Physical Force in the  
 New York City Police Department

Aaron Kohn

Bellevue University, DSC 630

Fall 2021

# Summary

Many encounters between police and civilian results in the use of physical force by the police officer. This study uses data collected by the NYPD in 2011 to see if there is a way to predict whether an encounter is likely to involve the use of physical force by a police officer, before any contact is initiated. The analysis will look at the variables collected and made available by the NYPD that are believed to be available at the earliest stages of an encounter.

The primary methods used to create predictions for this project are decision tree-based models. A better understanding of the plans for how the predictions are going to be used is necessary to be able to focus on improving the models.

The best performing model found in this study is a decision tree-based ensemble, using gradient boosting. The model improves accuracy and accurately predicts high levels of cases that include the use of force. The model provides the ability of reaching 80% of cases that involve the use force in just 40% of stops. If costly procedures are needed to reduce the use of force, it can be limited to less than half of stops and still affect 80% of cases that involve use of force.

# Introduction

Although the New York Police Department’s (NYPD) has discontinued its controversial stop and frisk program, stops and frisks occur all the time. Many daily encounters between police officers and civilian end in a stop and frisk. Even when an encounter is not classified as a stop and a frisk does not occur there are hundreds of interactions between police officers and suspects every day.

Many encounters involve the use of force by a police officer. If an officer can be made aware of the level of risk of an encounter before entering a confrontation, steps can be implemented to proceed with the necessary tools to ensure a safe resolution for all involved. This includes providing the officer with the tools to identify situations that are likely to result in the officer using force. The officers can be trained in techniques in how to conduct themselves during these interactions to ensure that they are resolved in the safest outcome possible. This benefits the individual stopped, the officer personally, the entire police force, and society in general.

Using the data collected by the NYPD models will be built to predict stops that may include the use of force. A focus will be placed on finding models with a high recall while maintaining acceptable levels of accuracy. Further understanding would be needed to determine exact levels of recall and accuracy that would make models actionable.

# Methods

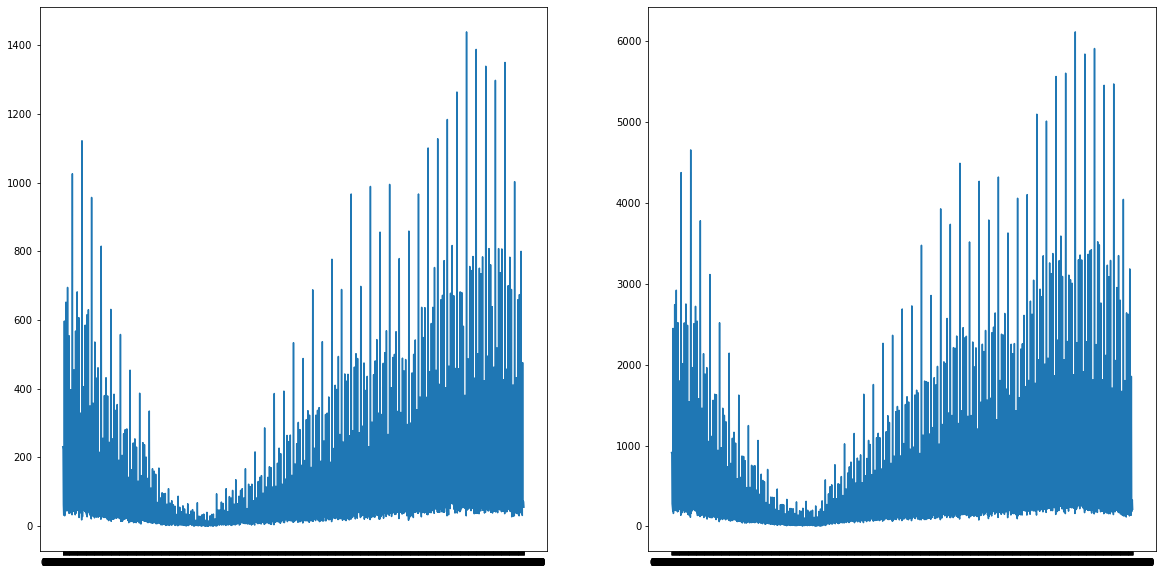
The NYPD has collected and made available data on stops conducted since the year 2003. Each year contains hundreds of thousands of encounters. This study focuses solely on the 2011 years data. This year was selected since it is believed to be the peak of New York Cities stop and frisk program, providing the largest amount samples in a single data set. Further validation of models can be performed on later years data. Testing the models’ life span can be tried as well to see how long the models maintain performance levels.

## Target Values

Identifying cases where physical force was used follows the NYPD’s classifications of physical force and includes the use of handcuffs, drawing a weapon along with pushing a suspect against a wall or the floor among others. The data set includes columns for each type of force. These columns will be combined into a new column indicating if force was used. A column indicating if multiple types of forces are used is added as well. The value is the number of force types used. This column has not been used in this study. Models will be created to predict the overall use of force.

## Variable Selection

In order to ensure that variables that are unknown at the time of model deployment do not affect the model, all variables need to be inspected to ensure that the information will be available to the officer prior to initiating an encounter. This requires manual inspection of each variable and an understanding of what it represents these variables are then removed from the data set that will be used for the model.

Time and date variables are removed in the final models. This study does not include time series analysis. If further steps are taken to implement the model it may be worthwhile to inspect the effects of time and dates and the use of force. Looking at the distribution of stops that include the use of force by time of day seems to follow a very similar distribution to all stops throughout the day.

**Stops That Include Use of Force**

**All Stops**

Ethical consideration must be given whether to include all variables in the final model. Tentatively, the study will not exclude any variables due to ethical considerations. This decision is made based on the purpose of this study which is to train police to be self-aware of their actions. If bias is discovered in current practice this should be addressed and decisions on ethical concerns can be made at that point. Removing these variables can lead to such issues to go undetected. It is important to remember that the focus of the study is on actions of the police officer and not that of the civilian. In no way does this data provide an accurate portrayal of the overall population. It is solely focused on those stopped by police.

Once initial modeling takes place important features will be identified. These features will be used to simplify the later models if limiting the number of variables doesn’t negatively affect the accuracy of the models.

Data Preparation

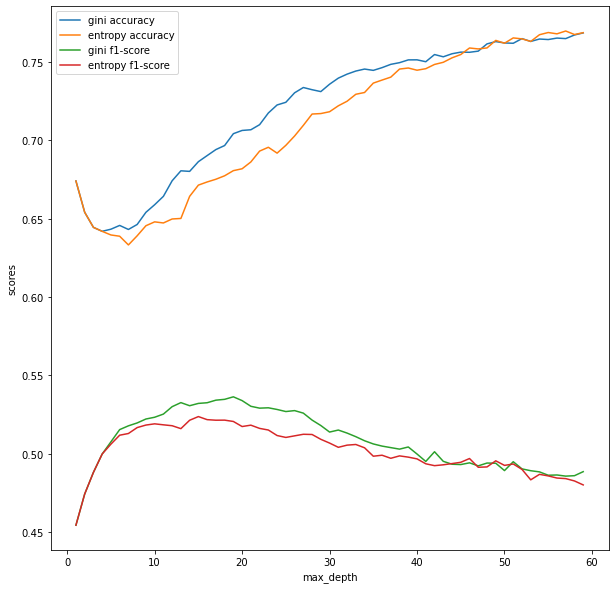
Histograms and bar plots are harnessed to quickly visualize the distribution of the individual variables. These also help identify outliers and mislabeled data. The age column which can be used in place of the dob column contains many wrongfully labeled data. When possible, correct age is calculated based off the dob and datestop columns. Care must be exercised when converting the columns to date time objects that dates are parsed correctly. Even after correcting and removing obvious mislabeled dates there still appears to be ages that are mislabeled. Further knowledge of data entry methods is needed to accurately address these errors.

Outliers need to be studied to ensure that variables are correctly labeled. Based on the data size it appears that missing data and outliers can be dropped without effecting the overall model. It is important to analyze outliers to see if there are any noteworthy patterns in these cases themselves.

All categorical data columns have been transformed to dummy encoding using the get dummies function. The models are trained on this data. As described above feature selection has been done to select important features from these variables.

## Model Selection

Initially a small sample of 50,000 observations was randomly selected. An 80/20 train test split was performed. The initial decision tree model achieved an accuracy of 0.7785470600040136 on the test data. The sample was increased to 150,000 and the model’s accuracy achieved was 0.7777591973244147. However, when selecting important features based on the model’s scores of over 0.00005 the number of features were reduced from 1906 to 618.

Successive models were run on the data to determine best depth to use for final model.Measures for overall accuracy and f1-score for instances where force was used were reported. Using these measures, a depth of 60 was determined best for accuracy and 16 for f1-scores. The results are reported in the following section.

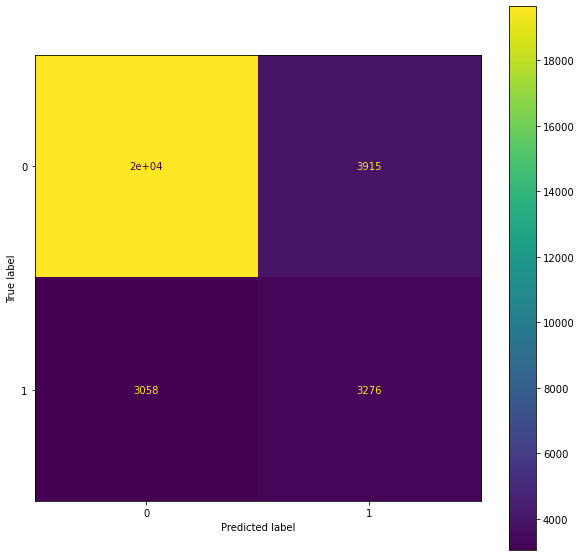
**Decision Model Scores**

Naïve Bayes and logistic regression models were attempted as well. The results for the logistic regression model are reported below. The Naïve Bayes algorithm was not able to predict all observations within the test data. It appears there were some values that did not appear in the training data.

# Results

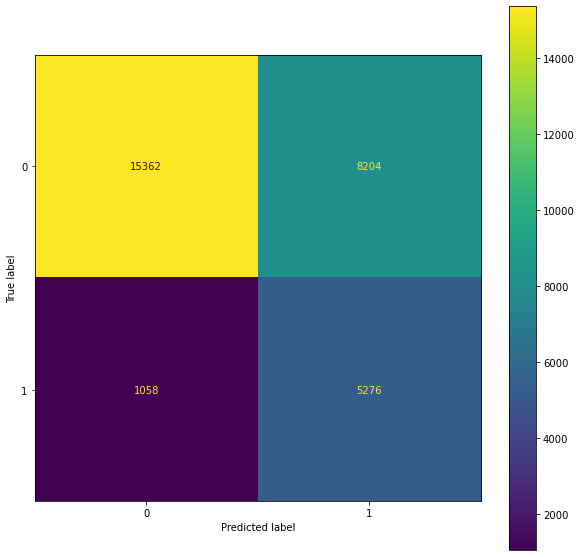
The decision tree model of a max depth of 60 which was selected for accuracy of 0.77. The recall for class 1 indicating the use of force was 0.52. It is important to keep in mind that class 0 is present at a rate of 78% in the data set.

**Results for Decision Tree. Max depth = 60**



Accuracy 0.77  
class precision recall f1-score  
0 0.87 0.83 0.85  
1 0.46 0.52 0.48

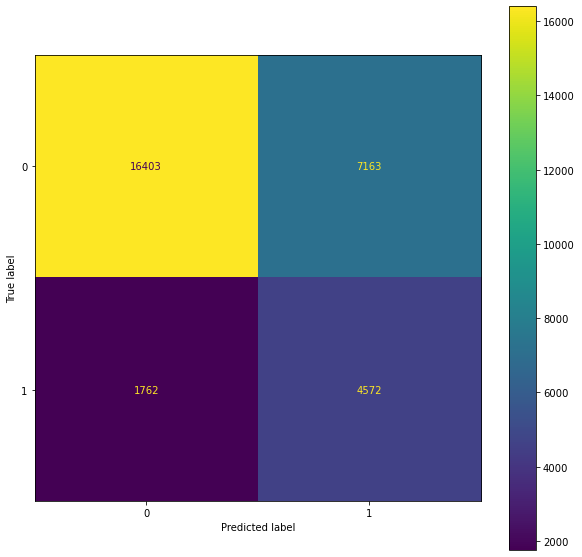
The model selected for highest f1-score was created using a depth of 16. F1-score was improved to 0.53 of 0.05. However, total accuracy fell to 0.69.



**Results for Decision Tree. Max depth = 16**

Accuracy 0.69  
class precision recall f1-score  
0 0.94 0.65 0.77  
1 0.39 0.83 0.53

The logistic regression model has an accuracy of 70% and f1-score of 0.51 for class 1.



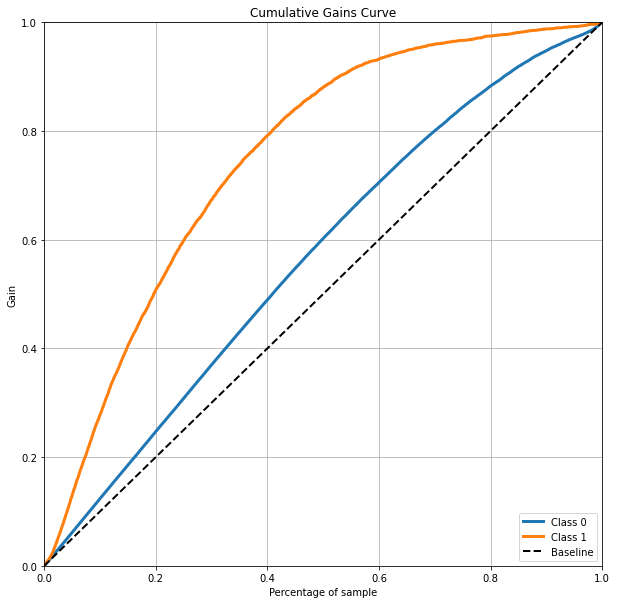
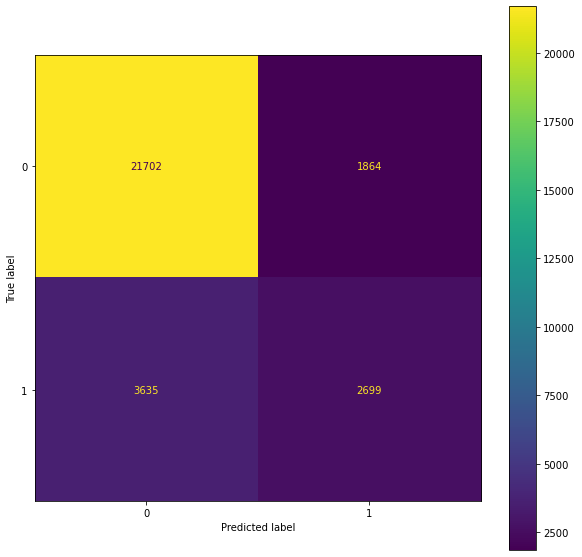
**Results for Logistic Regression**

Accuracy 0.70  
class precision recall f1-score  
0 0.90 0.70 0.79  
1 0.39 0.72 0.51

# Discussion/conclusion

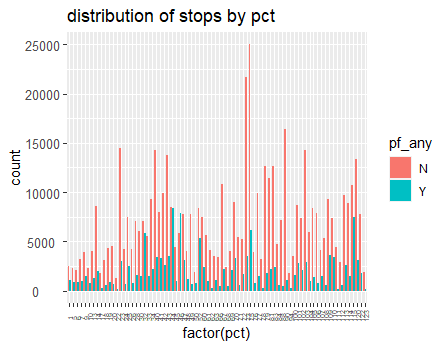
The models above do not achieve accuracies over 78%. If accuracy is used as the measure to determine the model’s performance, the above models do not provide much benefit. There may be gain based on precision, recall and f1-score. Further understanding on how predictions are to be used is still needed.

Using an ensemble method, particularly a gradient boosting classifier, improves accuracy to 82%. This is an improvement over the 78% found in the data. Looking at the gains chart shows the ability to capture a larger amount of the cases that included the use of force at lower percentages of the population.



**Results for Gradient Boosting Classifier**

Accuracy 0.82  
class precision recall f1-score  
0 0.86 0.92 0.89  
1 0.59 0.43 0.50

To further increase performance of models it may be worthwhile to train separate models for each precinct. The variation between percent of stops that included the use of violence is very large for the pct variable. Performing a chi-squared test returns an x2 of 47,498. 

# Acknowledgments

Although the work in this study has been performed solely by the author there are many that have provided input throughout the development of the project. A special thanks to Professor Alsaleem and Jake Rickord. The feedback and pointers they have provided greatly enhanced the outcome of the project.

# References

Data retrieved from <https://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk>

Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011

k.page