NYPD Stop and Frisk Data Analysis and Crime Prediction

Jainabou Barry Danfa
SI 671 Data Mining Final Project
University Of Michigan School of Information
Ann Arbor, MI
jainabou@umich.edu

Abstract

NYC stop-question-frisk policy has been the center of great debate in policing policies. The rise of predictive policing and data has caused underlying policies to become more prevalent. This project explores 2017-2018 data from stops in NYC and develops a predictive model to predict whether a person has a weapon on them or not. The model that performed the best was Gradient Boosting Classifier with a accuracy of .98. The features the model deemed important were mostly based on the officers suspicion of a weapon, indicating how bias can be baked into predictive models.

1 Introduction

New York City has implemented an infamous "Stop-Question-Frisk" policy, also known as stop-and-frisk, in efforts to get weapons and drugs off the streets. This policy has been widely controversial due to its effectiveness and legality. In one argument, stop-and-frisk reduces crime rates. Opponents argue that stop-and-frisk violates citizens fourth amendment rights and target specific demographics unfairly. This policy that has been in place since the 1990's has been centered in many policy debates and is still in practice today.

Substantial academic research has been completed on the racial disparities of the stop-and-frisk policy and the legal arguments supporting and opposing the policy. However, research is limited on the actual effectiveness of this policy and its implications when used in machine learning tasks. Many sectors in law enforcement are exploring ways to automate tasks and eliminate 'bias' in officer judgment of a potential crime. In this report we explore the most recent trends in the NYPD stop-and-frisk data and develop the best model for predicting if a person is in possession of a weapon or not. This will be followed by an analysis of important features in our prediction model and any bias that may be inherently baked into our model.

2 Related Work

In this section, I will discuss the previous works done around the stop-and-frisk policy to provide context to the analysis completed for this project.

2.1 Stop-and-Frisk Racial Disparities

In historical analysis of stop-and-frisk data from 2005-2010, we see that 89 percent of the stops involved nonwhites [5]. When someone was stopped, non-white suspects had a higher likelihood of being frisked compared to their white counterparts. In locations like Staten Island, where there is higher wealth, non-white suspects were stopped at a much higher rate, even though they represent the smallest percentage of residents in that area [5]. Many questions have been asked if these patterns are proof of racial disparities since there is no baseline to compare it to.

2.2 Stop-and-Frisk Legal Arguments

Many arguments have been made for stop-and-frisk in the legal courts. The argument against the policy argues that it violates a persons fourth amendment rights to have a valid reason to be stopped, not just 'suspicion' with sufficient burden of proof. Most things that have been cited by police as suspicion activity have been attributed to regular behavior of black males, causing racial profiling. Arguments for the policy state that officers operate colorblind and work to stop the crime in the city by investigating suspicion behavior before it happens. They say this has lead to a reduction in crime.

2.3 Predictive Policing

With the rise of data science and machine learning techniques automating task in all sectors, predictive policing has also been used to forecast crime for law enforcement. These techniques rely on historical data to build complex models that an predict crime before it happens based on features. This has received praise and criticism due to its ability to reduce costs for police departments but also bake in bias into decisions.

3 Data and Methods

3.1 Data Description

The dataset was accessed from New York City Police Department (NYPD) department with 22,637 stops from 2017-2018. After each stop, officers complete a UF-250 stop-and-frisk form that shows different aspects of the stop. The original dataset had 84 attributes for each stop. The main information used in our analysis is highlighted in table 1.

Table 1: Summary of Data Fields in stop-and-frisk form

Field	Description (μm)
SUSPECT BUILD TYPE	The suspects build
SUSPECT RACE	The suspects race
SUSPECT AGE	The suspects age
SUSPECT SEX	The suspects sex
DEMEANOR OF PERSON STOPPED	74 different categories of the officer reported demeanor of the susp
FRISKED FLAG	The suspect was frisked
SEARCHED FLAG	The suspect was searched
WEAPON FOUND FLAG	The suspect was found to have a weapon
SUSPECT ARRESTED FLAG	The suspect was arrested
ISSUING OFFICER RANK	The rank of the officer who issued the search
FIREARM FLAG	A firearm was found on the suspect
KNIFE CUTTER FLAG	A knife-cutter was found on the suspect
OBSERVED DURATION MINUTES	The time the suspect was observed before they were stopped
STOP DURATION MINUTES	The time the suspect was stopped
STOP LOCATION BORO NAME	The neighborhood the stop was conducted
SUMMONS ISSUED FLAG	A summons was issued during the stop
SUSPECTED CRIME DESCRIPTION	29 different categories of the crime suspected
PHYSICAL FORCE CEW FLAG	Physical force was used on the suspect using a conducted electron
PHYSICAL FORCE DRAW POINT FIREARM FLAG	Physical force was used on the suspect using by drawing a weapon
PHYSICAL FORCE HANDCUFF SUSPECT FLAG	Physical force was used on the suspect by handcuffing them
PHYSICAL FORCE OC SPRAY USED FLAG	Physical force was used on the suspect using by spraying pepper s
PHYSICAL FORCE RESTRAINT USED FLAG	Physical force was used on the suspect using by another restraint c

3.2 Data Analysis

The entire dataset was used to perform analysis on the characteristics of most stops. Correlation analysis was conducted between location. demographics, and being stopped and also changes over time. The results were showcased in various formats.

3.3 Prediction Methods

For the prediction models, different classifiers were compared for the category "weapon found" using the various fields in table 1. The models tested were Logistic Regression, Random Forest Classifier, Gradient Boosting Classifier, K-Nearest Neighbors, and Naive Bayes. These models were selected because of their various strengths for this specific dataset. The metric we use to judge the models performance is accuracy and F1 score. These two metrics will highlight the models performance on predicting whether a weapon will be found on a suspect and the presence of false positives and negatives.

Logistic Regression

Logistic Regression is a linear classification method that trains weights to minimize square error between ground truth labels and predicted labels. In our predictive task we can use logistic regression to classify if a weapon was found on the suspect.

K-Nearest Neighbors Algorithm

K-Nearest Neighbors is a non-parametic method used for unsupervised learning and supervised learning [4]. In supervised learning scenario, the method first records all training data with their labels and label test data according to its k-nearest neighbor. We use k-Nearest Neighbors algorithm to classify if a weapon was found on the suspect.

Random Forest

A random forest [7] is an ensemble classifier that trains multiple decision trees. Each decision tree is trained on a subset of original data. In predicting stage, each decision tree outputs a label and the final result is decided through a voting strategy. Random Forest largely improves test accuracy by reduce overfitting caused by a single decision tree.

Naive Bayes

Naive Bayes is based on the applying Bayes theorem to the network features. This model works well with features that are independent.

Gradient Boosting Classifier

Gradient Boosting works by minimizing the loss function for regression and classification problems. Having various ensembles of weak prediction models, the gradient boosting classifier usually has high predictive power.

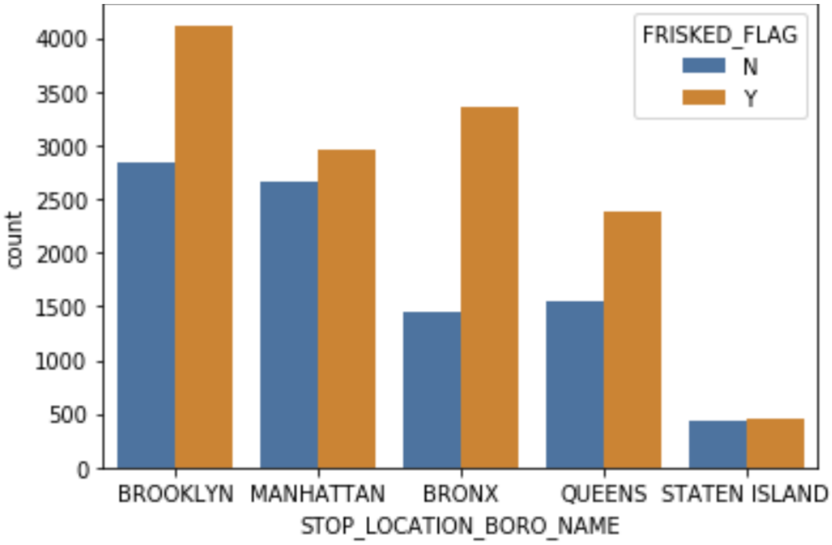
4 Results and Analysis

4.1 Data Analysis

Out of our 22,637 records,more searches happened compared to frisks. The frisk that occurred in Brooklyn and Manhattan are comparable while much more searched occurred in Brooklyn.

Blacks suspects made up most of the dataset, with Hispanics coming in the second place. As expected, the suspects age ranges from 18-60 years old, averaging 25 years old. It should be noted that the racial makeup of these neighborhoods is not as skewed as the makeup of who has been stopped in this dataset.

The data has less than 10 percent of stops where a weapon was actually found, although 30 percent of people stopped were arrested. The main suspected crime for the stop was possession of a weapon and robbery second and assault third. Less than a third of the suspected possession of a weapon were actually arrested for that. Most supervising officers that issued the stop were sergeants. Most stops were initiated by a radio call in and secondly self-initiated.



[htbp]

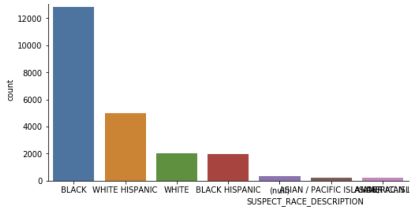


Figure 1: Race Distribution

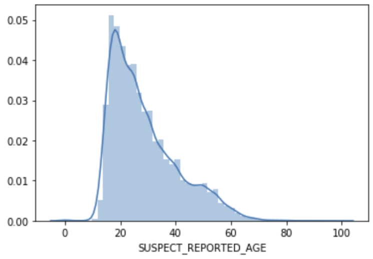
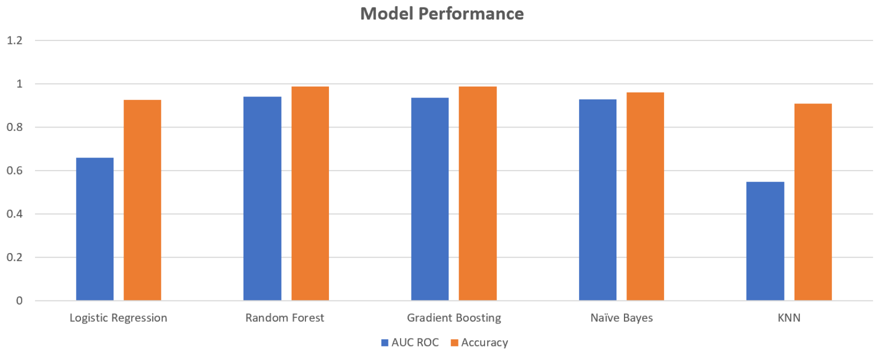


Figure 2: Age Distribution

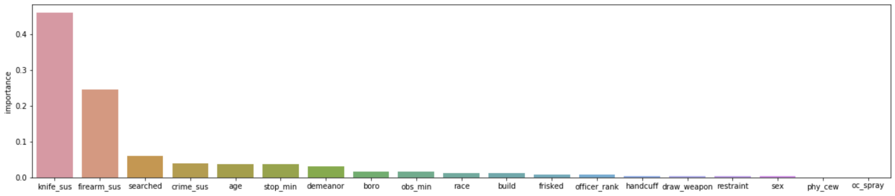
4.2 Model Performance

When testing various models to predict whether a suspect will have a weapon or not. Using the 24 features from the original dataset, we tested the predictive performance from our various models. We conducted a 80/20 split on the data. Overall the Random Forest and Gradient Boosting classifier has the best performance with AUC ROC scores and Accuracy.



The performance of Random Forest and Gradient Boosting being so high on this dataset does not surprise me given the way the data is set up. With the current high performance, I suspected some overfitting in the data but its performance on the test set was still comparable. Hyper-parameter tuning was not necessary due to the performance of the model with the baseline parameters.

The features that were considered more important on this data are mostly the suspect is thought to have a knife and/or firearm. These features are highly predictive, however not the best indicators to



use in a model since they are mostly based on the officers perception, which can be biased on other factors. Causing the judgement not to be biased. Also, the form is filled out after the officer knows if the suspect has a weapon on them or not. This could cause the reporting to be inaccurate.

5 Conclusion

Throughout this project, we saw the model that performed the best was Gradient Boosting Classifier with a accuracy of .98. The features the model deemed important were mostly based on the officers suspicion of a weapon, indicating how bias can be baked into predictive models. The nuance of policies and reporting can play a big factor into the data that is used for models and it is imperative that data scientist understand these nuances and use caution when completing 'plug-and-chug' models that could have life consequences for people.

References

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer–Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.

[4] https://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page

[5] https://www.rand.org/pubs/technicalreports/TR534.html