APMA4990: Modeling Social Data Final Project

Zachary Gleicher, Matt Piccolella, Edo Roth

Part 0: Project Background

For our project, we looked at New York City Stop and Frisk data.

What is Stop and Frisk?

Prior to 2013, Stop and Frisk was a policy under which the New York Police Department stop, interrogate, and search people with extremely vague criteria of suspicion. The policy reached it's peak in 2011, a year in which there were more than 685,000 stops. The Police Department claimed that the policy was effective at reducing crime; however, these claims were mostly uncorroborated[1] .In 2013, a court ruled that the practices of the NYPD were unconstitutional, violating Fourth Amendment rights. Since the ruling, Stop and Frisk has all but been stopped.

Why is the data important?

While Stop and Frisk has more or less stopped, the effects of the policy are still felt throughout New York[2]. With more and more occurrences of bias from police officers across the nation, it is becoming increasingly important to improve the ways that officers assess situations. Using data, we can visualize the impact of the policy, evaluate the biases of officers, and understand the inefficiencies and ineffectiveness of the policy.

Part 1: Obtaining the Data

Data for Stop and Frisk for each year from 2003 to 2014 is made available on NYC's website[3]. The data comes in large `.csv` files; each row represents a "stop" by a New York City officer, with each column representing a piece of information about the stop. The information for each stop is derived from a UF-250 form, which includes information about the stop; more than 112 features for each stop are recorded. For our project, we looked at only the data from 2012, which comes in a file called `2012.csv`.

Filtering the Data

To perform the analysis on the data, some filtering needed to be done. To do this, we created a script: `bin/parse_csv.sh`. The script loosely performs the following:

1. Picks out the columns we wanted to use (we picked ~30 of the 118 available)
2. Replaces 'Y' with '1', 'N' with '0', so our results are better interpreted by R
3. Filter out data with null or malformed data.

Generally, the data was pretty ready to work on, so this was a small part of our project.

Generating Coordinates of Stops

In order to conduct the analysis that we wanted to, we needed transform the `x` and `y` coordinates of each stop. The coordinates provided in the data come from the New York Long Island 3104 State Plan Coordinate system. The SPC is a Lambert conformal conic projection, which we had to then project back to latitude and longitude coordinates. We did this using `pyproj`, a Python library that can do the conversion. The code is available in `bin/convert_coordinates.py`. The output of this file needed to be parsed, as some of the outputs were invalid. We do this using `bin/reformat.sh`. These files were then output to `v11n/data/coordinates-pruned.csv`.

Part 2: Preliminary Data Analysis/Visualization

This part of our report follows the general Split/Apply/Reduce paradigm that we used throughout the course. This step involved R code to generate statistics about the data, which we then were able to understand visualize.

Objectives

- Answer the following questions:
  - Who is most affected by Stop and Frisk, and in what capacity are they affected?
  - When does Stop and Frisk happen the most?
  - Where in New York City do the stops occur?
  - For what what do officers stop people?
- Use visualization techniques to make the answers to these questions visible.

Analysis of Data

The code for the majority of the preliminary analysis of our data is found in `analysis/preliminary-stats.R`. The script uses `dplyr` to split, group, and analyze the data and `ggplot2` to plot the data.

This script creates four different graphs:

1. Number of Stops per Precinct
2. Number of Stops per Day
3. Number of {Stops, Arrests, Weapons} by Race
4. Number of Stops by Age

These graphs are available in `output/preliminary-analysis.pdf`.

In addition, we output the number of stops, arrests, and guns by race, which yield the following results:

```
      race  total arrests  guns
1    asian  17050   7.23% 1.13%
2    black 283991   5.65% 1.07%
3 hispanic 165049   6.44%  1.4%
4    white  50325   6.68% 2.05%
```

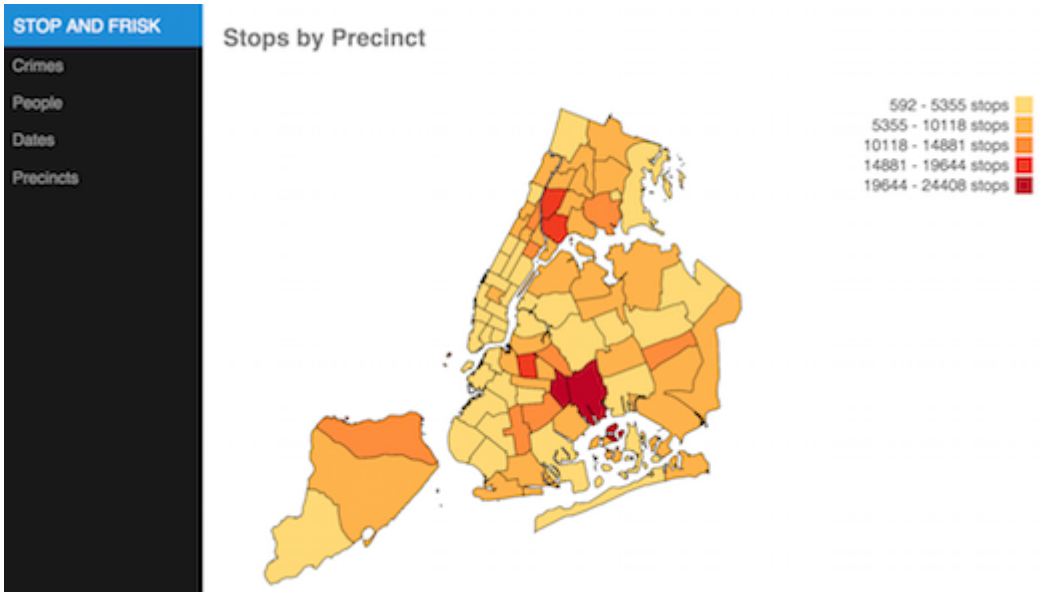
Additionally, we output the top 20 suspected crimes for which officers made stops:

```
detailcm incidences                                      label
1        20     129580              CRIMINAL POSSESSION OF WEAPON
2        85     116789                                   ROBBERY
3        14      60042                                   BURGLARY
4        46      44604                          GRAND LARCENY AUTO
5        31      37898                           CRIMINAL TRESPASS
6        45      29714                               GRAND LARCENY
7        27      26182          CRIMINAL POSSESSION  OF MARIHUANA
8         9      21175                                   ASSAULT
9        68      13505                               PETIT LARCENY
10       24      11788 CRIMINAL POSSESION  OF CONTROLLED SUBSTANCE
...
```

Visualization with D3.js

In addition to the visualization we did with `ggplot2`, `D3.js` allowed us to create visualizations that were web-based and interactive. The visualizations are based on data of four different categories: crimes, people, dates, and precincts. These include bar charts, pie charts, and a GeoMap. The visualizations are available at http://mattpic.com/stop-and-frisk.

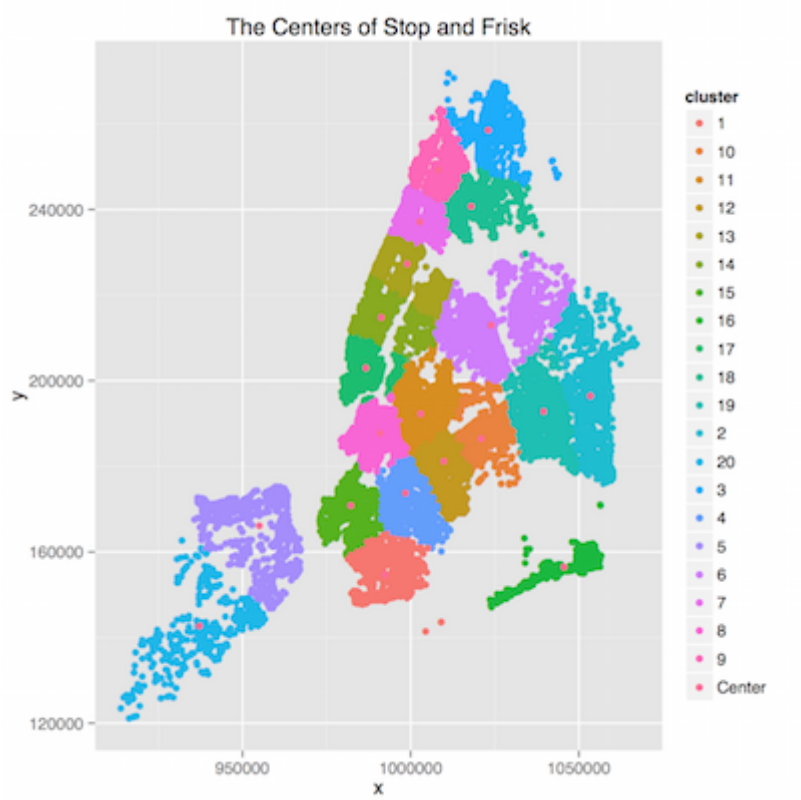
A sample of the visualizations is available here:



Part 2.5: Clustering

Initially, we had thought that clustering of the stops would be useful. We had thought that this would allow us to find the centers of Stop and Frisk in the city, which would then allow us to analyze what the centers corresponded with, whether it be police bias, income levels, etc. However, our initial results didn't show much, so we decided to move more toward prediction. However, we wrote the code to do clustering (both using a library and a function we wrote ourselves). Below is a

GeoMap with 20 centers plotted:



Part 3: Prediction Using Naive Bayes/Logistic Regression

This part of our report follows the analysis tools that we used in the second half of the class. Primarily, we used techniques of **classification**, which included Naive Bayes and Logistic Regression.

Objectives

Our main objective was to determine whether we can make more "effective stops." We will define an effective stop as one where an arrest is made after an officer makes a stop. In the dataset, about 94% of civilians are unnecessarily stopped. Only about 6% of people are arrested after being stopped. Our hypothesis is that we can make more effective stops by decreasing the percentage of innocent people who are stopped without substantially decreasing the number of effective stops. We plan to use Naive Bayes and logistic regression for classification and compare their results. Given that Naive Bayes and Logistic Regression classification produce results as a probability for arrest, we can examine possible probability thresholds to see if more effective stops can be made. (Note: our objectives for the project were based on previous research done by Microsoft Research[4],[5] .)

Choosing Features:

When an officer conducts a stop, he or she must fill out an UF-250 form which includes "yes or no" fields for an officer to identify why he or she decided to stop the suspect. We used these as the foundations for building our model. We then went through the dataset and looking for other useful features that could help predict arrests that are not necessarily "reasons for stop." Race is an

example of an additional feature. You can see the full list below. We made sure to exclude any features that would have happened after the stop. For example, the dataset has an indicator field that asks, "did the suspect have a weapon?" Although a “yes” response almost surely correlated with an arrest, this feature should not be used as given that the discovery of a weapon only happens after an officer has stopped and frisked an individual.

List of Reasons For Stop

Below you will see categories in the UF-250 form that officers can fill out as reasons for stop. All these categories are filled out with a “Y” for yes an “N” for no.

```
- Reason for stop - carrying suspicious object
- Reason for stop - fits a relevant description
- Reason for stop - casing a victim or location
- Reason for stop - suspect acting as a lookout
- Reason for stop - wearing clothes commonly used in a crime
- Reason for stop - actions indicative of a drug transaction
- Reason for stop - furtive movements
- Reason for stop - actions of engaging in a violent crime
- Reason for stop - suspicious bulge
- Additional circumstances - proximity to scene of offense
- Additional circumstances - evasive response to questioning
- Additional circumstances - associating with known criminals
- Additional circumstances - change direction at sight of officer
- Additional circumstances - area has high crime incidence
- Additional circumstances - time of day fits crime incidence
- Additional circumstances - sights or sounds of criminal activity
- Additional circumstances - report by victim/witness/officer
- Additional circumstances - ongoing investigation
```

Additional Features:

As described above, we included additional features which may not be technical reasons for stop, but do contribute to an officer’s decision to stop a suspect. When running our classification algorithms, the AUC increased by about 0.03. Below, you will see the additional features.

```
- Race: B: Black, W: White, A: Asian, P: Black Hispanic, W: White Hispanic
- Sex: F: Female, M: Male
- Build: H: Heavy, M: Medium, T: Thin, U: Muscular, Z: Unknown
- Height: 4: Tall, 3: Tall-Average, 2: Short-Average, 1: Short
- Age: Kept as numeric data
- Precinct: 76 different police precincts in the five boroughs
- Officer in Uniform: 0: Not in uniform, 1: In uniform
```

Cleaning Data Further

When looking at the data, we realized there were a lot of bad data points that were either clearly misentered, or not useful for our analysis. As discussed before, we first only selected those features that would be useful to classify whether an arrest should be made or not. We then filtered out

some points -- for instance, we had many values of age that that were over 200 and clearly misentered, so we limited our analysis to individuals under 100 years of age, a reasonable assumption as we only take out a miniscule percentage of values, most of which are likely incorrect. We also filtered out unknown values for race, sex, and build -- as we have so much data, we figured this would not assist in our analysis, and does not represent any significant portion of the population.

Percentages of data filtered out: Age: 0.02% Race: 3% Sex: 1.5% Build: 0.06%

Finally, we took the columns of height in inches and feet and combined them to form a total height column, categorizing them into discrete values. We adjusted for height in men and women, creating 4 categories of height based on these adjusted values.

```
4 = Tall, 3 = Tall-Average, 2 = Short-Average, 1 = Short
```

Naive Bayes

The first classification method we decided to try is the Naive Bayes algorithm.

Model

Given the imbalance of the data, with ~94% no-arrests, we realized that accuracy would not be the best metric for success. Instead, we chose to plot the ROC curve and compare the respective AUC values.

Additionally, we thought that the Naive Bayes classifier would perform better if we balanced the data. The reason for this assumption is that the prior distribution of the classes would lead to significantly higher posterior probabilities for no-arrest. We split the data into two groups: arrests and no-arrest. Then, in the classification task, we randomly sample 30,000 data points from each set in order to generate a balanced data set of arrests and no-arrests. After balancing the data, however, the AUC did not improve.

Most Decisive Features

In looking at the model, we examined the likelihoods for various features in order to see which differences would lead to a greatest deviance in the posterior probabilities. Most of the likelihood estimates remained fairly consistent for both classes.

Here are the features with the largest difference in likelihood between arrest and no arrest. All other features had differences less than 0.1.

Reason For Stop: Fits a Relevant Description

	No	**Yes**
No Arrest	0.8453423	0.1546577

	No	Yes
Arrest	0.6886918	0.3113082

Reason for Stop: Casing a Victim or Location

	No	Yes
No Arrest	0.6338087	0.3661913
Arrest	0.8301274	0.1698726

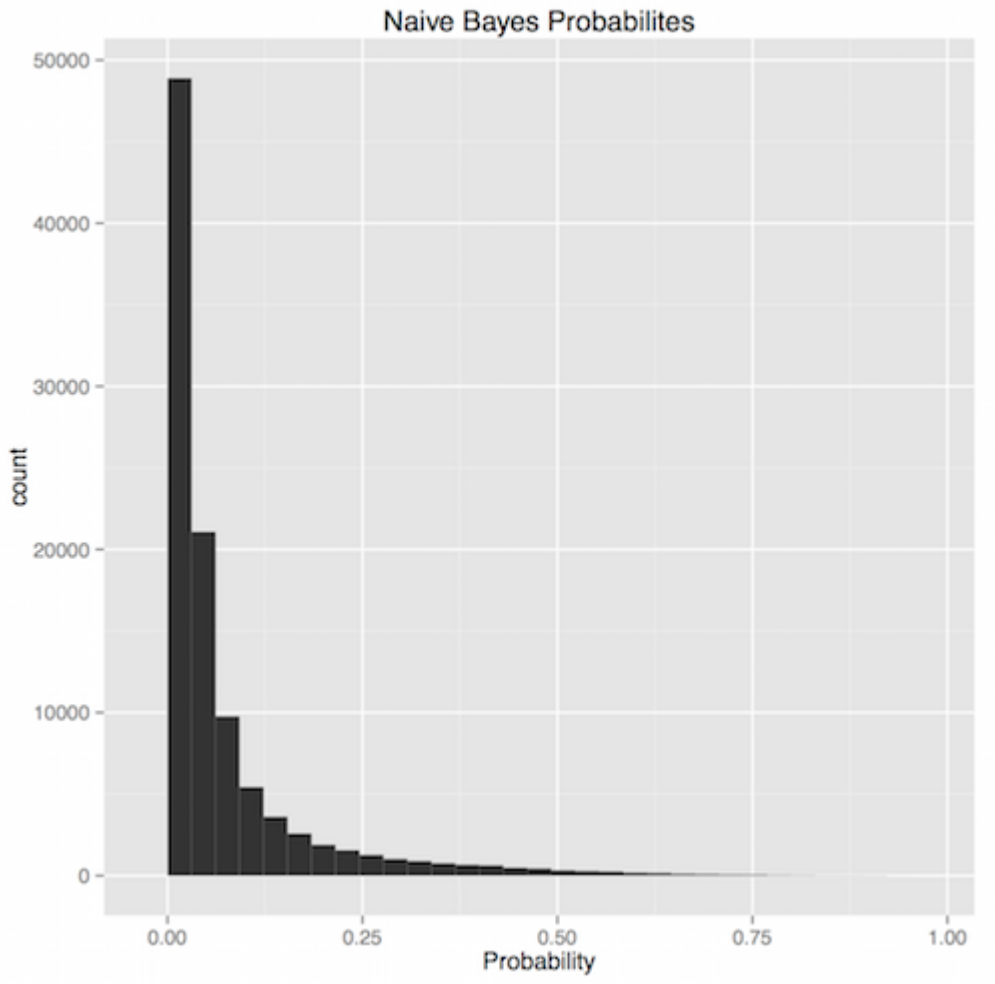
Additional Circumstances: Report by Victim/Witness/Officer

	No	Yes
No Arrest	0.8815675	0.1184325
Arrest	0.6976337	0.3023663

Was the Officer in Uniform

	No	Yes
No Arrest	0.2432714	0.7567286
Arrest	0.3628813	0.6371187

Probability Graph



In this graph, you will notice that there is higher probability mass towards predicting arrests. This is a result of the prior probability estimates for the no-arrest class being much higher than arrest.

ROC Curve

Plotting our ROC curve yields an AUC value of 0.7199524. It is important to use the ROC curve to evaluate the model because imbalanced data can lead to deceiving results when talking about accuracy. For example, if I had a dumb model that predicted no-arrest for every data point, we would get about 94% accuracy.

Logistic Regression

Additionally, we built a logistic regression model to classify the data as arrest and non-arrest. Since the feature space is small compared to the size of the dataset, we did not believe that a regularization method like lasso would lead to a dramatic improvement. We tested this hypothesis by implementing a logistic regression model with lasso using a loss function that maximized AUC, and it performed the same. The logistic regression model did slightly outperform the the Naive Bayes model by increasing the AUC by about 0.02 (AUC value of 0.7392117)

Here is the list of the 10 most predictive features for arrest given a stop:

```
"Get 10 best predictors for Arrest"
    pct42     pct25     pct19     pct44    pct105    pct102 cs_drgtr1      pct9
ac_rept1 cs_objcs1
0.5008073 0.5639895 0.6069064 0.6222683 0.6400148 0.7004087 0.8372146 0.8652883
0.9516741 1.0960916
```

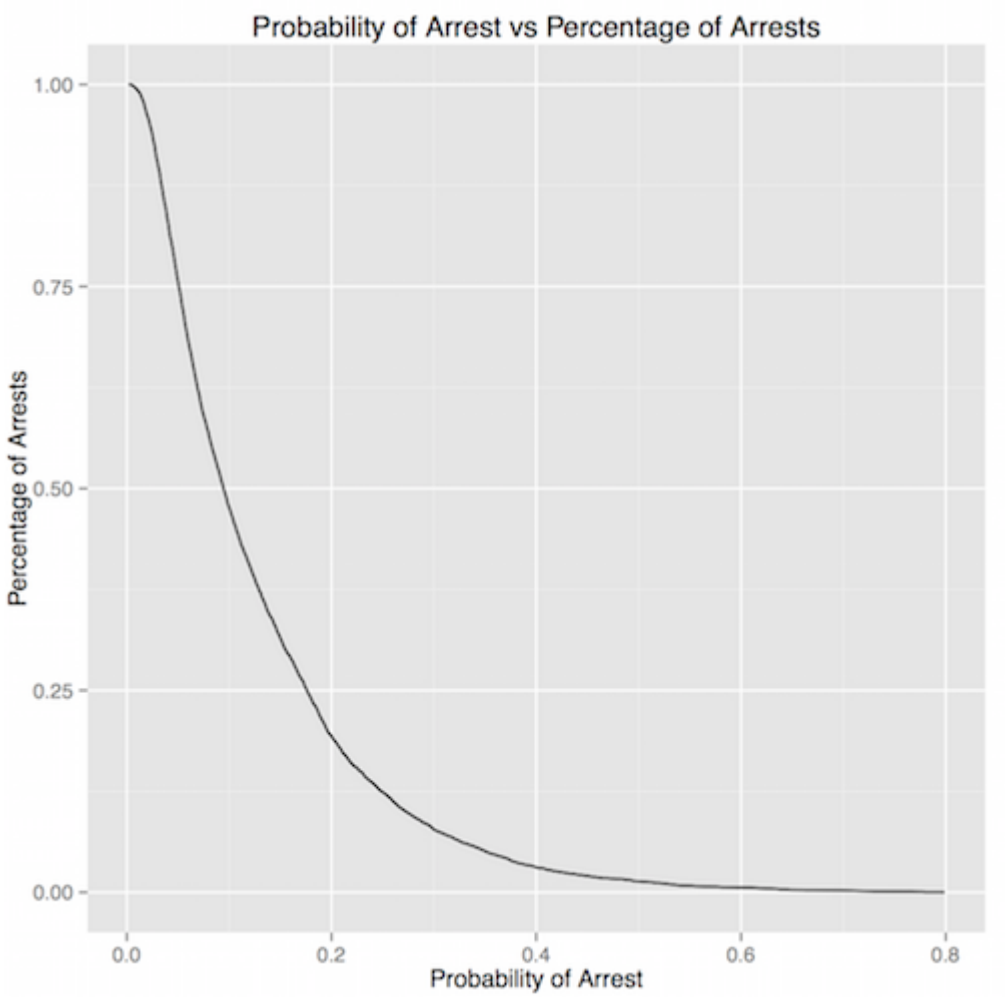
As you will notice, the most predictive features include precincts, but also include, “having a suspicious object,” “report by victim, officer, or witness,” and “actions indicative of a drug transaction.” We noticed that the most predictive precincts did not correlate with the most crime heavy areas in NYC, but rather they were in relatively safe neighborhoods. For example, police precinct 9 includes the East Village and NOHO. Perhaps this can be interpreted as police officers being more confident in stopping potential criminals.

Adaboost

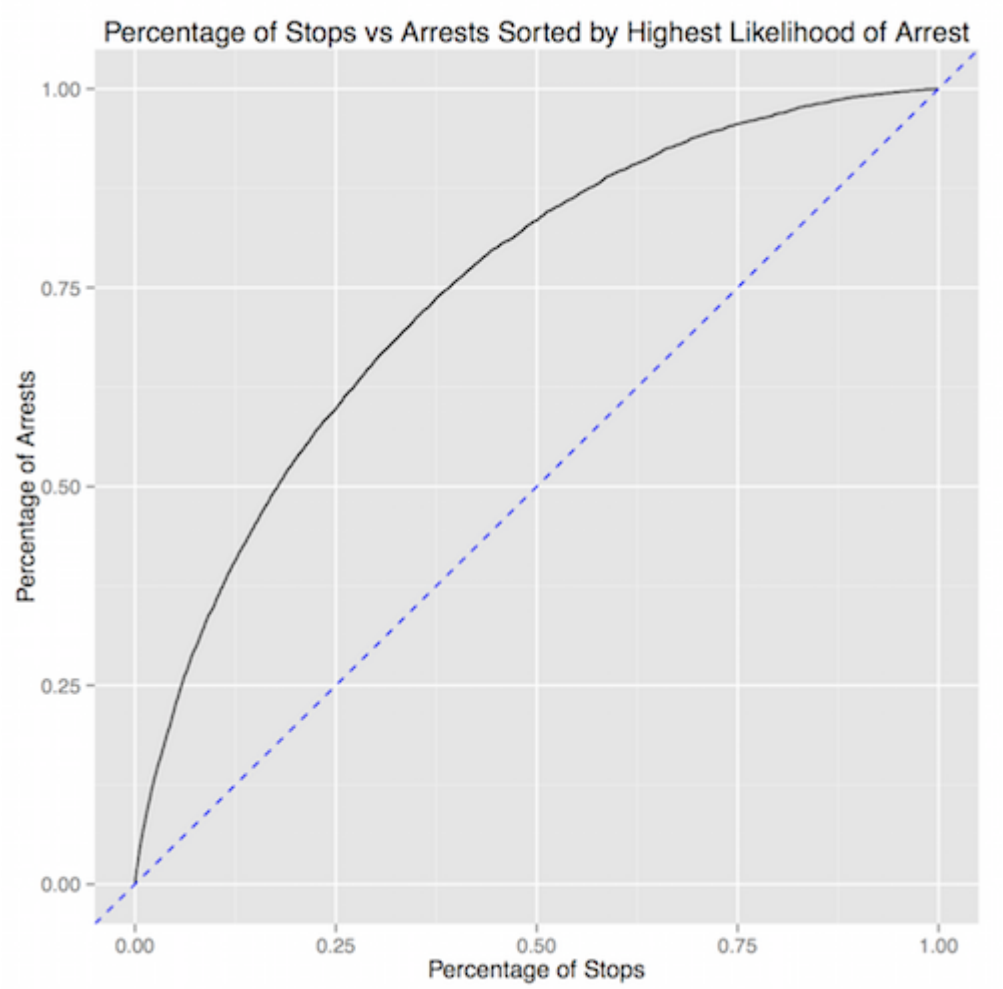
Additionally, we implemented the adaboost algorithm to see if we could get better results. Adaboost did not perform as well and had trouble working on such a large dataset.

Results

Using logistic regression, we successfully built a model that could predict more effective stops. In the graphs below, you will see that by tweaking the threshold for the probability of arrest, you can significantly reduce the number of ineffective stops.



This graph shows the predicted probability of arrest from the logistic regression model versus the percentage of arrests made. As you can see, the curves falls very quickly, which is due to the fact that the model predicts non-arrests much more heavily.



Here, you can see a graph comparing the percentage of stops vs the percentage of arrests where the data has been sorted by the the highest likelihood of arrest. Intuitively, we want to increase the area between the blue dashed line and the black curve. More area means that our classification algorithm substantially reduce ineffective stops. This graph shows that you can reduce the number of innocent stops by 50% while only reducing the number of guilty stops by about 20%. With the combination of both graphs, a probability threshold can be set in order to find a better trade off between minimizing the number of stops of innocent people, while still maximizing the number of stops of guilty people.

Sources

- [1] http://www.washingtonpost.com/blogs/the-fix/wp/2014/12/03/new-york-has-essentially-eliminated-stop-and-frisk-and-crime-is-still-down/
- [2] http://www.nytimes.com/2014/09/20/nyregion/friskings-ebb-but-still-hang-over-brooklyn-lives.html
- [3] http://www.nyc.gov/html/nypd/html/analysis_and_planning/stop_question_and_frisk_report.shtml
- [4] https://5harad.com/papers/frisky.pdf
- [5] https://ds3.research.microsoft.com/doc/sqf.pdf