Analysis and Prediction of NYC Crime Data

Haitong Lin (hl929), Yaxuan Huang (yh798), Pingyue Pan (pp452)

1. Abstract

The high frequency of crime is a prominent reality of metropolitan life. In the post-pandemic New York City, new challenges in urban management, public health, and financial downturn have led to higher crime rates that increase the public's concern. This paper aims to predict patterns of crime type in New York City by building various machine learning models, including Random Forest, AdaBoost, and Logistic Regression. Our best-performing model is Random Forest with an accuracy of 0.49.

2. Motivation

The disparate, seemingly random degrees of safety in and around the city have become a growing source of concern as crime rates in New York City have increased since the COVID-19 outbreak. As current New York City residents, we are interested in the past, present, and future of the city's crime statistics because they directly affect our daily lives.

For this project, we are motivated by the public's concern and our own desire to identify the factors that lead to the occurrence of a specific level of crime (Misdemeanor, Violation, and Felony). To elaborate on our process, we first analyzed the previous crime data in the city to explore the patterns indicated by the data. Then, we delved into investigating the factors associated with crime patterns in NYC and developed an algorithm to predict the instance of a particular type of crime. We hope that our project could provide insights into crime forecasting and smart policing.

3. Context

While machine learning has been commonly applied as a tool for predictive policing, early efforts in crime prediction have been controversial. The area of concern has been the structural biases in police enforcement and its complex association with crime in society. One of the most recent state-of-art research that takes into account such structural biases was done by data and social scientists from the University of Chicago. Focusing on the city of Chicago, the research team developed a crime forecasting algorithm by learning patterns in spatiotemporal features from public data on violent and property crimes, successfully predicting future crimes one week in advance with approximately 90% accuracy (Rotaru et al. 2022).

However, in the literature review stage for this project, we did not come across many projects that focused on crime prediction in New York City. Most articles or Kaggle projects examined crime rates in NYC using Exploratory Data Analysis or statistical analysis from a social science perspective. On the other hand, projects that developed prediction models mostly took place before the pandemic and focused on predicting the occurrence of crime, rather than the type of crime. Namely, the paper titled "Prediction of Crime in Neighbourhoods of New York City using Spatial Data Analysis" (Almuhanna et al. 2021) used 2.2M criminal records from 2006 to 2019 to train a crime classification algorithm using Random Forest and XGboost that reached 0.503 and 0.52 prediction accuracy respectively. This particular project provides some of the highest accuracies we have observed in predicting crimes in New York City.

Since current literature presents a gap in examining crime type patterns in post-pandemic NYC, the temporal scope of our project is set from 2019 to 2022. The analysis and methods we used in this project were not groundbreaking, but they provide more relevant insights into predicting crimes in the present day and near future. Given the limited post-pandemic data and constraints on computing power, our project sets out to explore different prediction methods and aims to replicate the existing model's accuracy. In providing an updated, more pertinent prediction model for the post-pandemic NYC world, our project also aims to identify potential improvements in data collection.

4. Methods

4.1 Baseline Model Selection

Our project uses the Logistic Regression Model as a baseline model since we are building an Offense Level classifier. As covered in class, Logistic Regression is a supervised learning, classification algorithm that is used to assign instances to a discrete set of classes. It finds a linear decision boundary and relies on the probabilistic interpretation of parameterized Bernouilli distribution. The model is also a common baseline model for a classification problem with a reasonable performance.

4.2 Model Selection

In our model selection process, we compared common classification models by using a 10-fold cross-validation. After observing the accuracies, we found that Random Forest and AdaBoost had the highest accuracies.

The Random Forest model is a supervised machine learning algorithm used to solve both classification and regression problems. It builds multiple decision trees and merges them together to obtain a prediction of higher accuracy and stability. Notably, it is also known as an **ensemble learning algorithm** (also known as **meta-learning**) that seeks better prediction performance by combining predictions from various models. To highlight, Random Forest belongs to the model family of **Bagging decision trees**, meaning that it improves the algorithm by reducing its variance through averaging multiple models trained on random subsets of the data. In doing so, Random Forest relies on an optimizer that functions on the greedy approach of rule addition, followed by pruning processes. As the Random Forest requires little data pre-processing while achieving relatively high prediction accuracy, it is one of the most popular machine learning algorithms in classification problems.

Similar to Random Forest, **AdaBoost** (Adaptive Boosting) is also an ensemble learning method that was initially created to increase the accuracy of binary classification algorithms. At a high level, AdaBoost utilizes an iterative approach to learn from the flaws of the weak classifiers, while pruning them into stronger, more efficient models. Whereas Random Forest uses the Bagging technique to improve the performance of its algorithms, AdaBoost uses **Boosting** to reduce the bias errors that arise when models are unable to identify relevant patterns in data. Unraveling the essence of AdaBoost, it targets the problem of a single classifier's inability to predict the class of an object. However, when a classifier is grouped with multiple weak classifiers, each classifier can progressively learn from one another in terms of wrongly classified objects, and the resulting model would be one with higher predictive accuracy.

5. Setup

5.1 Feature Selection

We conducted exploratory data analysis to figure out the dataset features we want to use in our model. We started by looking at crimes in different locations and time.

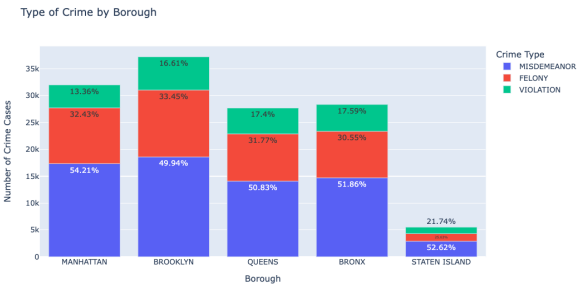

Figure 1. Type of crime by borough

Figure 2. Type of crime by hours

Regarding Crime Type by Borough, we observed that Brooklyn and Manhattan have both the highest numbers of crime cases and the highest percentages of Felony crimes. As for the type of crime by hours, although the number of crime cases during the midnight hours of 3-5 a.m. is lower compared to other hours, their percentage of Felony is the highest among all hours (40.4% - 41.57%). In contrast, during the afternoon hours of 14-17 p.m., the number of crime cases are the highest, but the percentage of felony is lower as compared to midnight hours.

We also explored Crime Types by Victim Age and Victim Sex. We observed that the Victim Age group of 25-44 years old has the highest proportion across all three types of crimes. As for Victim Sex, there is not a clear pattern.

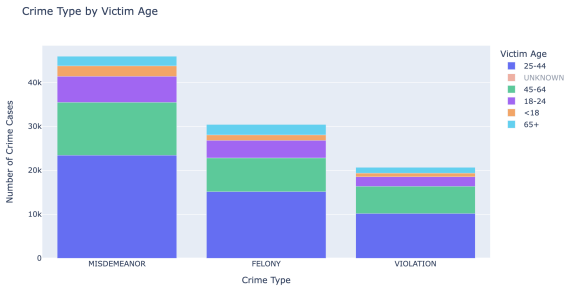

Figure 3. Crime time by victim age

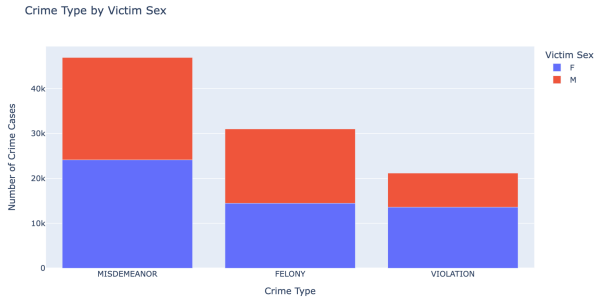

Figure 4. Crime by victim sex

5.2 Preliminary Experiment

From EDA, we figured out that the dataset features that have a correlation with crime can be divided into three categories: location, time, and the characteristics of the victim. For the first set of experiments, we decided to use the characteristics of the victim to predict the level of crime that occurred to them. We decide to use hour, borough, victim name, sex, and age group as our features for this trial, and we also decided that we would develop our methods based on classification algorithms because it's the best fit for the data types in our dataset. We selected Logistic Regression to be our baseline model, and also tested models including Linear Discriminant Analysis, knn, Naive Bayes, AdaBoost, and Neural Network. We calculate their accuracy using 10-fold cross-validation and drew a box plot to visualize and compare the accuracy. As the graph shows, Logistic Regression, LDA, Ada boost, and the Neural Network have the highest accuracies.

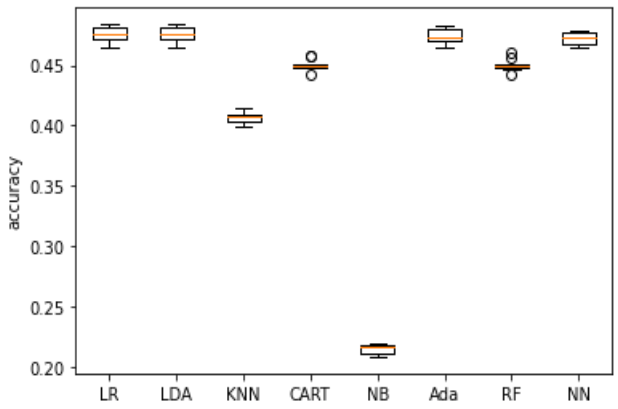

Figure 5. Model comparison

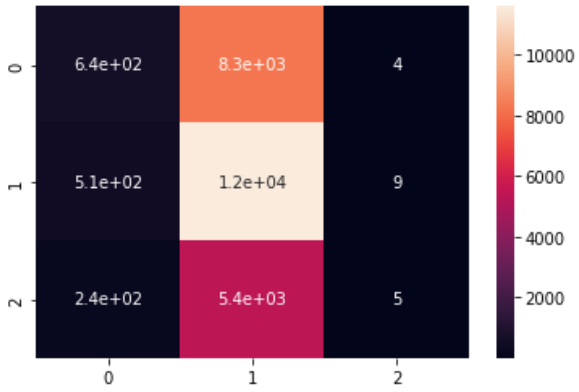

Figure 6. Baseline confusion matrix

Choosing Logistic Regression, we first used data points from 2019-2021 as the training dataset, and datapoint in 2022 as the test dataset, fitting them using Logistic Regression. It turned out that the accuracy in the training dataset is 0.4764, and in the test dataset is 0.4584. We also use the confusion matrix to visualize the classification., from which we could see that the data classified as class 2 is small, which might indicate an imbalanced dataset.

5.3 Modifications

Based on the limitations found in the preliminary experiment, we applied some modifications to our dataset. In the preliminary experiment, we selected a fixed proportion of data from each year, which guaranteed that we took into account the information at different times but failed to consider it as a predictive feature. Additionally, we observed that there might be an imbalance in the dataset because the number of different levels of crime is imbalanced by nature. Therefore, we researched for methods to solve these issues and modified our methods by sampling different datasets for our model. We first revised our feature selection process, and decided to add more features for our future experiments for a more comprehensive prediction. Then, the first additional dataset we added is to remove the time constraint and just randomly sampled 10% of the original dataset, which we think will help solve the first problem we mentioned earlier. For the second additional dataset, we sampled an equal amount of data for each category of crime, and the total amount of data is also about 10% of the original dataset.

We also improved our data preprocessing methods. We first check for the missing value in our datasets and remove the rows with null values for better results. Then, we added left joined with the holiday dataset to create the feature Holiday, which is a binary value to show us whether the date is a holiday or not. We also controlled the category to remove some outliers and used label encoder to turn text to numerical values. We believe that our modifications will help improve the performance of our future experiments.

With the modified datasets, we iterated our experiments and selected AdaBoost to be our best model. Details of our final model can be found in the results and discussion section below.

6. Results and Discussion

As discussed in the previous section, the two datasets we used for the final experiments are the 10% sampled dataset (referred to as dataset 1 below), and the dataset stratified with the level of crime (referred to as dataset 2 below). The datasets we used in our final experiments have the following features: Neighborhood, Hour, Year, Month, Holiday, Status, Borough, Victim_Sex, DoW, Victim_Agegroup,Victim_Race. We also did a visualization to display their importance.

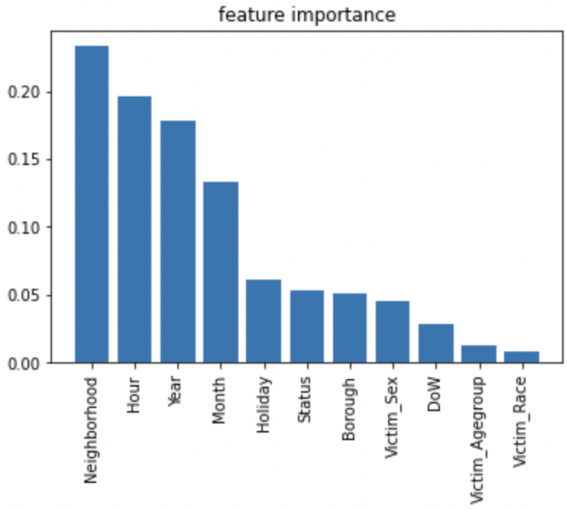

Figure 7. Feature importance

We first ran experiments with our baseline model, Logistic Regression. For dataset 1, we recorded a training accuracy of 51%, and a testing accuracy of 50.8%. For dataset 2, the training accuracy 40.8% was and the testing accuracy was 41.6%, which might indicate an underfit in our data. We also found out that the accuracies in dataset 1 are very similar to the results we got in the preliminary experiments, and the problem regarding imbalance remained. Using dataset 2 can improve this problem (which can be seen in the comparison of confusion matrices below), but would lead to a decrease in accuracy.

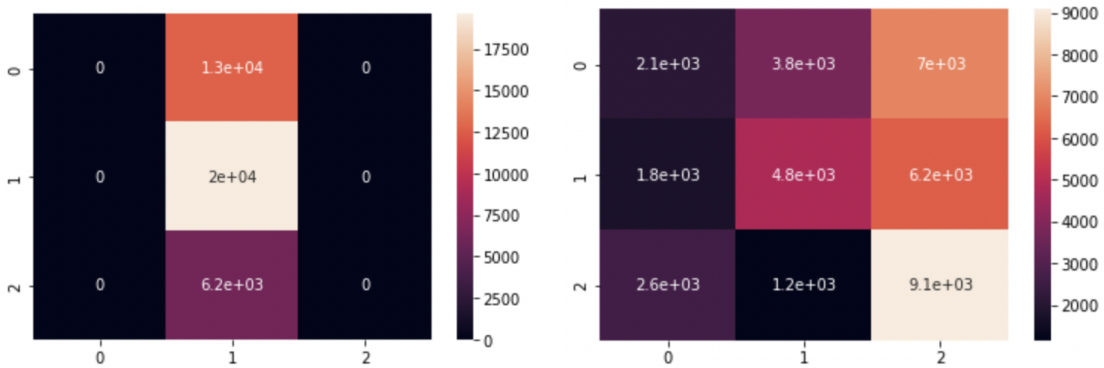

Figure 8. Confusion matrices for different datasets

We then focused on getting better classification models and working on improving their performance on dataset 2. We compared common classification models by using 10-fold cross-validation, and we found out that AdaBoost (46.99%) and Random Forest (44.68%) have the best performances.

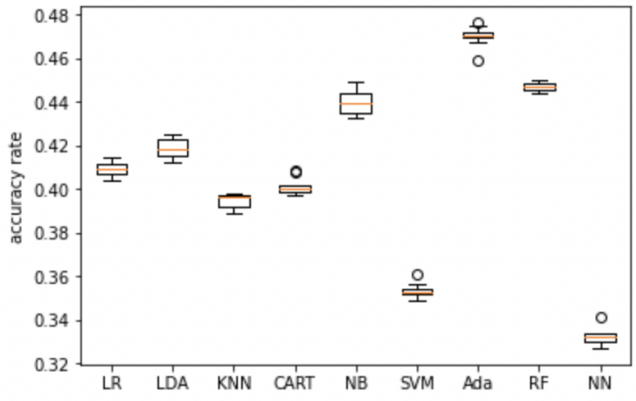

Figure 9. Model comparison

For Random Forest, we used GridSearch to tune our model, and the best performance we recorded included a max depth of 10 and 200 estimators. The training accuracy is 49.95% and the testing accuracy is 49.16%. However, this algorithm might possess an overfitting problem, since in some of our trials, the training accuracies are significantly higher than the testing accuracies. The classification report can be found below.

	precision	recall	f1-score	support
0	0.49	0.35	0.41	12835
1	0.64	0.27	0.38	12836
2	0.46	0.85	0.60	12883
accuracy			0.49	38554
macro avg	0.53	0.49	0.46	38554
weighted avg	0.53	0.49	0.46	38554

Figure 10. Classification report of Random Forest

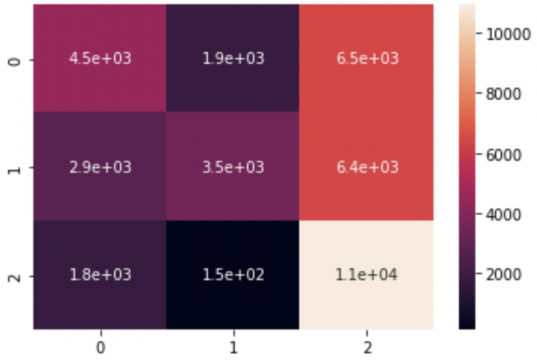

Figure 11. Random Forest confusion matrix

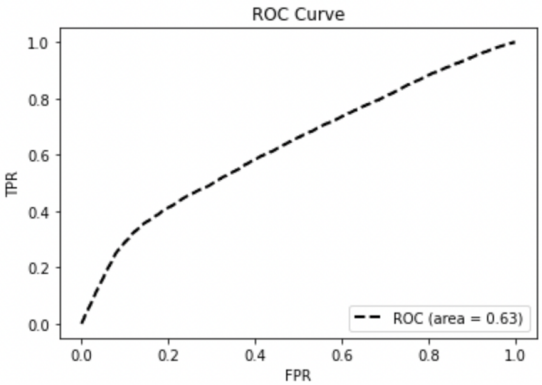

Figure 12. Random Forst ROC curve

For AdaBoost, we also used GridSearch to tune our model, and the best performance we recorded has 300 estimators and a learning rate of 1.0. The best training accuracy is 47.36% and the best testing accuracy is 48.51%. The classification report can be found below.

	precision	recall	f1-score	support
0	0.49	0.35	0.41	12835
1	0.64	0.27	0.38	12836
2	0.46	0.85	0.60	12883
accuracy			0.49	38554
macro avg	0.53	0.49	0.46	38554
weighted avg	0.53	0.49	0.46	38554

Figure 13. Classification report of AdaBoost

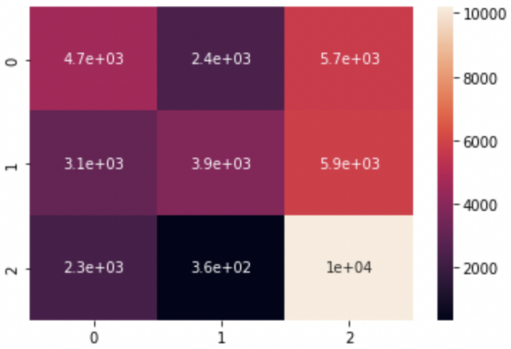

Figure 14. AdaBoost confusion matrix

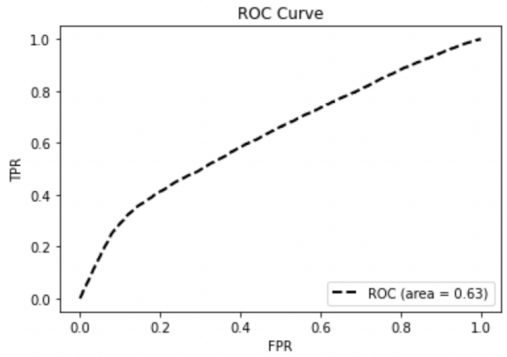

Figure 15. AdaBoost ROC curve

The accuracies for our model are not very satisfying, even if we spent much effort in tuning and re-process our datasets. We discussed our work and agreed on some reflections and takeaways for this project. We have chosen a very challenging area and used an innovative approach in this domain space. Crime is complicated and hard to predict by nature. It is not something we can randomly guess, and we were also challenged by the lack of existing literature on this topic. However, we believe that our project can serve as preliminary results and sources of inspiration for future work in this area. Looking forward, we think one modification we have is that we can add in more features including weather, and the profile characteristics of the criminal. In other words, we think our work might serve as a source of suggestion for the police department on what additional information they could include in their record, and could also be applied in areas like smart policing. In short, we have learned a lot from this project and we look forward to future iterations and expansion on this topic.

References

Almuhanna, Abrar A., et al. "Prediction of Crime in Neighbourhoods of New York City Using Spatial Data Analysis." *IEEE Xplore*, 1 Apr. 2021, ieeexplore.ieee.org/document/9425120. Accessed 12 Dec. 2022.

NYC Open Data. "NYPD Complaint Data Current (Year to Date) | NYC Open Data." *Data.cityofnewyork.us*, 9 Oct. 2022, data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-Year-To-Date-/5uac-w243. Accessed 13 Nov. 2022.

Rotaru, V., Huang, Y., Li, T., Evans, J., & Chattopadhyay, I. (2022). Event-level prediction of urban crime reveals a signature of enforcement bias in US cities. Nature Human Behaviour. https://doi.org/10.1038/s41562-022-01372-0