Table of Contents

Executive summary

This project focus on studying New York City crime data to assist local police department to understand crime patterns and better deploy resources to each area accordingly. The project is a real world imitated case study. Our team act as consultants to provide constructive advices to our client, the New York City Police Department with assistant from the Deloitte mentors.

Background

The New York City Police Department (NYPD) seeks for assistant from analytics company to obtain both data science and visualization advisory to identify high risk neighborhood so as additional police forces would be allocated to ensure public safely. Additional, NYPD looks for insight analysis on social pulse of crime for mitigate crime rate in long term. While investigating crime requires increased special service resources and officer hours to pursue justice coupled, increase community risk, and negative citizen sentiment, violent crimes, which cause severe social impact and law enforcement activities, should be considered primarily on crime analysis in this project.

Objectives

According to the NYPD requirements, there are three objectives should be achieved by the end of the project:

- Developing a crime risk model to increase the predictive capability of future crime
- Establishing some full visualization capabilities to dynamically manipulate classified data to guide the strategic deployment of precinct resources.
- Integrating demographic factors within the BI platform to provide constructive advices on resources deployment and insight finding of social pulse to police and city officials.

Process

The entire project was deployed in 5 stages:

- **Problem identification**: To fully understand the project requirements, our team discussed with Deloitte team to identify the problems NYPD is facing and determined the objectives client wants to achieve

- **Data collection**: Besides the NYPD complaint data provided by client, our team also obtain a demographic dataset of each zip code in NYC. The demographic factors included in the dataset are population, education rate, poverty rate, etc. The NYPD complaint data is used for classified class and conduct crime pattern analysis. The Demographic dataset would be primary used for analyzing social pulse of high risk neighborhood.
- **Data preparation**: Two datasets were cleaned and reorganized to be more interpretable and fit to the models. The detail steps will be discussed in later chapter.
- **Models building**: The clean datasets were run on 6 models to select the one that results the best performance. The best models have prediction accuracy about 85% for both day and night period when the false positive rate is also under controlling.
- **BI platform development**: The model output is visualized by Tableau. This platform could be used for NYPD to visualize high risk neighborhoods directly and identify relevant demographic and historic crime related to each neighborhood that may have crime event happened at any given time.

Finding

Based on the results from our model:

- Significant violent crime predictors: violent crime density, all kinds of crime density, number of crime occurred in past six months in adjacent areas and number of months since last crime event
- Demographic factors of high risk neighborhood: high population, high poverty rate, high renter-occupied rate, low education rate, low median age, low median income
- High risk neighborhoods: downtown and midtown of Manhattan, Flushing in Queens, central of Bronx

Advice

- NYPD should allocate more police resources to high risk neighborhoods and pay extra attention to areas that have contain most of the common demographic factors we have identified

- City leadership should allocate extra education funding to district that has low education rate. Improving education rate will increase median income therefore decrease poverty rate, which will reduce crime rate effectively
- For low population density neighborhood, they should construct certain percentage of subsidized housing to disperse high population density neighborhoods. Besides, such policy also helps to solve social issues such as high poverty rate and low education rate

Conclusion

For our project, we analyzed NYC crime data and built models to identify high-risk neighborhoods and corresponding demographic factors. Due to time and data limit, we only focus on violent crime, which usually causes most severe social impact and demand more public resource. For modeling building, we ran data on several machine learning models based on day and night period and selected the one that has the best performance based on prediction accuracy and false positive rate. Besides, we implemented our results on a BI platform to display significant demographic factors for each zip code in NYC. This BI model also allows users to import extra data to make prediction in the future.

Based on our analysis, we summarized the high-risk neighborhoods, common demographic factors for these neighborhoods, significant variables for violent crime prediction in NYC and suggestions on how to improve public safety. Our analysis will be helpful for NYPD on predicting future crime and deploying police resource accordingly. Besides, NYC city leadership could pay attention to current high-risk neighborhood demographic factors and try to reduce violent crime rate through solving social issues in long term.

Project Overflow

Data Exploration

- Platform: R, Tableau
- Required package: data.table, dplyr, lubridate, ggplot2, ggmap
- Related files: Exploration.R, Descriptive.twb
- Data: NYPD_Complaint_Data_Historic.csv, summary.csv

Predictor Generation

- Platform: R
- Required package: sqldf, data.table, dplyr, sp, lubridate, RJSONIO, Rcurl
- Related files: predictor_generation.R, preparation_function.R
- Data: NYPD_Complaint_Data_Historic.csv, district.csv, demographic_cal.csv

Data Preparation

- Platform: R
- Required package: dplyr, moments, caret, ROSE
- Related files: Preparation (violent day).Rmd, Preparation (violent night).Rmd
- Data: VIOLENT CRIME_day.csv, VIOLENT CRIME_night.csv

Model Building

- Platform: Python
- Required package: h2o, numpy, pandas
- Related files: Day_model.ipynb, Night_model.ipynb
- Data: violent_day_test.csv, violent_day_train.csv, violent_night_test.csv, violent_night_traincsv

BI Platform

- Platform: Tableau
- Related files: BI platform.twb
- Data: violent_visual.csv

1. Purpose statement & Introduction

1.1. Introduction

New York attracts millions of visitor each year, which brings huge pressure for NYPD and city leadership to keep the city safe due to dense and diversified population. Hence, the capability of predicting crime and identify high risk neighborhoods are important for NYPD. For this project, our purposes is to analyze crime pattern from the past NYC crime data and develop a model to predict crime rate in the future. Besides, we will establishing full visualization capabilities to dynamically manipulate classified data to guide the strategic deployment of precinct resources. While investigating crime requires increased special service resources and officer hours to pursue justice coupled, increase community risk, and negative citizen sentiment, we decided to focus on analyzing violent crimes, which cause severe social impact and law enforcement activities.

1.2. Project Workflow

The entire project was deployed in 6 stages:

- **Problem identification**: To fully understand the project requirements, we discussed with Deloitte team to identify the problems NYPD is facing and determined the objectives client wants to achieve.
- **Data collection & exploration**: Besides the NYPD complaint data provided by client, we also obtain a demographic dataset of each zip code in NYC. The demographic factors included in the dataset are population, education rate, poverty rate, etc. The NYPD complaint data is used for classified class and conduct crime pattern analysis. The Demographic dataset would be primary used for analyzing social pulse of high risk neighborhood.
- **Data preparation**: Two datasets were cleaned and reorganized to be more interpretable and fit to the models. The detail steps will be discussed in later chapter.
- **Models building**: The clean datasets were run on 6 models to select the one that results the best performance based on accuracy and False Positive Rate.
- **BI platform development**: The model output is visualized by Tableau.

- **Advice**: We provide advice on improving public safety to NYPC and city leadership according to the model results.

1.3. Data Sources

There are two data sources used for this project:

- NYPD Complaint Historic Data: data.cityofnewyork.us
- Census American Community Survey (2010): factfinder.census.gov

NYPD complaint crime data is a government-published data set. It stores information of all types of crime occurred from 2006 to 2016. It is used for conducting spatial analysis and identifying historical crime trend and patterns. This data set contains detailed information including exact time, actual location and crime description for each crime event. For our project, we will mainly focus on violent crime since it usually causes severe social effect and therefore require more police resources. After careful inspection, 5 original crime descriptions are defined as violent crime - offense against public safety, arson, dangerous drugs, dangerous weapon and intoxicated & impaired driving.

Demographic data from the U.S. 2010 Census American Community Survey are included as part of the violent crime analysis in order to determine some external factors that may affect crime and form a relationship between crime and social factors. Demographic data includes data related to population, age, income, education, renter-occupied rate, etc. for each zip code in NYC.

2. Data Exploration

We used R and Tableau to have an overview for the violent crime distribution in term of location and time.

2.1. Spatial analysis

Heat map is often used in crime density charting. The logic behind heat map is that for each area that is showed to have high density, there are a great deal of crimes happened both in this area and its adjacent area. The heat map of violent crime density is shown in Figure 2.1. In the heat map, darker color represents higher violent crime density. Approximately the entire Bronx, downtown and midtown of Manhattan, and central of Brooklyn belong to high crime density districts.

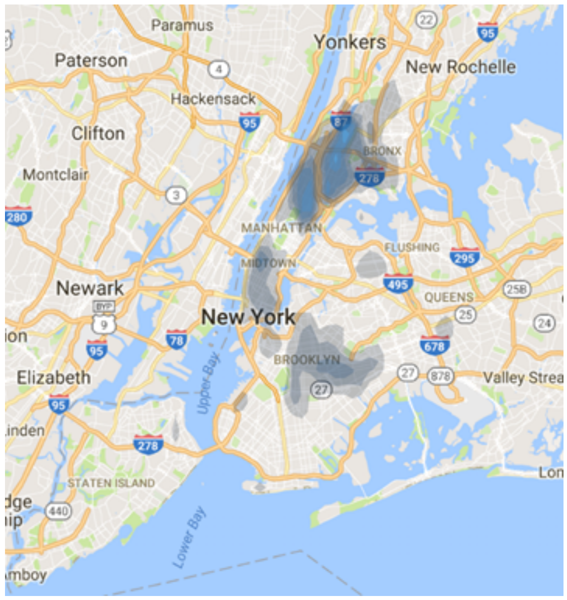


Figure 2.1: Heat Map of Violent Crime Density

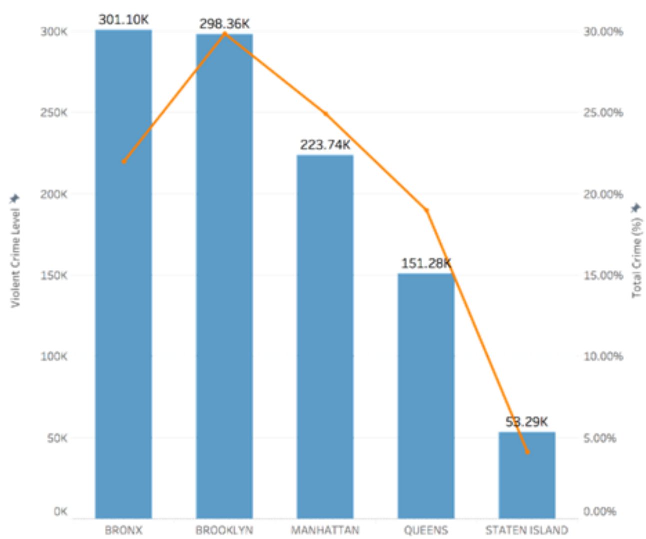


Figure 2.2: Violent Crime Frequency by District

In order to identify the unique pattern of violent crime. Violent crime level is summarized into five districts and is compared to the district distribution of all kinds of crime. The result is shown in Figure 2.2. The blue bars represent the total number of violent crime in each district. The orange line represents the crime percentage contributes to the all kinds of crime in NYC by each district. In all kind of crime, Brooklyn has the highest crime percentage while the percentage of

Bronx is much lower than Brooklyn. However, the number of violent crime for these two district is approximately the same. We can conclude that Brooklyn is a high-crime-density area for all kinds of crimes. But for Bronx, larger portion of crime happened here is violent crime compares to other districts. Since violent crime often causes more severe social impact, this indicates that NYPD should pay more attention or deploy more police resource to Bronx.

2.2. Time distribution

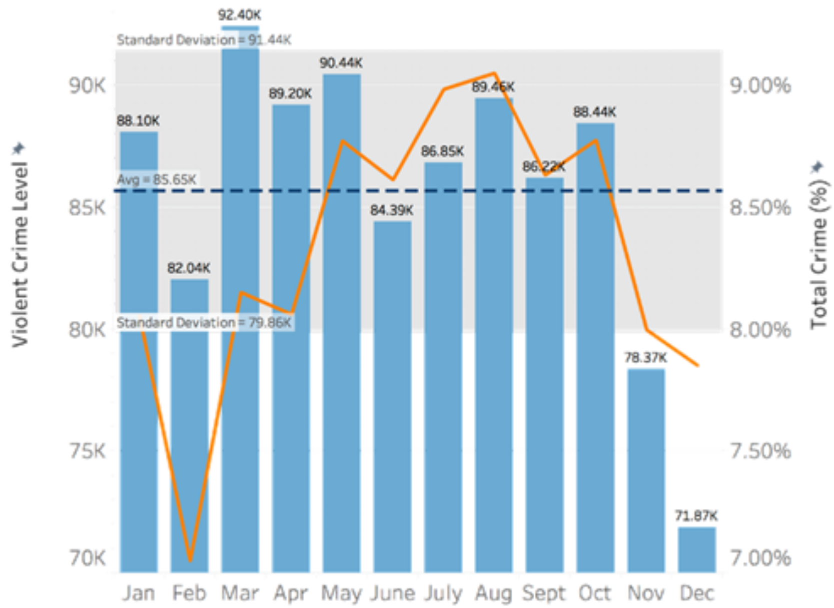


Figure 2.3: Monthly Distribution

Figure 2.3 shows the monthly distribution of both violent crime and total crime. According to the graph, number of violent crime for December is much lower than that of the rest of the months, which contradict to the common believe that there are more crime happen during holiday season. Additionally, the graph also shows the obvious difference in the distribution of violent crime and all kinds of crime. That is. The number of violent crime is much lower than the crime percentage line in November and December while much higher than the percentage line from January to May, which leads to a conclusion that compared to other kinds of crime, violent crime is more likely to happen from January to May but less likely to happen in November and December.

Similarly, hourly distribution is shown in Figure 2.4. Violent crime has the similar trend with all kinds of crime. The crime level reaches the minimum level in the early morning. It continues to increase from morning and starts to decrease during night time. However, we can also find that the distribution of violent crime differs from that of all kinds of crime. It can be seen that for all kinds of crime, the crime occurrence peaks in afternoon and starts to decrease in evening. However, the crime level continues to increase from noon to night and the number keeps at a high level until midnight. It can also be concluded that violent crime has higher variance in hourly distribution because the violent crime level exceeds the percentage line in several hours.

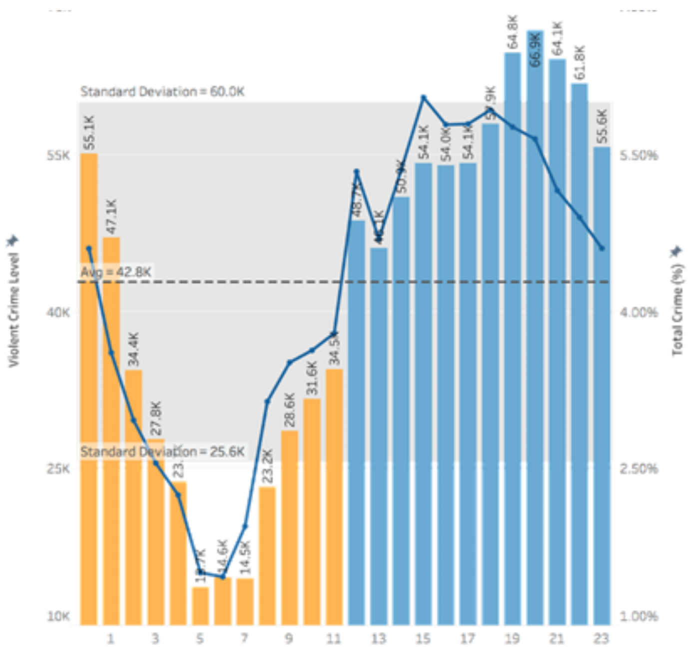
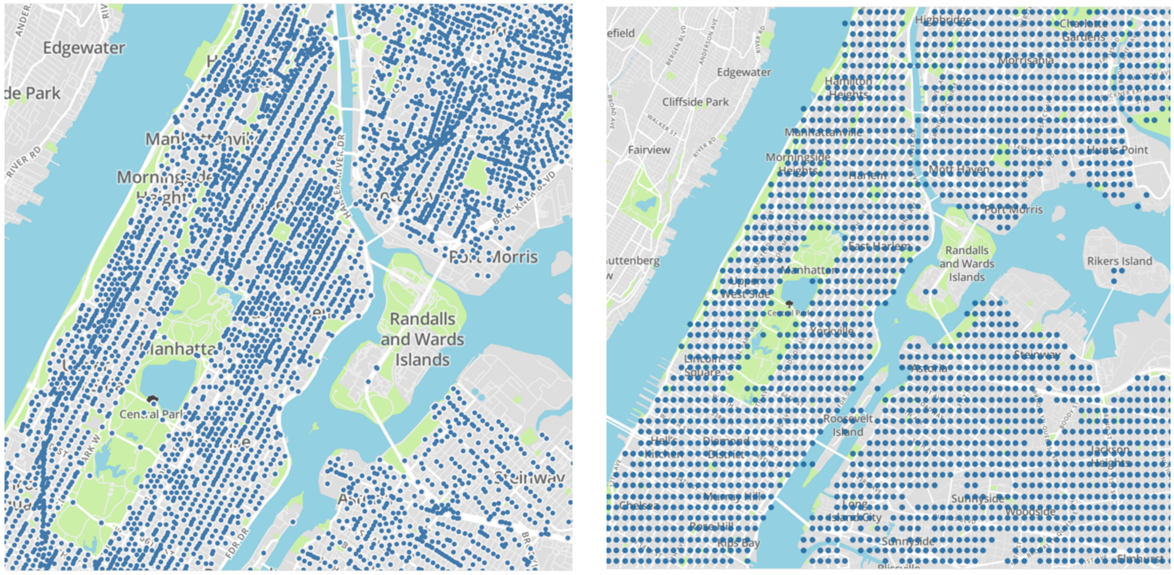


Figure 2.4: Hourly Distribution for Violent Crime Versus All Crime

3. Data Preparation

3.1. Grid Transformation

It's widely used in crime related research to develop predictive model based on square grid. For prediction purpose, The actual location in original data set is aggregated into 200m*200m square grid. The transformation is shown in Figure 3.1. The number of crime is summarized into each square grid and the prediction followed will be based on each single square grid.



(a) Before transformation **(b) After transformation**

Figure 3.1: Grid Transformation

3.2. Predictors Generation

3.2.1. Relevant calculation

Given the spatial crime data set of violent crime event location (x_i,y_i), a common method of calculating crime density is to use the Gaussian kernal density function

$$g(x,y) = \sum_{x_i,y_i} \frac{1}{2\pi\sigma^2} \times exp(-\frac{(x_i - x)^2 + (y_i - y)^2}{2\sigma^2}) \times \sum_t N_{x_i,y_i,t}$$

Density is calculated for each single square grid. In this formula, the density of a specific square grid is decided by both the number of crime happened within the grid and the crime happened in the adjacent grid. The weight of adjacent grid is decided by the distance to central grid. Specifically, theta is set to be 0.6. To avoid leakage, crime density calculation only includes crime event happened in 2006 to 2014, which consists of the training set. The density is calculated in two levels – total crime and violent crime.

Crime history within six months is also calculated. For a specific time *t*, Historic crime happened in the same grid the sum of number of crime happened within recent six months with weight considered to be exponentially based on time

$$p(x,y,t) = \sum_{i=1}^{6} exp(-i)\times N_{x,y,t-i}$$

For historic crime happened in the adjacent grid, the level is both time and distance weighted

$$a(x,y,t) = \sum_{x_i,y_i} \frac{1}{2\pi\sigma^2}\times exp(-\frac{(x_i - x)^2 + (y_i - y)^2}{2\sigma^2})\times p(x_i, y_i, t)$$

3.2.2. Demographic Variables

Since the demographic information is in zip code level but the predicting unit is grid. So we need to find the zip code of each grid. We used Google API in R to find the zip code of central location for each grid. Then the demographic information is assigned to each single grid. For information in percentage or numbers that cannot be divided, like foreign-born rate and median income, the same number is assigned to grids within each zip code. For information in number that can be divided, like population, the number is averagely distributed to each grid.

3.3. List of Predictors

Category	Variable	Description
Crime History	Total_density	Density of all kinds of crime
	Crime_density	Density of violent crime
	Last_occurred	Time since last crime event (in months)
	Previous_occurred	Weighted number of crime happened in past six month in the same grid
	Adjacent_previous_occurred	Weighted number of crime happened in past six month in adjacent grids
	Last_month	Number of crime happened in last month

	Population	Number of population
	Education rate (%)	Percentage of high school graduate or higher
	Median income	Median household income
	Foreign born rate (%)	Foreign born rate
	Poverty rate (%)	Percentage of individuals below poverty level
	15-19 (%)	Percentage of residents who are 15-19 years old
	20-24 (%)	Percentage of residents who are 20-24 years old
Demographic Variable	25-29 (%)	Percentage of residents who are 25-29 years old
	30-34 (%)	Percentage of residents who are 30-34 years old
	Median age	Median age of residents
	White (%)	Percentage of residents who are White (alone or in combination)
	Black or African American (%)	Percentage of residents who are Black or African American (alone or in combination)
	Asian (%)	Percentage of residents who are Asian (alone or in combination)
	Hispanic or Latino (%)	Percentage of residents who are Hispanic or Latino (alone or in combination)
Social economy	Housing	The number of total housing units

	Householder	Number of residents who own a house
	Living-alone householder (%)	Percentage of householder who live alone
Others	Manhattan, Bronx, Brooklyn, Queens	District dummy variables (with Staten Island as reference level)
	Jan, Feb Mar…	Monthly dummy variables (with December as reference level)

Table 3.1 List of predictors

3.4.Data Preparation on Independent Variables

After all independent variables are generated, data preparation proceeded in the following five steps.

1) Check variable skewness

We firstly checked the skewness of all independent variables to determine whether the variable is normally distributed.

2) Transformation

For those highly skewed variables, log function is used to transform them

$$X' = \log(X + C)$$

For nonzero variables, C would be assigned zero to do log transformation. When the variable contains some zero values, C is considered to be a very small but positive number. Specifically, referred to some research, C should be approximately half of smallest number that is not zero. Therefore, for variables that are continuous, like density, C is set to be 0.001. For variables that are discrete, like last_occured, C is set to be 0.5.

3) Standardization

All independent variables were standardized to enforce same scales on a set of variables in case variables with large values would incorrectly dominate the training process.

$$X' = (X - \min(X))/(\max(X) - \min(X))$$

4) Dataset partition

The whole dataset was partitioned into two parts: train set and test set. In crime prediction, it's better to split the data set based on time so that we can easily test the prediction result for crimes happened more recently. The original data set contains historic crime data from 2006 to 2016. Train set is from June 2006 to December 2014 and is used to train a model. Test set is from January 2015 to December 2016 and is used to measure how well the model performs at making predictions in new data.

5) Under sampling on train set

We also did under sampling to our train set to make sure the dependent variable is balanced.

4. Predictive Modeling with H2O

Prediction models are built with H2O package in Python.

4.1. Model training

Six supervised models are trained independently

- Logistic regression
- Random forest
- Neural network
- Stacked ensemble (Logistic regression + Random forest)
- Gradient boosting decision tree
- Gradient boosting decision tree ensemble

Each training process also includes four cross validation process to ensure the coefficient is accurately estimated. In neural network and gradient boosting decision tree, hyper parameter search is used to determine the best parameter for each model.

4.2. Threshold selection

In violent crime prediction, false positive rate should be carefully minimized because a false positive prediction may lead to a waste of resource or may cause public panic. The models output the likelihood of a crime event happened for each grid. The threshold is determined to be the maximum probability that ensure the false positive rate in validation set does not exceed 10%.

4.3. Model Performance Comparison

When selecting the best model, high overall prediction accuracy is consider to be the first criteria. In addition, the false positive rate and false negative rate should not be so high.

Model	Accuracy	Sensitivity	Specificity	FPR	FNR
Logistic	84.03%	55.20%	88.69%	11.31%	44.80%
Random Forest	84.20%	51.81%	89.32%	10.68%	48.19%
Neural Network	83.26%	46.63%	89.18%	10.82%	53.37%
Stacked Ensemble	78.98%	68.94%	80.60%	19.40%	31.06%
GBM	84.38%	55.06%	89.12%	10.88%	44.94%
GBM Ensemble	83.43%	58.62%	87.44%	12.56%	41.38%

Table 4.1: Model Comparison - Day

Table 4.1 shows the model performance of violent crime happened in day time in test set. The prediction accuracy of random forest ranks the second and is 0.18% lower than the accuracy of GBM. The false positive rate of random forest is 0.2% higher than that of GBM. Therefore, random forest is the best model for predicting violent crime happened in day time.

Table 4.2 shows the model performance of violent crime happened in night time in test set.it can be seen that Random Forest model also does the best job among the six models for night period since it has the highest prediction accuracy (85.25%) and the lowest false positive rate (10.14%) in the meantime.

Model	Accuracy	Sensitivity	Specificity	FPR	FNR
Logistic	84.98%	54.01%	88.95%	11.05%	45.99%
Random Forest	85.25%	49.25%	89.86%	10.14%	50.75%
Neural Network	83.46%	48.72%	88.27%	11.83%	51.28%
Stacked Ensemble	81.78%	62.92%	84.20%	15.80%	37.08%
GBM	85.21%	54.12%	89.20%	10.80%	45.88%
GBM Ensemble	84.18%	57.49%	87.60%	12.40%	42.51%

Table 4.2: Model Comparison - Night

4.4. Confusion matrix for the best model

The confusion matrix of two best model performance is shown in table 4.3 and 4.4

		Actual	
		False	**True**
Predict	**False**	304758	26578
	True	36446	28578

Table 4.4: Confusion matrix of the best models for day

		Actual	
		False	**True**
Predict	**False**	315702	22857
	True	35616	22185

Table 4.4: Confusion matrix of the best models for night

From the confusion matrix, we see that most of predictions are accurate. Most grids that do not have crime happened are predicted accurately. For grids that have crime event happened, the model can predict about half of them right.

4.5. Variable importance

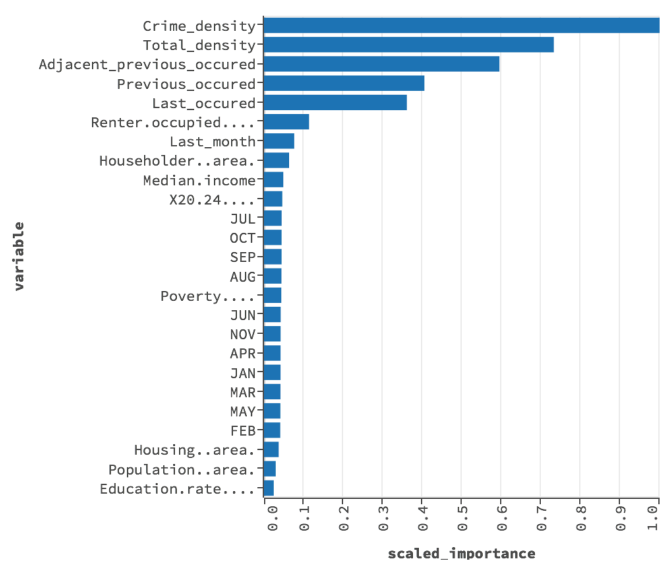


Figure 4.1: Variable importance of best model for day

The variable importance of the best model for day is shown in Figure 4.1. Predictors which are most useful for predicting the violent event during day are crime density, total density, number of crime occurred in past six months in adjacent grids, last occurred? and number of months since last crime event. All of those most important variables in our day model are crime history related and it seems reasonable to us. Similarly, the variable importance of the best model for night is shown in Figure 4.2. Predictors which are most useful for predicting the violent event during night are the same as those in night model.

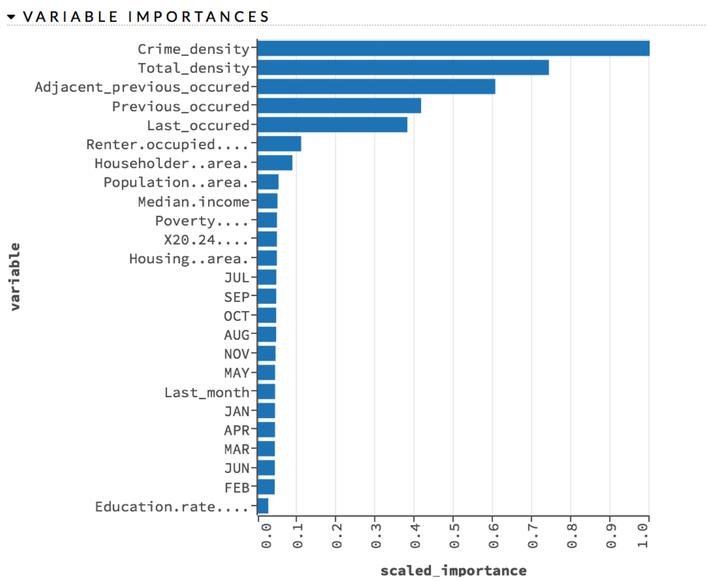


Figure 4.2: Variable importance of best model for night

5. Model Application

5.1. BI Platform

BI platform is built for NYPD to get the relevant information of areas that may have crime event happened potentially. Figure 5.1 shows the platform built using Tableau to visualize the prediction for year 2015 and 2016. The filter in the top can be used to select the period, year and month to see the prediction of violent crime distribution in NYC. The dots on the map represents violent crimes predicted to happen in the prediction period, year 2015 and 2016 and the colors of the spots represent district information. NYPD can get the detailed information of each single grid by clicking the spot.

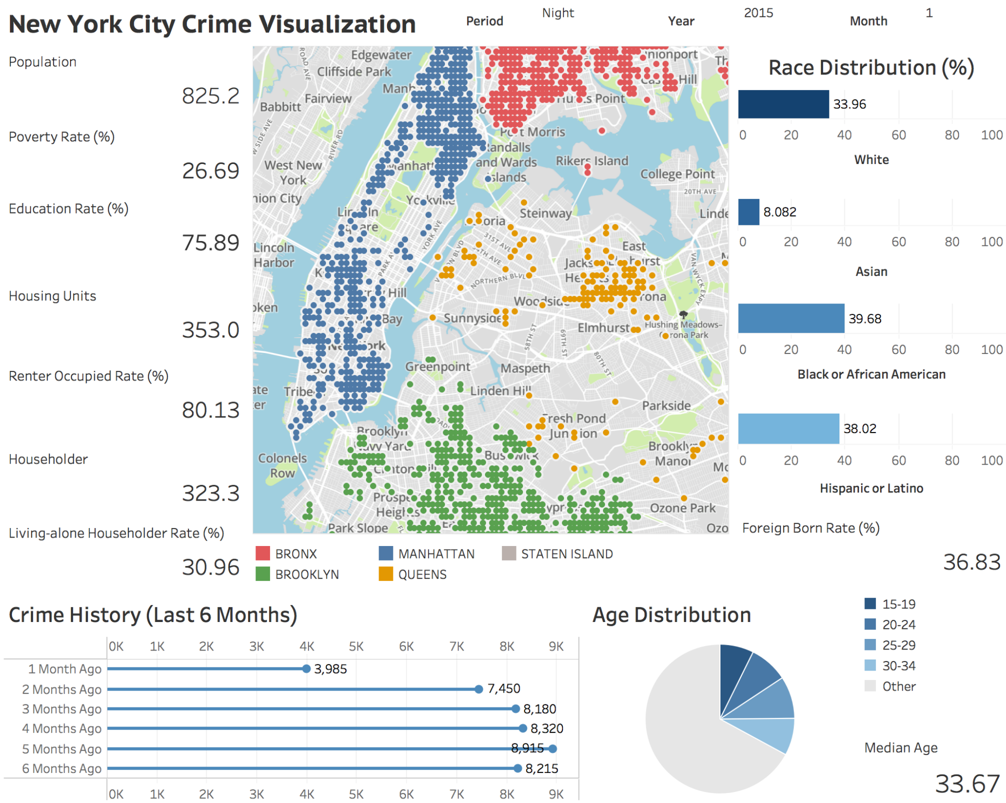


Figure 5.1: BI Platform Visualization

Figure 5.2 shows an example of clicking a spot in Manhattan. The basic demographic information in this clicked spot is shown on the top left, including population, median income, education rate, etc. Race information is shown on the top right, from which it can be seen that over 76% of people living here are White. The historic crime information in past six months is shown in the bottom left. There're several violent crime happened recently but the number is relatively low. The age group distribution is shown in the bottom right, which indicates that residents here are pretty young. This BI platform enables NYPD to form a better understanding of each neighborhood. It's also published in Tableau Public and can be seen from https://public.tableau.com/profile/jingbo.zhang#!/vizhome/NYCCrimeBIPlatform/Dashboard1?publish=yes.

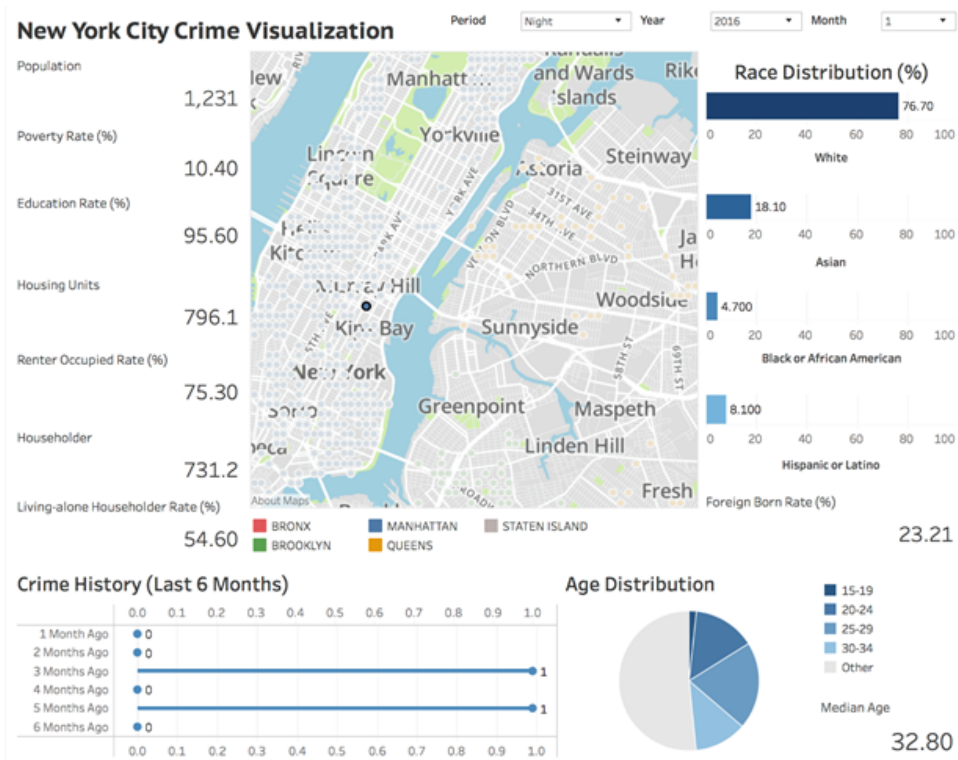


Figure 5.2: BI Platform Visualization for a specific spot

5.2. Potential Social Pulses

High-risk neighborhoods are defined for those grids predicted to have crime event in over 80% of months in our test period, year 2015 and 2016. By comparing the distribution of demographic information, some potential social pulses are summarized from the prediction model. It shows that 68% of high-risk neighborhoods have poverty rate over 25% but the percentage for all neighborhoods is only 19%. The same pattern applies to renter-occupied rate. From the prediction result, high renter-occupied rate is another common demographic information for high-risk neighborhoods. High-risk neighborhoods also have low education rate and low median age compared to all neighborhoods.

	poverty rate > 25%	education rate > 80%	Renter-occupied rate > 75%	median age > 35
High-risk neighborhoods	68%	29%	80%	25%
All neighborhoods	19%	66%	25%	64%

Table 5.1: Potential Social Pulses

5.3. Model application in the future

The predictive model in this project and the BI platform can also be used by NYPD for crime analysis in the future. If some new crime information is available in the future, those data can be imported into predictive model to predict areas where crime may happen and help identify high-risk neighborhoods.

In the next step, NYPD can use the BI platform to visualize the prediction distribution for violent crime and identify the demographic factors for high-risk neighborhood predicted. This enables them to better deploy police source and help to reduce violent crime. NYPD should deploy additional resources to high-risk neighborhoods identified by the model, especially at peak accident times.

For city leadership, they could allocate more education fund to high-risk neighborhoods. Also, for areas with low population density, they could construct subsidized housing. By doing these, they could reduce income gap and education gap for high-risk neighborhoods step by step.

6. Conclusion

For our project, we analyzed NYC crime data and built models to identify high-risk neighborhoods and corresponding demographic factors. Due to time and data limit, we only focus on violent crime, which usually causes most severe social impact and demand more public resource. For modeling building, we ran data on several machine learning models based on day and night period and selected the one that has the best performance based on prediction accuracy and false positive rate. Besides, we implemented our results on a BI platform to display significant demographic factors for each zip code in NYC. This BI model also allows users to import extra data to make prediction in the future.

Based on our analysis, we summarized the high-risk neighborhoods, common demographic factors for these neighborhoods, significant variables for violent crime prediction in NYC and suggestions on how to improve public safety. Our analysis will be helpful for NYPD on predicting future crime and deploying police resource accordingly. Besides, NYC city leadership could pay attention to current high-risk neighborhood demographic factors and try to reduce violent crime rate through solving social issues in long term.

References

Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, 61, 115-125. doi:10.1016/j.dss.2014.02.003

Clarke, R. V. G., and John Eck. Become a Problem-Solving Crime Analyst: in 55 Small Steps. *UCL Jill Dando Institute of Crime Science*, 2003.

Eck, John E., et al. "Finding Crime Hot Spots through Repeat Address Mapping." *Analyzing Crime Patterns: Frontiers of Practice*, pp. 49–64., doi:10.4135/9781452220369.n5.

Gerber, Matthew S. "Predicting Crime Using Twitter and Kernel Density Estimation." *Decision Support Systems*, vol. 61, 2014, pp. 115–125., doi:10.1016/j.dss.2014.02.003.

Kahle D. and Wickham. H. Ggmap: Spatial Visualization with ggplot2. *CONTRIBUTED RESEARCH ARTICLES*, The R Journal Vol. 5/1, June 2013.

Mohler, George. Marked Point Process Hotspot Maps for Homicide and Gun Crime Prediction in Chicago. *International Journal of Forecasting*, vol. 30, no. 3, 2014, pp. 491–497., doi:10.1016/j.ijforecast.2014.01.004.

Rummens, Anneleen, et al. "The Use of Predictive Analysis in Spatiotemporal Crime Forecasting: Building and Testing a Model in an Urban Context." *Applied Geography*, vol. 86, 2017, pp. 255–261., doi:10.1016/j.apgeog.2017.06.011.

Xue, Yifei, and Donald E. Brown. "Spatial Analysis with Preference Specification of Latent Decision Makers for Criminal Event Prediction." *Decision Support Systems*, vol. 41, no. 3, 2006, pp. 560–573., doi:10.1016/j.dss.2004.06.007.