MACHINE LEARNING FINAL PROJECT

QUICK DECISION on Crime

(Threat-based Machine learning model to help NYPD)

Prasanth Bathae Kumaresh
Nancy Yang
Hao Zheng

Table of Contents

Background and Motivation *4*

Discussion about questions *5*

Limitations of the model **6**

Discussion about our supervised learning workflow and models *7*

Feature selection **7**

Gender 7

Race 7

Location 8

Time period 8

Age 8

Feature pre-processing **8**

Hyperparameter tuning **9**

Model selection **9**

Quantitative choice 10

Qualitative choice 10

Human choice 10

Communication of results, and advice to a non-expert *11*

Model Usage **11**

Sample results 11

Summary **12**

APPENDIX *13*

Table of Figures

Figure 1. Distribution of race 5
Figure 2. Sample scenario results 12
Figure 3. Incidence of murders among all crimes 15
Figure 4. Distribution of Gender 15
Figure 5. Distribution of Boroughs 16
Figure 6. Distribution of Time periods 16
Figure 7. Distribution of Age groups 17
Figure 8. Heatmap of crimes 17
Figure 9. Flowchart - Part A 18
Figure 10. Flowchart - Part B 18
Figure 11. Feature importance – Black model (Descending order) 19
Figure 12. Feature importance - Rest (non-black) model (Descending order) 20

Table of Tables

Table 1. Categorization of time of day 8
Table 2. Models for Hyperparameter tuning 9
Table 3. Different scenarios for the best model 10
Table 4. Feature Selection 13
Table 5. Feature Preprocessing 14
Table 6. Quick Decision - Drop-down menu options 14

Background and Motivation

For a society and by extension, a country to be successful and prosperous, the key areas it has to have are a robust system of upholding the rule of law, maintaining property and personal rights, having appropriate checks and balances to mitigate any one area of the governments taking control. In that regard, the Police Department (PD) has an important job of maintaining law and order in the city and providing a safe social environment and a peace of mind to all its residents.

As the population grows and cuts to public funding increases, PDs are increasingly strained for resources in resolving all crimes in their midst. Shortage of manpower leads to some crimes not being followed up to their finish and end up as cold cases. This lack of human resources is magnified in a big city such as New York City. The New York Police Department (NYPD) is constantly looking for ways to improve their operations in order to effectively tackle all the potential crimes happening around them.

When the NYPD receives a 911 emergency call, the operators take down the details of the incident and pass on the message to the officers on the beat. They go to investigate the issue and depending on the seriousness of the incident, backup officers are requested by those on the ground. This process takes time, which is the most important factor in reducing or eliminating crimes.

Ideally, the NYPD would like to send out 4 officers or more to all incoming 911 calls. It is not possible, as it is cost intensive and not an efficient usage of resources. Instead of having a standard procedure of sending 2 officers to the potential crime scene, it is beneficial to have a prediction system which can identify serious crimes such as potential of a murder occurring, so that NYPD can send in more officers. This efficient allocation of resources based on the seriousness of the potential crimes will go a long way in reducing the number of serious crimes without sacrificing the focus on non-violent crimes.

We used the crime database from the NYPD[1] and downloaded their csv file as the source for our investigation. This file has all the crimes committed in the various boroughs of New York City (NYC) in addition to the demographic features of the criminals.

One thing which jumps out when we look at the data is the disproportionate number of ‘black’ (~75%) people arrested than all the other races (Figure 1) .Our model has to remove the obvious bias in the criminal justice system and provide a solution commensurate with the crime and not just based on the race of the assailant(s). We will be using demographic information such as age, gender and race along with the location and time of day in our prediction model. We have curated our features accordingly from the available input dataset.

[1] https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic

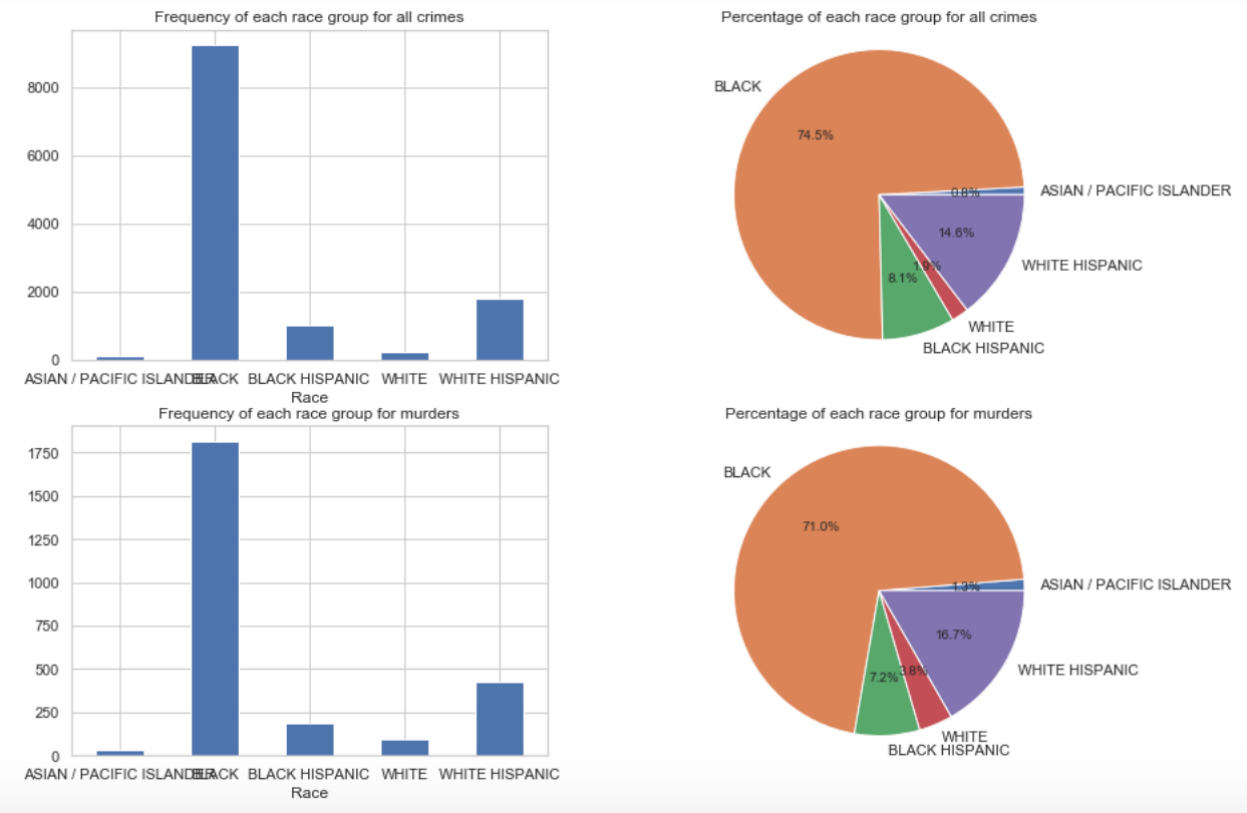

Figure 1. Distribution of race

Discussion about questions

Business question we are trying to answer is:

> How should the NYPD efficiently allocate their officers to incoming 911 calls which describe demographic features of the person causing the disturbance along with the location and time of the potential crime?

Our statistical question to address the business question is:

> Can a statistical model give prediction probabilities of a potential violent crime (murder) based on the demographic and locational information described in a 911 call?

The statistical model can be used as a useful supportive tool in decision making. The model uses all the information that can be easily gathered from a phone call and can give a probability almost immediately after they are fed to the underlying model. Based on the available resources, the NYPD can decide on the cutoff for probabilities to decide on the number of officers to be assigned to that case. Reports that predict a higher chance of a murder will get a

SWAT team sent while the ones at the lower end will receive the standard 2 officers to check on them. This scaled allocation of resources will ensure differentiated treatment based on the threat level, such that more officers are not unnecessarily deployed for non-serious crimes. This statistical model will go a long way in optimizing the available human resources without reducing the quality of service to the public and can be really helpful in improving the overall efficiency of the police department.

Limitations of the model

However, this model is not perfect, there are issues that can't be explained by the statistical model:

1. Statistical methods predict probabilities based on previous crime data and they don't necessarily give appropriate weightage to the recent events. For example, location of the crime is one of the features used in the model. In real world scenarios, "dangerous places" in the city might change over time due to the gangs moving around the city as a result of increased police activities. This would lead to inaccuracies in the predicted probabilities of violent crimes
2. Superficial features are not the sole features that dictate serious crimes. According to Olivia Goldhill[2], societal pressure is a major behind the scenes contributor to someone becoming a criminal. Societal pressure comes from cultural norms which can vary over time periods and cultures, and also from stress experienced by that person. Those attributes are difficult to model, since under the right circumstances, most people can commit a crime.
3. Our model helps in the efficient allocation of human resources and not necessarily to lower the costs to PD. Incremental cost in sending 2 more officers from 2 to 4 is different from 4 to 6, as the latter case involves having 2 patrol cars instead of one. Furthermore, more research should be done to evaluate the cost-benefit analysis of increasing the team size of the officers, its impact on the morale of the officers, reduction in potential crimes etc.
4. Statistical models will inherit the biases in the underlying data. If no action is done to rectify it, model predictions will further propagate the systematic biases, thereby forming a closed feedback loop

[2] Goldhill, O. (2018, August 3). A neuroscientist who studies rage says we're all capable of doing something terrible. Retrieved from https://qz.com/1348203/a-neuroscientist-who-studies-rage-says-were-all-capable-of-doing-something-terrible/

Discussion about our supervised learning workflow and models

Feature selection

The first step of feature selection is to use our intuition to check for relevant variables for our model by looking at the original data set. We are going to check for the probability of murder happening from our dataset. We eliminated several unrelated variables such as ID of the crime case, date of the crime, victim information, jurisdiction_code etc. (Table 4).

Then we do exploratory data analysis (EDA) by plotting the features used in the model to find any trends (or) outliers in the data, the results of which are given below. We compared the features for the overall crime (entire) data with that of 'murder' data. Almost 19% of our data has incidence of murder in it (Figure 3).

Note: We have excluded 'Unknown' values in our EDA to have a better understanding of the underlying data.

Gender

Gender plots show the majority of crimes committed are by men (Figure 4). One interesting observation is the slight increase in % of women when the overall crimes are limited to just murders. This can lead to a bias in the model for women so much so that when the perpetuator is a female, the model can predict that they are more likely to commit a murder.

Race

Race has a significant difference between various races (Figure 1). Data says that Black people are more likely to commit a crime than all the other races combined. This might be due to the "stop and frisk" policy by the NYPD where the majority of people that were stopped and searched from people of African-American (Black) or Latino background. This can be seen in the disproportional % of those races in criminal data when compared with their population in the city [3,4,5] We have to address this bias when we build our model.

[3] "Stop-and-Frisk Data". *New York Civil Liberties Union*. January 2, 2012. Retrieved November 30, 2019.

[4] QuickFacts for New York City / New York State / United States, United States Census Bureau. Retrieved February 9, 2017

[5] O'Neill, James P. "Crime and Enforcement Activity in New York City" (PDF). *NYPD*. Retrieved November 30, 2019.

Location

The plots show Brooklyn and Bronx have the highest concentration of crimes and murders from the 5 New York City Boroughs (Figure 5).

Time period

Time period is another interesting indicator of both crime and murder. As per our intuition, both crime and murder are more likely to happen during the sleeping hours (10PM - 6AM) (Figure 6). This indicates that crime reports received in sleeping hours should be taken more seriously.

Age

Younger people (18-24 and 25-44) have the highest proportion of crimes and murders in the city. Proportion of older people (45-64 and 65+) increases when the overall crimes are limited to only murders (Figure 7).

Generally, for crimes that are less likely to involve a murder, the NYPD will only send out few officers first, and further reinforcements are sent if needed; for crimes that are more likely to involve a murder, the department will sent out SWAT team or more officers to the place where a crime or a possible crime is being reported.

Feature pre-processing

Some of the features are renamed to more descriptive names (Table 5).

Following preprocessing are done:

- Observations with NA values are removed and re-indexed
- Age_groups equal to 1020, 940 and 224 are removed
- Time and Hour are calculated from the renamed 'Time' feature
- Hour is further categorized into 3 groups:

HOUR	PERIOD
22 - 0	Sleeping hours
6 - 18	Working hours
18 - 22	Afterwork

Table 1. Categorization of time of day

- Heatmap of crimes is created using google maps with latitude and longitude data (Figure 8).

Hyperparameter tuning

Our entire process is shown in a flowchart (Figure 9 and Figure 10)

Our entire dataset was divided into 80% training and 20% testing with a random seed of 123. We used 'Pipeline' in combination with 'GridSearchCV' with 10 folds to find our best model using training data. It was run with the following models with different hyperparameters to find the best combination which yielded the lowest training and cross-validation error rates. We have limited ourselves to models which gave us probabilities for predictions, as we will use these probability values to make our recommendations to the client (Table 2).

MODEL NAME	HYPERPARAMETER TUNING
Logistic regression	c = [0.01, 0.1, 1]
Multinomial Naive Bayes	alpha = [0.01, 0.1, 1, 10]
Random Forest	max_depth = None n_estimators = [50, 100, 150, 200]

Table 2. Models for Hyperparameter tuning

Model selection

Best model with the lowest error rate was

Logistic regression model with c = 0.01

This best model is then applied in different scenarios (Table 3)

Option #	SCENARIO	TRAINING ERROR RATE	VALIDATION ERROR RATE	NOTES
1	Grouping all data together	0.18638	0.18638	Training data is used as is
2	One model for each race	Asian = 0.31459 Black = 0.19591 Black Hispanic = 0.17896 White = 0.37368 White Hispanic = 0.2357 Unknown = 0.05936	Asian = 0.31333 Black = 0.19591 Black Hispanic = 0.17894 White = 0.37363 White Hispanic = 0.2357 Unknown = 0.05934	Model built and tested for each race individually
3	One for Black, rest grouped together	Black = 0.19591 Rest = 0.16868	Black = 0.19591 Rest = 0.16868	Black crimes are kept separately while the rest of the races are

				combined
4	Model having same row count for each race as that of a smallest race	0.20599	0.20601	Row counts are normalized by taking the same # of rows for each race (= # of rows in the smallest group)

Table 3. Different scenarios for the best model

Quantitative choice

Training errors and cross-validation errors for the chosen Logistic regression model were compared for all scenarios. The objective was to choose a scenario who yielded the lowest cross-validation error rate without adding unnecessary complexity in the model. Based on the results, Option # 3 was selected as it had one of the lowest error rates, provided us an opportunity to assign different sets of thresholds for black and non-black crime being reported to estimate the danger level.

Test error rate using our final model is 0.19838

Qualitative choice

Our decision to divide our model into ‘black’ and ‘non-black’ perpetuators is driven by the huge population of ‘black’ entries in our data (figure). Significant (71-75%) percentage of the population in the dataset is black, so we did not want our model to be biased, therefore we created two models to address this issue - one model where the assailants are black and another when they are from a different race. By having different sets of thresholds, we can reduce the bias in the input dataset and send the correct number of officers to the crime scene.

Human choice

Our logistic regression model helps to predict the probabilities of a murder happening given our input features. This is more intuitive as we can see the impact of individual parameters on the overall result. By having all the input feature variables as dummy variables, we can see the impact of each variable on the overall probability of a prospective murder.

Impact of different features on both ‘Black’ and ‘Rest (non-black)’ models are shown in Figure 11 and Figure 12.

Communication of results, and advice to a non-expert

The number of police officers sent and whether or not to send a SWAT team should be based on the threat level of the suspect. The threat level of the suspect in the model is defined by the suspect's predicted intention of committing a murder. Our model uses collected demographic features including age, gender, race along with location and time of day to calculate the probability of the suspect committing a murder, and use the probability to assign the threat level of the suspect.

NYPD answers all crime reports with an initial investigation group of at least 2 officers. As the threat level of suspects increases, more officers are deployed. For suspects with the highest threat level, SWAT team will be sent to neutralize the threat. The default concept behind the model is that the NYPD wants to send 2 officers to 60% of low threat crime reports; 4 officers to 25% (60-85%) of medium threat crime reports; 8 officers to 10% (85-95%) of high threat crime reports and SWAT team to 5% (>95%) of extremely high threat crime reports. If the NYPD wants to change the percentage, they can avail the services of a programmer to tune the appropriate percentages in the model.

Model Usage

Police officers will have five features which have their drop-down lists (Table 6). After inputting these five variables and selecting "OK", the model will return the threat level and recommended number of officers to be sent to that location.

This model can become an extension in the current 911 call handling system with a self-describing name such as "**Quick Decision**". When the telephone operators receive the phone call, and record age, gender, race, time zone when receiving the call and the location of the suspects, they can be linked to the "Quick Decision" tool in order to avoid re-entering the same data. It can be further expanded by using telephone area code and current time to prepopulate the location and time features, so that it will save valuable time in the data entry process. At the end of the data collection, our tool will show the predicted probability of murder and the appropriate number of officers to be sent.

Sample results

Here are the results from some sample scenarios:

```
The suspect is M BLACK in QUEENS during time period of Working hours with age about: 65+
The probablity of murder is : 0.19
The threat level is Low. We should send 2 people.
```

```
The suspect is M BLACK HISPANIC in BROOKLYN during time period of Afterwork with age about: 45-64
The probablity of murder is : 0.23
The threat level is Medium. We should send 4 people.
```

```
The suspect is Unknown_Sex BLACK in BROOKLYN during time period of Afterwork with age about: 25-44
The probablity of murder is : 0.31
The threat level is extremely High. We should send SWAT team.
```

Figure 2. Sample scenario results

Summary

This model provides a quick recommendation while adding little overhead. Combined with the field experiences of the officers and their recommendations, this model can be periodically tuned to minimize the error in predicted probabilities. Thus, this dynamic model can assist in improving operations of the NYPD, eliminating waste, assigning the right number of officers to the threat level without compromising on quality, hence it is a Win-Win for everyone involved - NYPD, General Public and the Government.

APPENDIX

Variable in the original dataset	Keep in the prediction model?
Incident_key	
Occur_date	
Occur_time	✅
Boro	✅
Precinct	
Jurisdiction_code	
Location_desc	
Statistical_murder_flag	✅
Perp_age_group	✅
Perp_sex	✅
Perp_race	✅
Vic_age_group	
Vic_sex	
Vic_race	
X_coord_cd	
Y_coord_cd	
Latitude	✅
Longitude	✅

Table 4. Feature Selection

ORIGINAL FEATURE NAME	RENAMED FEATURE NAME
OCCUR_TIME	Time
BORO	Location
PERP_AGE_GROUP	Age_group
PERP_RACE	Sex
STATISTICAL_MURDER_FLAG	Is.Murder
Latitude	
Longitude	

Table 5. Feature Preprocessing

FEATURE	DROP-DOWN LIST OPTIONS
Age	1. “<18”, 2. “18-24”, 3. “25-44”, 4. “45-64”, 5. “65+” and 6. “Unknown”
Gender	1. “Male”, 2. “Female” and 3. “Unknown”
Location	1. ”Bronx”, 2. Queens”, 3. “Brooklyn”, 4. “Manhattan” and 5. “Staten Island”
Time Zone	1. “Sleeping Hour (22 - 0)”, 2. “After work (22 - 0)” and 3. “Working Hour (6 - 18)”
Race	1. “Black”, 2. “White”, 3. “Black Hispanic”, 4. “Asian /Pacific Islander”, 5. “White Hispanic”, 6. “American Indian /Alaskan Native” and 7. “Unknown”

Table 6. Quick Decision - Drop-down menu options

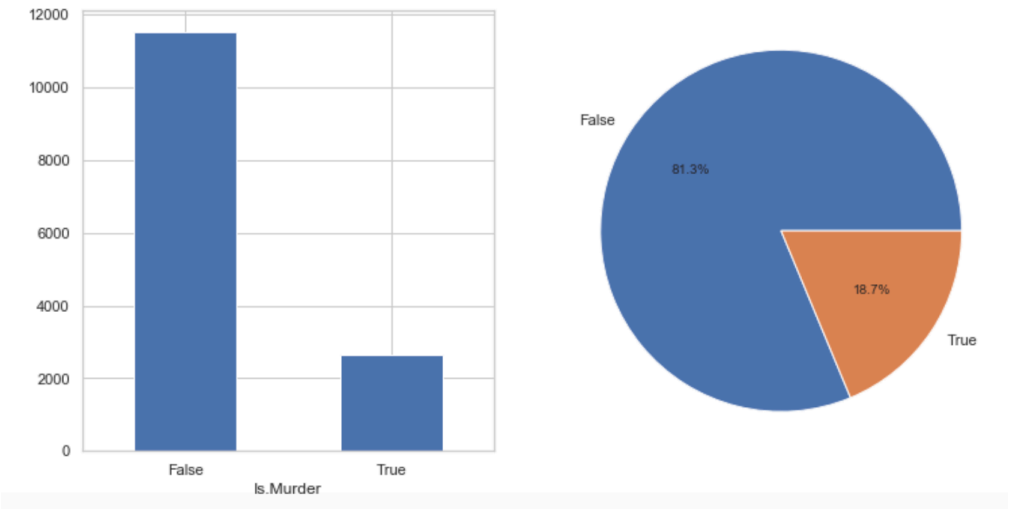

Figure 3. Incidence of murders among all crimes

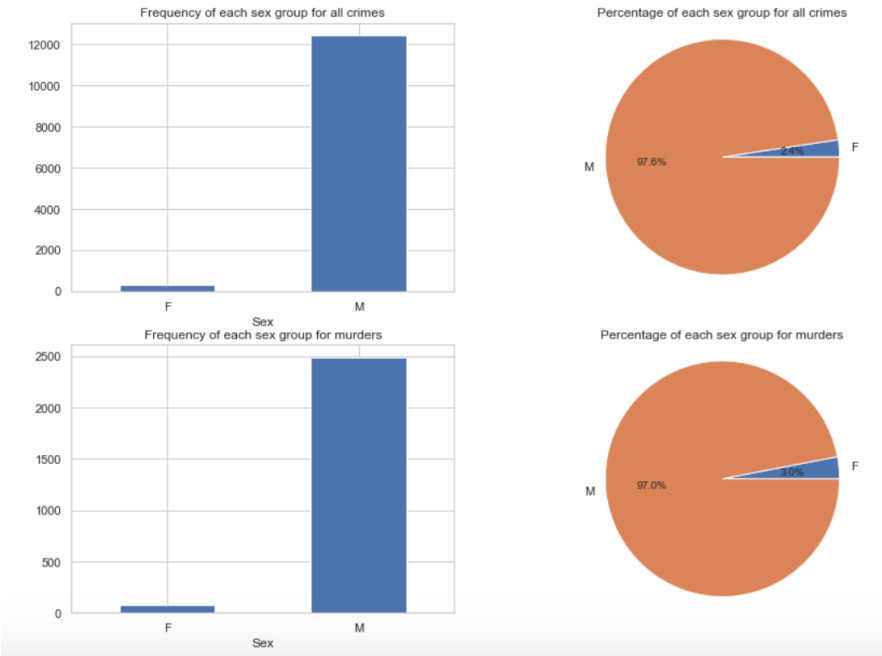

Figure 4. Distribution of Gender

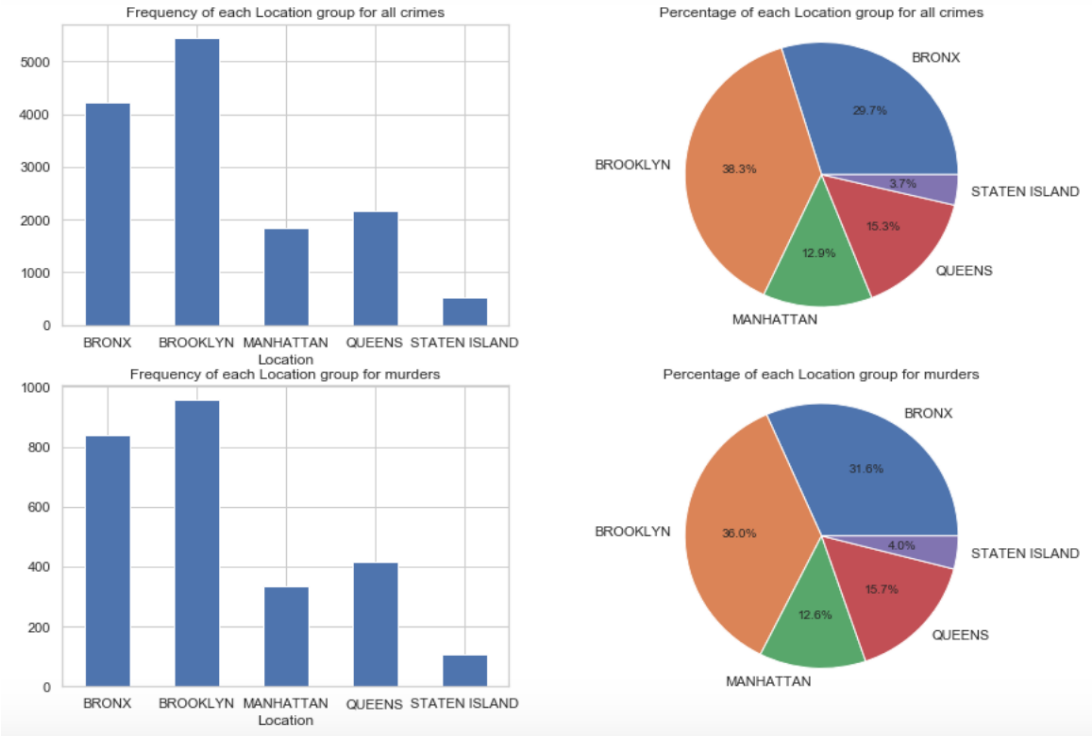

Figure 5. Distribution of Boroughs

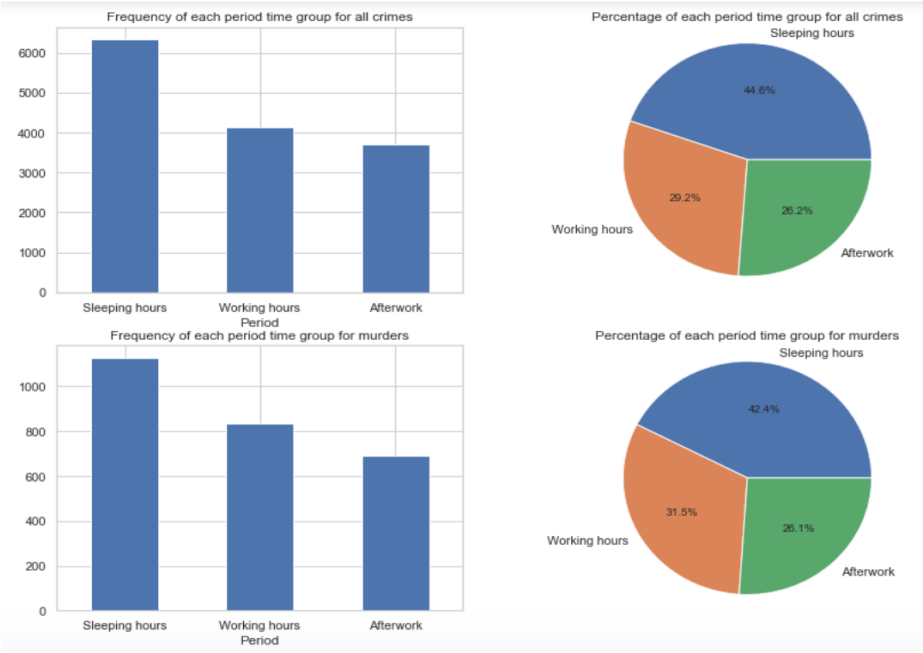

Figure 6. Distribution of Time periods

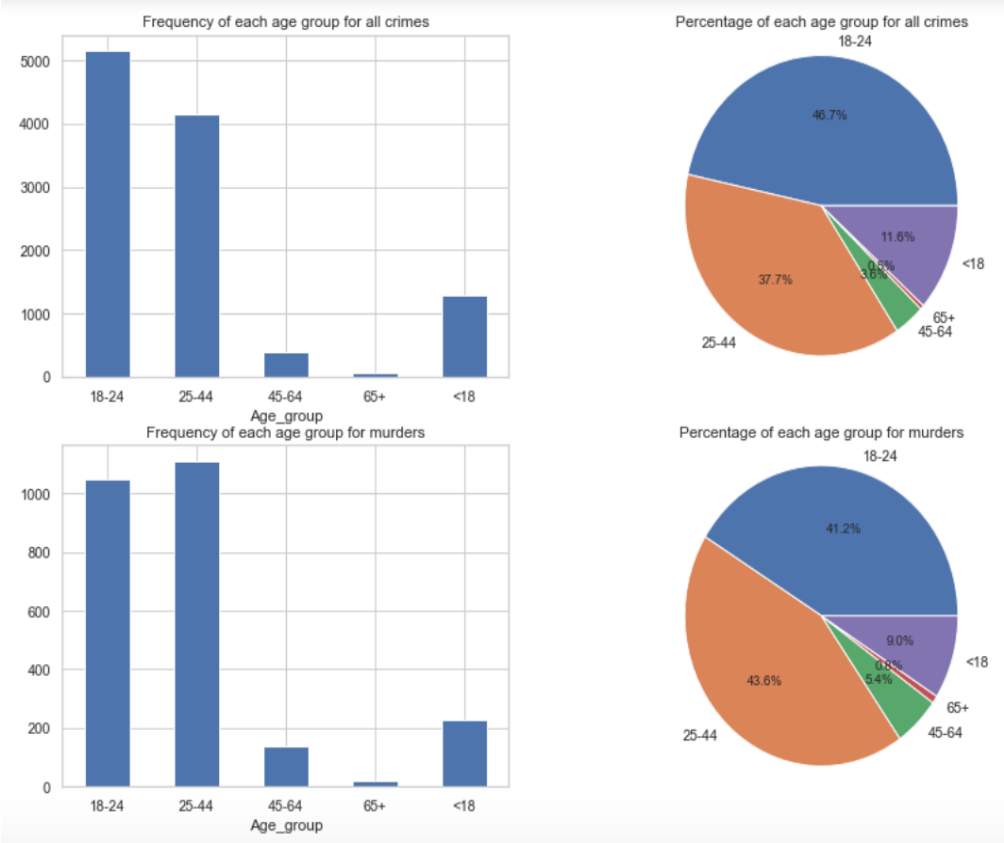

Figure 7. Distribution of Age groups

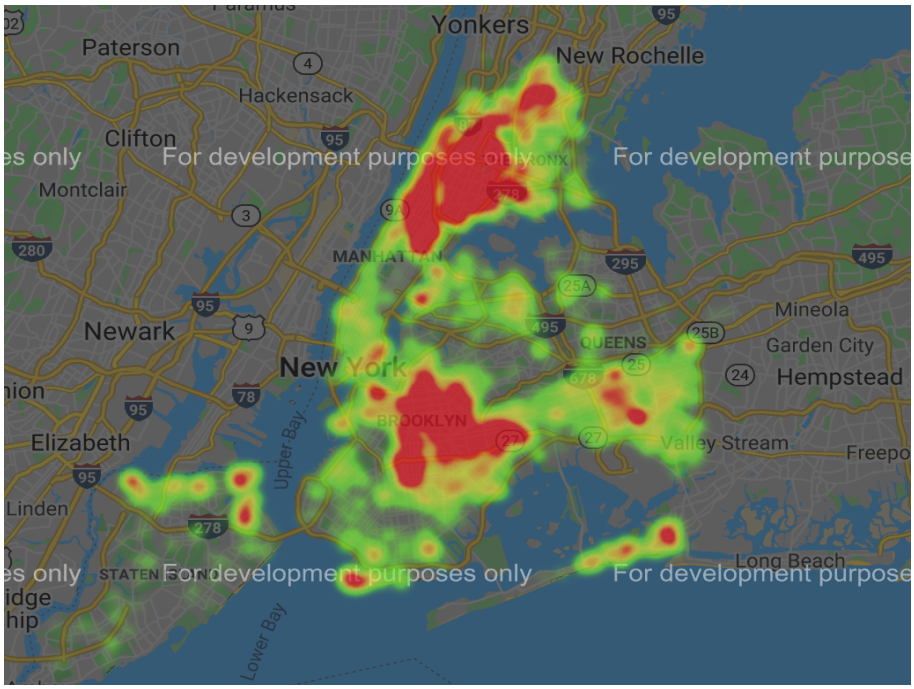

Figure 8. Heatmap of crimes

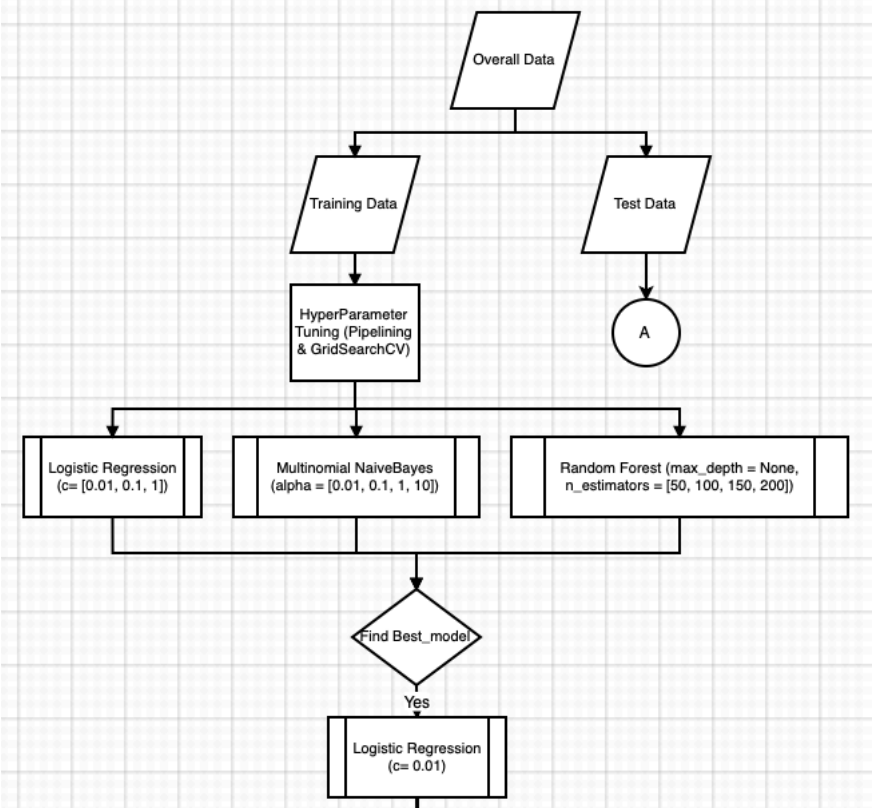

Figure 9. Flowchart - Part A

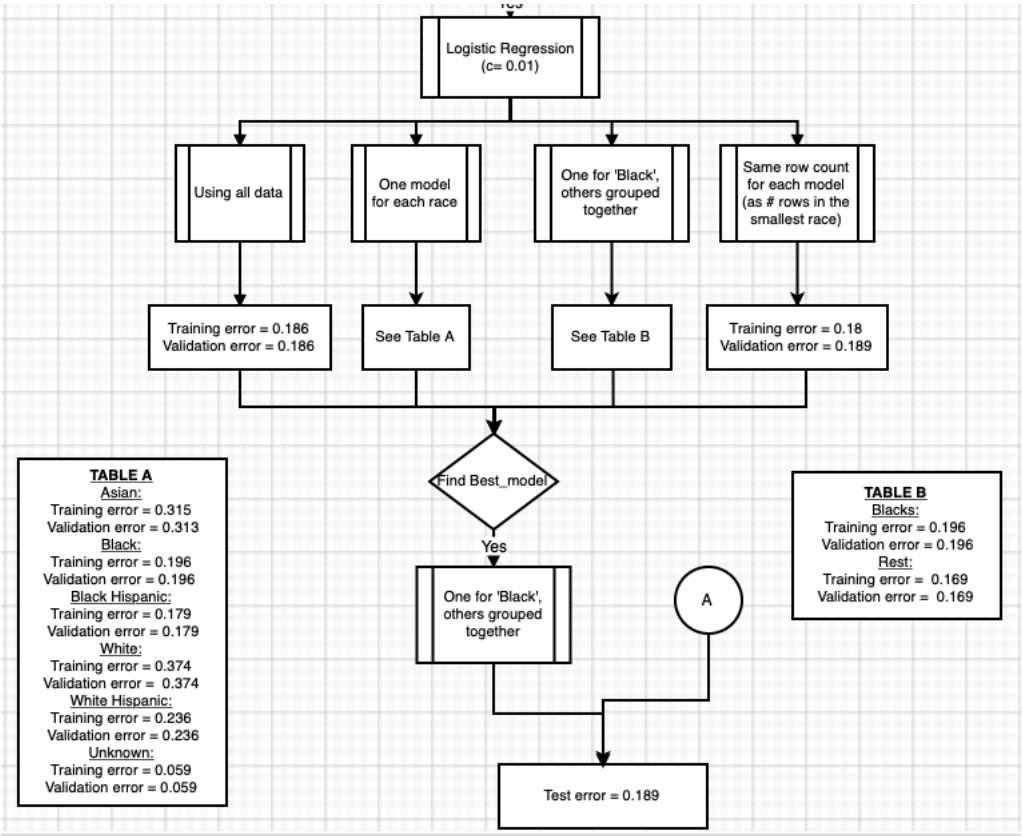

Figure 10. Flowchart - Part B

	names	coefs
6	25-44	0.331989
7	45-64	0.143514
8	65+	0.024570
5	18-24	0.011831
19	WHITE HISPANIC	0.000000
18	WHITE	0.000000
17	Unknown_Race	0.000000
16	BLACK HISPANIC	0.000000
14	ASIAN / PACIFIC ISLANDER	0.000000
0	BRONX	-0.000957
13	Unknown_Sex	-0.028874
11	F	-0.056970
4	STATEN ISLAND	-0.074903
2	MANHATTAN	-0.096966
9	<18	-0.102802
3	QUEENS	-0.109048
20	Afterwork	-0.120293
21	Sleeping hours	-0.132465
1	BROOKLYN	-0.141581
22	Working hours	-0.170698
12	M	-0.337612
15	BLACK	-0.423455
10	Unknown_Age	-0.832557

Figure 11. Feature importance – Black model (Descending order)

	names	coefs
7	45-64	0.162451
18	WHITE	0.139982
6	25-44	0.109384
14	ASIAN / PACIFIC ISLANDER	0.048800
8	65+	0.015837
15	BLACK	0.000000
4	STATEN ISLAND	-0.002916
22	Working hours	-0.025218
19	WHITE HISPANIC	-0.032346
11	F	-0.040076
9	<18	-0.060362
3	QUEENS	-0.067366
5	18-24	-0.094071
2	MANHATTAN	-0.121417
1	BROOKLYN	-0.152159
0	BRONX	-0.195111
16	BLACK HISPANIC	-0.200496
12	M	-0.216711
20	Afterwork	-0.236383
21	Sleeping hours	-0.277369
13	Unknown_Sex	-0.282182
17	Unknown_Race	-0.494910
10	Unknown_Age	-0.672209

Figure 12. Feature importance - Rest (non-black) model (Descending order)