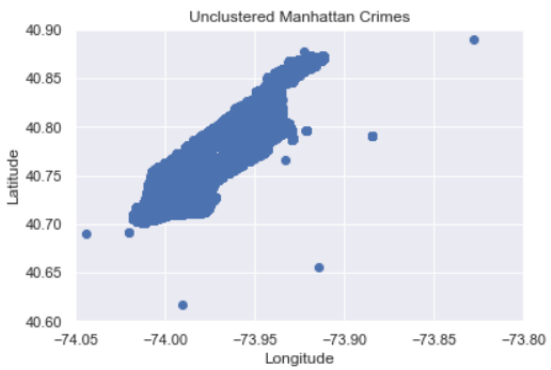
	Precision	Recall
"0": Purple	68%	98%
"1": Red	74%	10%

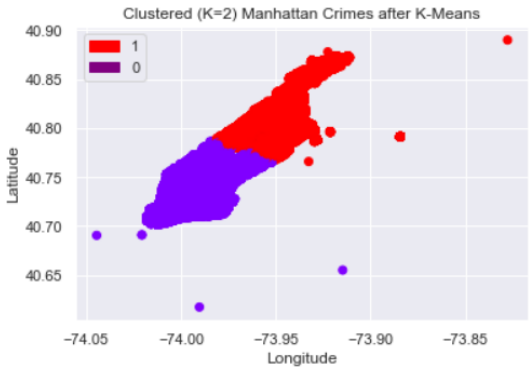
Both accuracy metrics for the three models ranged between 67 and 70 percent for the Bronx, a slight dip from Staten Island. In addition, the precision rates were more stable for the Bronx for both labeled classes. However, we did continue to see extremely high recall rates for one class and very low recall for others, with the "0" class having excellent recall rates and the confusion matrix shows why, as few observations were predicted as "1" compared to the other class. This once more points to issues

the model faces due to one class having far more observations and might suggest that additional clusters could lead to more consistent recall rates.

5.2.3. Manhattan

We now move to the third of the five boroughs, Manhattan. After once more implementing k-means, we compare the un-clustered plot to the clustered data. As we see below, the northeastern part of Manhattan, consisting of 194,942 observations, is labeled as "1" and the southwest "0". The "0" class is made up of 243,618 observations:





We now implement the same three models (k=661 for k-NN) and show the results below:

Tables: Model Results for Manhattan

Logit Manhattan	Predicted:0	Predicted: 1	Performance
Actual: 0	55,812	17,547	
Actual: 1	22,215	35,994	
Accuracy			69.8%
10-fold CV accuracy			69.8%

	Precision	Recall
"0": Purple	72%	76%
"1": Red	67%	62%

k-NN Manhattan	Predicted:0	Predicted: 1	Performance
Actual: 0	56,597	16,762	
Actual: 1	25,558	32,561	
Accuracy			67.8%
10-fold CV accuracy			67.9%

	Precision	Recall
"0": Purple	69%	77%
"1": Red	66%	56%

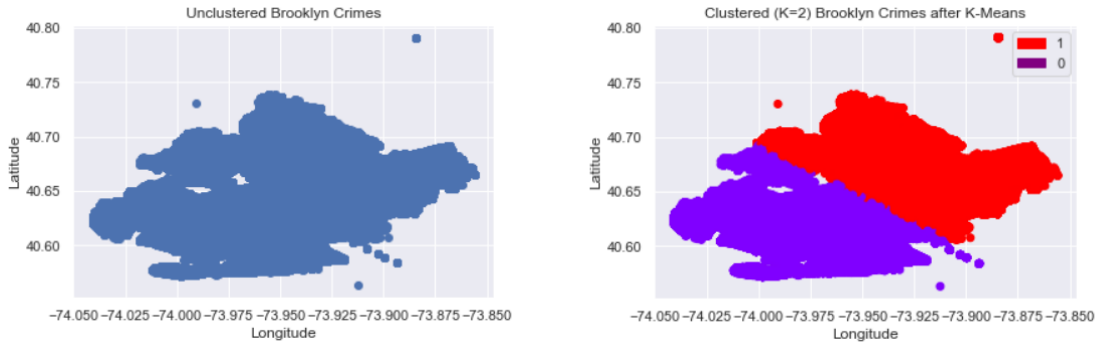
LDA Manhattan	Predicted:0	Predicted: 1	Performance
Actual: 0	59,037	14,322	
Actual: 1	26,078	32,131	
Accuracy			69.3%
10-fold CV accuracy			69.4%

	Precision	Recall
"0": Purple	69%	80%
"1": Red	69%	55%

Manhattan had very similar accuracy to the Bronx, which also means they performed worse in this aspect compared to Staten Island. However, Manhattan far outperformed the prior boroughs discussed in recall, showing a far more stable and consistent models, regardless of the technique used. In general, the three models predicted crimes in southwest Manhattan, but the difference between classes was not as lopsided as before. This might owe to the fact that Manhattan had far more observations, which allowed for the model to be trained more effectively and in turn produced better results.

5.2.4. Brooklyn

We now turn to NYC's most populous borough, Brooklyn, and perform k-means to generate the clusters:



In this case, the east and northeast of Brooklyn are labeled "1" (348,069 observations), with the rest (173,553 observations) "0". We then run the same models (k=721 for k-NN) and show the results:

Tables: Model Results for Brooklyn

Logit Brooklyn	Predicted:0	Predicted: 1	Performance
Actual: 0	19,915	32,037	
Actual: 1	10,003	94,532	
Accuracy			73.1%
10-fold CV accuracy			73.1%

	Precision	Recall
"0": Purple	67%	38%
"1": Red	75%	90%

k-NN Brooklyn	Predicted:0	Predicted: 1	Performance
Actual: 0	15,448	36,504	
Actual: 1	6,820	97,715	
Accuracy			72.3%
10-fold CV accuracy			72.4%

	Precision	Recall
"0": Purple	69%	30%
"1": Red	73%	93%

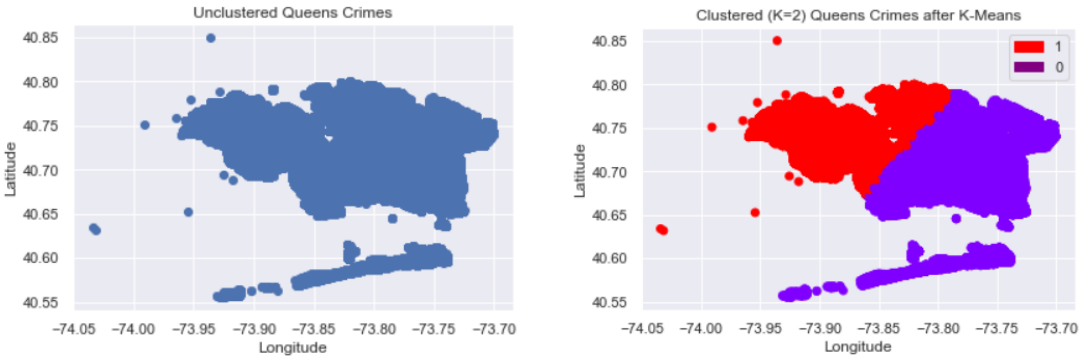
LDA Brooklyn	Predicted:0	Predicted: 1	Performance
Actual: 0	8,766	43,186	
Actual: 1	2,991	101,544	
Accuracy			70.5%
10-fold CV accuracy			70.4%

	Precision	Recall
"0": Purple	75%	17%
"1": Red	70%	97%

The overall accuracies of the three models ranged between 70 and 73 percent, slightly higher than the Bronx and Manhattan, but lower than Staten Island. We also see once more that the recall rates for one class were much lower than the other class, with "1" often being successfully classified as seen in their recall rates. For "0", recall rates were lower, with the Logistic Regression and k-NN model performing better in this regard. The precision rates were consistent all around, but these models once more bring up concerns about too many observations being assigned to one class by K-means.

5.2.5. Queens

The last of the five boroughs, queens, is also the largest borough by area in the city. With this, we plot the data and its transformation to clusters:



For Queens, the eastern and southern parts (180,742 observations) were labeled "1", with the rest of the borough (179,725 observations) "0". The k-means algorithm most evenly clustered this borough based on the number of observations in each class, a problem that had arose in the previous boroughs discussed. With this in mind, we implement our models (k=599 for k-NN) one last time:

Tables: Model Results for Queens

Logit Queens	Predicted:0	Predicted: 1	Performance
Actual: 0	34,446	19,472	
Actual: 1	13,410	40,913	
Accuracy			69.6%
10-fold CV accuracy			69.3%

	Precision	Recall
"0": Purple	44%	13%
"1": Red	49%	84%

k-NN Queens	Predicted:0	Predicted: 1	Performance
Actual: 0	33,721	20,197	
Actual: 1	13,284	40,939	
Accuracy			69.0%
10-fold CV accuracy			68.8%

	Precision	Recall
"0": Purple	72%	63%
"1": Red	67%	76%

LDA Queens	Predicted:0	Predicted: 1	Performance
Actual: 0	37,519	16,399	
Actual: 1	17,437	36,786	
Accuracy			68.7%
10-fold CV accuracy			68.3%

	Precision	Recall
"0": Purple	72%	63%
"1": Red	67%	76%

The accuracy of the models was in line with the Bronx and Manhattan, but focusing on the precision and recall rates, the logistic regression was the worst performing model overall. Even though it had the highest recall rate for the "1" class, its recall rates and the precision of the model were still low. The k-NN and LDA models had far higher and more consistent precision and recall rates. This suggest that these models were well suited to Queens, but also appears to confirm the benefit of k-means more evenly clustering the data, a key issue in some of the other boroughs.

6. Conclusion

Our analysis of crime in New York City provided insightful and potentially useful information that can be useful in the never-ending battle against crime. Predicting when a crime will happen and what crime will occur is arguably impossible, but the use of modern statistical models and the increasing availability and accuracy of data offers significant promise in combatting crime.

A 2015 paper on crime analysis (Jayaweera et al. 2015) stated that the increasing global population and complex geographical diversity of crime makes proper data collection and analysis critical for law enforcement to understand a crime landscape that constantly changes. That was a primary goal of our regularization techniques, which looked to choose variables that had a large impact on the level of crime. The models appeared to show a strong seasonal component, with the summer having a higher level of crime compared to the winter months. In addition, we gained insight on certain demographic and location indicators that appeared to contribute to the crime level. This is useful for law enforcement and policy makers alike because it can help to more efficiently allocate resources that are often scarce.

The borough-by-borough analysis using classification techniques focused on location prediction, leveraging modern techniques to do what could be incredibly useful, but what is also very difficult. The models had solid accuracy, but in several boroughs, the model's effectiveness did appear to be hurt by one class having far more observations. In Manhattan and Queens, where classes were more evenly divided, the model performed far better overall. This might suggest a benefit of additional clusters that might bring more consistent and stable models in terms of precision and recall, two very important metrics for classification techniques.

Crime analytics is a growing area, and its applications are extensive, with the potential for far more analysis that was out of the scope of this paper. For example, it might be possible to leverage the raw GPS coordinates, potentially leading to location prediction on an individual level. In addition, unsupervised algorithms could be incredibly useful in identifying trends in crime (Reddy et al. 2021). Such applications, along with our models, show the promise and effectiveness of crime analytics in fighting crime.

7. References

[1] Bureau of Justice Statistics. 2012. "Victimizations Not Reported to the Police, 2006-2010". Archived | Victimizations Not Reported to the Police, 2006-2010 | Bureau of Justice Statistics (ojp.gov)

[2] City of New York. 2021. "NYPD Complaint Data Historic". https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i

[3] Isuru Jayaweera; Chamath Sajeewa; Sampath Liyanage; Tharindu Wijewardane; Indika Perera; Adeesha Wijayasiri. 2014. "Crime Analytics: Analysis of Crime Through Newspaper Articles". IEEE. https://ieeexplore.ieee.org/document/7112359

[4] Shraddha Reddy; Mohan Rao; Nikhila K. 2021. "Analysis of Crime". ICICNIS 2020. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3769891

[5] State of California Department of Justice. 2014. "2014 Computational Formulas". https://oag.ca.gov/sites/all/files/agweb/pdfs/cjsc/stats/computational_formulas.pdf

[6] U.S Census Bureau, 2018, "2017 American Community Survey 1-year Estimates". https://www.census.gov/programs-surveys/acs/technical-documentation/table-and-geography-changes/2017/1-year.html

[7] U.S Census Bureau, 2019, "2018 American Community Survey 1-year Estimates". https://www.census.gov/newsroom/press-kits/2019/acs-1year.html

[8] U.S Census Bureau, 2020, "2019 American Community Survey 1-year Estimates". https://www.census.gov/newsroom/press-kits/2020/acs-1year.html

[9] U.S Census Bureau. 2021, "2020 Census Redistricting data press release". https://www.census.gov/newsroom/press-kits/2021/2020-census-redistricting.html

[10] U.S Census Bureau, 2021. https://data.census.gov/cedsci/