Examining the NYPD Arrests Dataset

Binxin Liu (bl642), Mengjia Xia (mx233)

Abstract

In New York City, the stop-question-and-frisk program is a NYC Police Department (NYPD) practice of temporarily detaining, questioning and searching civilians on the street. However, this program became the subject of a racial profiling controversy. The executive director of the New York Civil Liberties Union once pointed out no meaningful progress was made in reducing racial disparities in the program. In this report, we fitted a logistic regression classifier and a kernel SVM classifier to predict which types of suspects would be arrested. F_1 scores of the two classifiers are 0.825 and 0.861 and AUC values are 0.915 and 0.852 respectively. We found that carrying contrabands and weapons is a sign of being arrested. Besides, the arrestment decision also depends on time, location, and the economic situation. Furthermore, based on demographic parity and equalized odds, we found the unfairness in the decision of arrestment over sex and race, which supports the existence of disparities.

Contents

1. Introduction

The *stop-question-and-frisk* program is a New York City Police Department (NYPD) practice of temporarily detaining, questioning, and at times searching civilians on the street for weapons and other contraband [1]. Every time a police officer stops a person in NYC, the officer is supposed to fill out a form recording the details of the stop, of which we used in modeling will be discussed later (see, e.g., UF-250 Form Stop, Question and Frisk Report Worksheet 2016 Version). The forms were filled out by hand and manually entered into an NYPD database until 2017 when the forms became electronic.

However, this program became the subject of a racial profiling controversy. To test whether such disparities exist in who is stopped and arrested by police on street, we obtained Stop, Question and Frisk Data from NYPD website, which records every stop, question and frisk effected in NYC ranging from 01/01/03 to 12/31/18. We used classification predictive modeling to see which types of suspects would be arrested with a higher probability and measured fairness in terms of NPR. We found the unfairness in the decision of arrestment over sex and race, which supports the existence of disparities.

The rest of the report is organized as follows. An

[1] https://en.wikipedia.org/wiki/Stop-and-frisk_in_New_York_City

exploratory data analysis is given in Section 2. Problem is formulated in Section 3.4. We present details of linear and non-linear modeling in Section 4 and 5, followed by our interpretations of the results in Section 6. Finally, we conclude the report in Section 7.

2. Exploratory Data Analysis

2.1 Overview

For each recorded sample in Stop, Question and Frisk Data, it has 112 features, including dates, times, locations, physical features of the suspect, crime suspected, details about what happened at the time and so on. There are more than 5 million records during the 16 years.

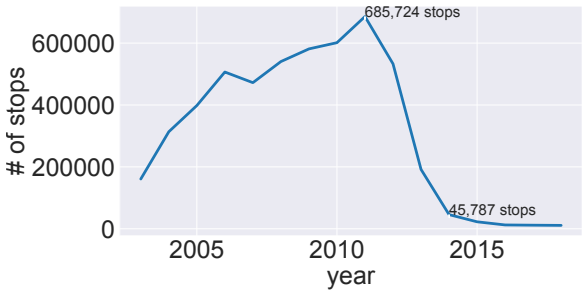


Figure 1. Number of Reported Stops by Year

Since Mayor de Blasio came into office in January 2014, the number of reported NYPD stops has drastically declined in New York City, as shown in Figure 1. Considering time efficiency of training our models, the personnel change in 2014 as well as the advent of electronic form in 2017, we investigated the records from 2014 to 2016 (80754 records).

To incorporate the influence of the economics situation in the New York City, we acquired the monthly Consumer Price Index (CPI) and S&P 500 index data from Wind Financial Terminal. We used the 1st lag term of CPI and S&P 500 index to capture the change of economic situation in New York City.

2.2 Data Cleaning

Since the data was manually extracted from forms, some variables are messy and abnormal because officials' handwriting was not easy to identify. For example, we found that the age of one suspect was recorded as 1 and there are missing values in some columns.

As the number of missing values is small, we directly **dropped rows with missing values**. To deal with **outliers**, we **winsorized 1%** on age, weight, and height. Besides, we dropped rows with values of sex, race, hair color, eye color and build labeled as unknown. After that, we had 75343 observations in total.

The variable "crimsusp" (crime suspected) is an important indicator of arrestment decision. However, the values were messy due to (1) different abbreviations used by different officials, (2) typo, and (3) the use of Penal Law code instead [2]. We **manually matched values that represent the same crime type**, resulting in 53 types. Note that if there exists more than one crime suspected, we only kept the first crime.

Please see Table 1 and to get a feel for the structure of the data set. A more detailed exhibition data (after feature engineering) can be found in our GitHub Repository.

These features can roughly be divided into 8 categories: (1) date, time, location of the event, (2) reasons for being searched, frisked or stopped, (3) demographic features of the suspect, (4) actions taken by the police officer, (5) details about the suspect, like type of identification, whether carrying certain kinds of weapons, etc., (6) additional circumstance such as the suspect's attitude, sights or sounds of criminal activities, etc., (7) economic situation such as CPI and S&P 500 index, and (8) others.

2.3 Data Analysis & Visualization

Figure 2 shows the number of arrestment and non-arrestment **across the time** and its corresponding arrest-made rate. Though the total number of stops decreased, the arrest-made rate increased greatly. Arrest-made rate during the second half of the year was higher than the first half, especially in November and December, the holiday season. In addition, we can see that the arrest-made rate reached its peak during the commuting time (7:00-9:00 and 18:00-20:00). Though most stops happened at midnight, the arrest-made rate was not that high compared with the daytime.

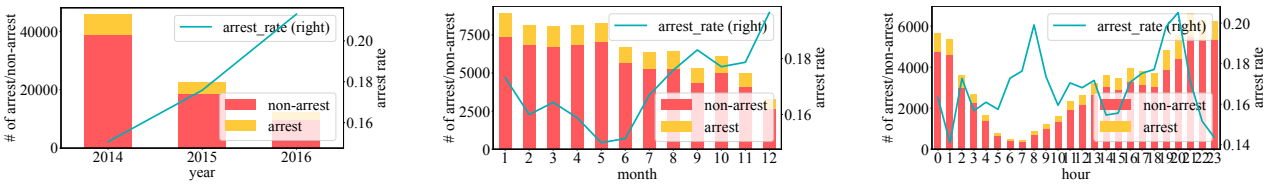
Figure 3 shows the **geographical distribution** of these observations. Color turning purple to red means the higher frequency of stops, questions or frisks, while bigger red circles mean higher arrest-made rate. As we can see, arrest-made rate varies a lot in different precincts.

We also investigated arrest-made rate across several **suspect features** which was visualized in Figure 4. The arrest-made rates of older suspects are illustrated to be higher, while arrest-made rates over different weights

[2] For example, for the crime type of "Criminal Trespass", it is recorded as "CRIM TRESS", "CRIM TRESPASS", "TRESSPASS", "140.1", etc.

arstmade	timestop	crimsusp	frisked	searched	...	sex	race	age	weight	haircolr	eyecolor
0	1410	16	1	0	...	M	B	18	150	BK	BR
1	930	32	1	1	...	M	B	31	160	BK	BR
0	1260	16	1	0	...	M	B	16	160	BK	BR

Table 1. Data Set Example



(a) By Year **(b)** By Month **(c)** By Hour

Figure 2. Arrest Across Time

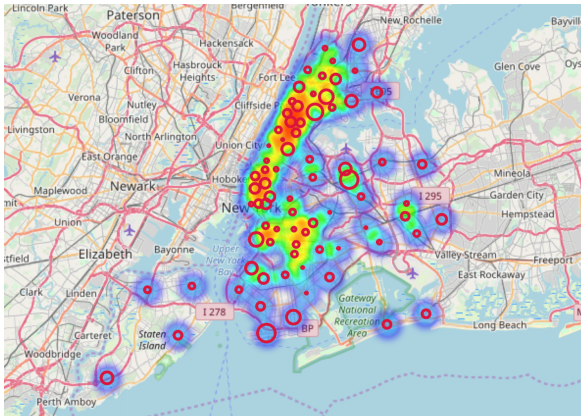


Figure 3. Arrest Across Location

and heights did not show significant disparity because the number of suspects with extreme weights/heights is very small. The arrest-made rate of females was much higher than that of males, but there was far fewer women in our samples. In terms of race, the rates of P (Black Hispanic), Q (White Hispanic), and Z (other) were the highest. For the rest of features, there was no significant difference taking the small number of samples with certain features in consideration.

Figure 5 shows the **correlation between dominant features** whose correlation coefficient with respect to the predicted variable were larger than 0.1. They were listed in Table 2. Here, we applied Pearson Correlation for pair of continuous variables and Cramer's V for pair of categorical variables. Only 13 features were selected, which means that the majority of features could not linearly explain the variation.

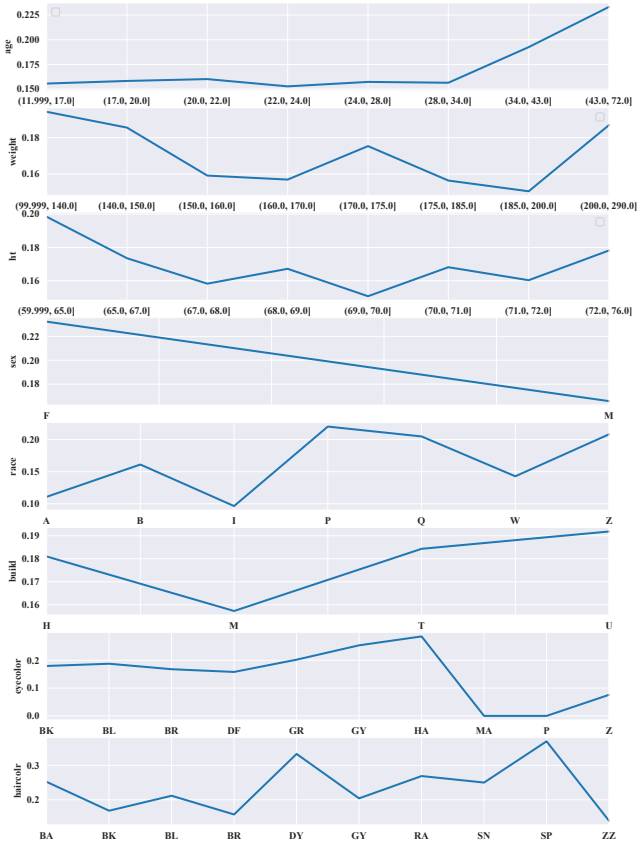


Figure 4. Arrest-made Rate over Suspect Features

sex: F:female, M:male

race: A:Asian, B:Black, I:American Indian, P:BlackHispanic, Q:White Hispanic, Z:Other

build: H: heavy, M:medium, T:thin, U:muscular

haircolor, eyecolor: BK:black, BL:blue, BR:brown, DF:two different, GR:grenn, GY:gray, HA:hazel, MA:maroon, PK:pink, VI:vilvet

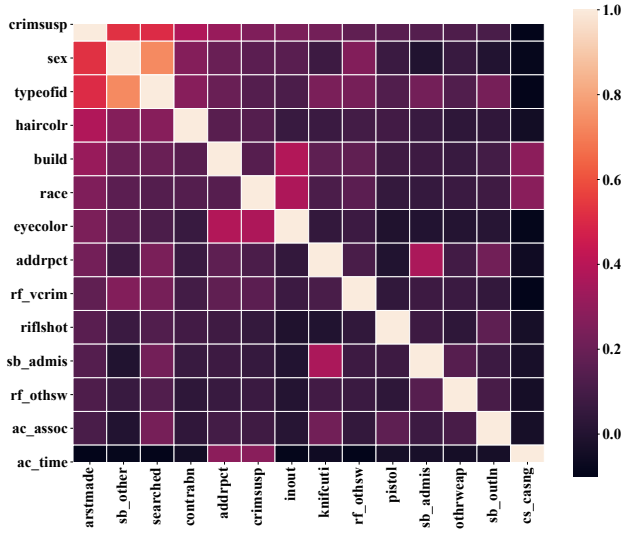


Figure 5. Dominant Features Correlation Heatmap

3. Problem Formulation

Prediction of arrestment decision is to figure out what kind of stop, question, or frisk and what kind of suspect will have a higher probability/ability to be arrested. We attempted to train models to grasp these probabilities/abilities.

3.1 Feature Engineering

Before modeling, we performed feature engineering, which can substantially boost machine learning model performance by providing a well-structured form of data.

Specifically, we used (1) **binary encoding** for the Boolean features. (2) **One-hot encoding** was applied to the categorical features (crimsusp, addrpct, sex, race, hair color, eye color, build, etc.). Meanwhile, for the sake of further stratified split on data set, we (3) **merged some categories together** in features such as crimsusp, eyecolor, and haircoler. To grasp the non-linear relationships, we applied the technique of (4) **polynomial transformation** on timestop, CPI, and stock market index.

Those continuous variables such as year, month, height, weight, and economic data were **standardized** before being used. This procedure was indispensable and imposed a significant influence on the final results.

After feature engineering, there were 207 features in total. Based on such a large number of features, regularization seemed to be necessary to prevent over-fitting.

3.2 Validation Method

To validate our models, we applied **stratified cross validation** with 5 folds. Stratification here is to ensure that each fold has the same feature structure with the whole data set. For example, if there were about 18% female samples in total, then in each fold, there were about 18% female.

For each fold, we used the 80:20 principle to split the train-test data set. Again, it was a **stratified split** rather than a simple random split.

Feature	Description	Type
sb-other	basis of search: other	boolean
searched	was suspect searched	boolean
contranb	was contraband found on suspect	boolean
addrpct	stop address precinct	Categorical
crimesusp	crime suspected	Categorical
inout	was stop inside or outside	boolean
knifcuti	was a knife/cutting instrumts found	bollean
rf-othsw	reason for frisk: other susp of weapons	boolean
pistol	was a pistol found	boolean
sb-admis	basis of search: admission by suspect	boolean
otherweap	was other weapons found	boolean
sb-outln	outline of weapon	boolean
cs-casng	reason for stop: casing a victim or location	boolean

Table 2. Dominant (High Correlation) Features

3.3 Model Performance & Fairness Metrics

Following standard performance metrics are used for evaluating our classifier: **Accuracy (ACC), Precision (PPV), Recall (TPR) (from Confusion Matrix), AUC, and F_1-score.**

$$ACC = \frac{TP+TN}{P+N}, \quad PPV = \frac{TP}{TP+FP}$$
$$TPR = \frac{TP}{p}, \quad F_1 = 2\frac{PPV \cdot TPR}{PPV+TPR}$$

Note that AUC and F_1-score are overall metrics for performance of the classification model.

In addition to performance metrics, fairness is also an important aspect to be explored. Donna Lieberman, Executive Director of the New York Civil Liberties Union, said: "New York City is safer than ever, but we have made no meaningful progress in reducing the **racial disparities** in who is stopped by police on the street."[3]

[3]https://www.nyclu.org/en/press-releases/nyclu-releases-report-

Though many definitions of fairness have been proposed in the literature (see [1] for an overview), we used **demographic parity** and **equalized odds**, which capture the "group fairness".

The classifier satisfies demographic party if the prediction is independent of the protected attribute:

$$\mathbf{P}(\hat{y}|a=1) = \mathbf{P}(\hat{y}|a=0) = \mathbf{P}\hat{y}.$$

The classifier satisfies equalized odds if the prediction is independent of the protected attribute a conditional on the outcome y:

$$\mathbf{P}(\hat{y}|y,a=1) = \mathbf{P}(\hat{y}|y,a=0) = \mathbf{P}\hat{y}|y.$$

Here, we used false positive rate to measure in what extent we harm the innocent suspects who should not be arrested. This report focused on fairness with respect to **sex and race**, which will be shown in the later part.

3.4 Implementation

Throughout the report, we implemented all models with Python and the Scikit-Learn [2] library.

4. Linear Modeling: Logistic Regression

4.1 Model Overview

As we said in the problem formulation part, the problem was to build models that can explain what kind of suspects or events feature leading to arrestment with higher possibilities. This aligns perfectly with the logistic regression, which naturally has a probability explanation. Logistic regression transforms the continuous value obtained from linear regression into a Bernoulli random variable, which is shown below.

$$\mathbf{P}(y=1) = logistic(w^T x)$$
$$\mathbf{P}(y=0) = logistic(-w^T x)$$

where,

$$logistic(u) = \frac{exp(u)}{1+exp(u)}$$

4.2 Train with Raw Data

First of all, we trained a logistic regression classifier on all cleaned data (75343x207). The result from cross validation was displayed in Table 3. The fairness performance of this model was shown in Table 4.

From Table 3, the TPR and F_1 are pretty low compared to other metrics, which means that this model performed poorly on finding positive samples out of which should be positive. However, such poor performance was mostly caused by the fact that the data set was unbalanced: only 16% of samples were positive!

		1	2	3	4	5	avg
TPR	**trn***	.630	.629	.628	.630	.630	.629
	tst*	.626	.623	.619	.627	.627	.624
PPV	**trn**	.820	.823	.823	.819	.820	.821
	tst	.817	.818	.822	.814	.814	.817
ACC	**trn**	.914	.914	.914	.913	.913	.913
	tst	.913	.912	.012	.912	.912	.912
F_1	**trn**	.713	.713	.712	.712	.712	.712
	tst	.709	.707	.706	.708	.708	,708
AUC	**trn**	.915	.915	.916	.917	.915	,916
	tst	.913	.915	.912	.905	.915	.912

Table 3. Performance of Model on Complete Data set
* trn is short for train, tst short for test

Sex*	**F**	**M**					
coef	-.14	-.48					
FPR	.035	.027					
Race*	**A**	**B**	**I**	**P**	**Q**	**W**	**Z**
coef	-.14	-.04	-0.44	.11	.06	-.27	0.08
FPR	.01	.03	.01	.03	.03	.03	.01

Table 4. Fairness of Model on Complete Data set
* F:female, M:male, A:Asian, B:Black, I:American Indian, P:Black Hispanic, Q:White Hispanic, Z:Other

As for how fair was this model, Table 4 told us that the model pushed female, Hispanic (P, Q) to be arrested much more than male and other races. When checking FPR over different sexes and races, we can find that this model was surprisingly fair, there did not exist significant differences across sexes and races.

4.3 Train with Balanced Data

Given that the raw data set was unbalanced, the number of arrested was far less than the unarrested, it is reasonable to train a new model based on a balanced data set. We used **under-sampling** to balance the data. Specifically, we kept all the samples labeled as 1 and part of the samples labeled as 0 (unarrested) to ensure that the number of positive samples equals that of the negative one. We also used stratified splitting to maintain the feature structure. After balancing the data set, we got 24670 samples.

Table 5 displayed the average of results from 5-fold

analyzing-nypd-stop-and-frisk-data

cross validation. The overall performance increased considerably and reasonably after the data set was balanced!

	TPR	PPV	ACC	F_1	AUC
trn	.787	.877	.855	.830	.920
tst	.782	.872	.850	.825	.915

Table 5. Performance of Model on Balanced Data set

According to coefficients of demographic features in Table 6, suspects with different sex or race had a disparate probability of being arrested. Such demographic disparity was inevitable if we desired a better classification performance. Meanwhile, the FPR varied greatly from female to male and across the races. Such phenomena could be explained by the disparity which was a fact supported by our data analysis before and what was said by the executive director, Donna Lieberman. Balanced data helped the model learn such disparity so that the performance of the model boosted while the fairness of the model declined.

Sex	**F**	**M**					
coef	.14	-.09					
FPR	.12	.09					
Race	**A**	**B**	**I**	**P**	**Q**	**W**	**Z**
coef	-.08	-.01	-.01	.19	.10	-.21	0.05
FPR	.03	.10	.20	.12	.11	.09	.16

Table 6. Fairness of Model on Balanced Data set

4.4 Fitting Analysis

Based on the results shown above, there was little evidence supporting that the model was over-fitted or under-fitted so that we claimed the model was properly fitted. Specifically, the performance metrics on both train and test sets were relatively high. However, here, we would like to add an l_1 penalty to the loss function in order to firstly test our claim on the good fit and secondly encourage sparsity and therefore interpretability. We still trained the model on the balanced data set.

From Table 7, the performance was not improved essentially by using a regularization. This could be a sign that our model was not over-fitted at all despite a large number of features used.

	TPR	PVV	ACC	F_1	AUC
trn	.783	.878	.853	.828	.920
tst	.780	.874	.850	.824	.915

Table 7. Performance of l_1 Penalized Model

Changes in terms of fairness in Table 8 shows that adding l_1 regularization increased unfairness across races and sexes.

Sex	**F**	**M**					
coef	0.01	-0.23					
FPR	.13	.09					
Race	**A**	**B**	**I**	**P**	**Q**	**W**	**Z**
coef	-.10	-.02	-.00	.14	.06	-.21	.00
FPR	.02	.09	.20	.12	.11	.08	.16

Table 8. Fairness of l_1 Penalized Model

5. Non-linear Modeling: Kernel SVM

5.1 Model Overview

In this section, we tried to use non-linear models to extract more information from the data set and thus improve performance. We used Kernel SVM models. As a warm-up, let us briefly review Kernel SVM models. The goal of SVM is to find a hyperplane with the largest distance to the closest training examples. The optimization problem of soft-margin SVM is formulated as follows:

$$\begin{aligned} &\min_{\mathbf{w},b,\xi} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\xi_i \\ &s.t.\ \ y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i \quad \forall i \in [n] \\ &\qquad \xi_i \geq 0 \quad \forall i \in [n], \end{aligned}$$

where C is a parameter that controls trade-off between margin and training error. If data is not linearly separable, we can incorporate kernel methods to construct a non-linear boundary. Using kernels, we map the input space into the feature space, which usually has a higher dimensionality. Four types of kernels are often used:

- Linear: $k(\vec{a},\vec{b}) = \vec{a}\cdot\vec{b}$
- Polynomial: $k(\vec{a},\vec{b}) = [\vec{a}\cdot\vec{b}+1]^d$
- Radial Basis Function (RBF): $k(\vec{a},\vec{b}) = e^{-\gamma[\vec{a}-\vec{b}]^2}$
- Sigmoid: $k(\vec{a},\vec{b}) = \tanh(\gamma[\vec{a}\cdot\vec{b}]+c)$

Note that when we fitted the logistic classification models, we manually calculated the quadratic and cubic terms of some features to incorporate non-linearity. We removed such features in this section.

5.2 Model Performance

We implemented SVM with 4 different kernels and set C as 1. The results were reported in Table 9. Though Poly SVM outperformed others on the training set, its performance on the test set showed a sign of over-fitting. Based on the results, we chose the RBF kernel. Compared with the best logistic model (on balanced data without regularization) we trained before, the F_1 and TPR increased while AUC decreased.

We also reported fairness of the RBF kernel model in terms of FPR in Table 10. Though FPR is low among difference sexes or races, the differences still show the unfairness. We do not report coefficients of these features because the classification boundary is defined in a higher dimensionality. Lack of interpretability is a drawback of non-linear models.

		Linear	Poly	RBF	Sigmoid
TPR	**trn**	.842	.921	.910	.819
	tst	.833	.830	.864	.831
PPV	**trn**	.840	.843	.893	.829
	tst	.839	.840	.862	.834
ACC	**trn**	.839	.918	.907	.816
	tst	.831	.833	.858	.830
F_1	**trn**	.840	.923	.912	.812
	tst	.830	.827	.861	.828
AUC	**trn**	.830	.911	.901	.808
	tst	.823	.826	.852	.824

Table 9. Performance of Kernel SVMs

Sex	**F**	**M**					
FPR	.15	.08					
Race	**A**	**B**	**I**	**P**	**Q**	**W**	**Z**
FPR	.04	.08	.01	.07	.10	.09	.01

Table 10. Fairness of RBF-SVM Model

6. Result Interpretation

Finally, in this section, we interpret the results of our models. We start with the results of the logistic regression with l_1 penalty. There were 36 features whose coefficient was 0 such as crimsusp_9 (Criminal Possession of a Forged Instrument), 15 (Driving while Intoxicated), 24 (Homicide), and CPI. Table 11 displayed the features with relatively large coefficients. No detailed explanation was reported since we applied a more advanced model, Random Forest, to do this work.

Feature	Coef	Feature	Coef
contrabn	3.107	pistol	2.986
knifcuti	2.233	otherweap	2.049
sb_other	1.880	addrpct_42	1.732
addrpct_40	1.541	searched	1.440
crimsusp_50	1.373	addrpct_19	1.294
sumissue	-3.145	addrpct_121	-1.423

Table 11. The Top Important Features from l_1 logistic regression

We utilized the scikit-learn Random Forest Library, which implements the **Gini Importance**. Gini Importance (also known as Mean Decrease in Impurity (MDI)) calculates each feature importance as the sum over the number of splits (across all tress) that include the feature, proportionally to the number of samples it splits [4].

Table 12 shows the top 5 important features, which are whether the suspect was searched, whether the suspect was searched based on other reasons, whether contraband was found on suspects, was the stop inside or outside and what time did the event happen. Among all the suspected crime types, Misdemeanor (32), Criminal Trespass (50), Felony (16), Criminal Possession of a Weapon (12) and Other (34) are the most relevant as shown in Table 13.

Feature	Importance
searched	0.123727
sb_other	0.093401
contrabn	0.05665
inout	0.038324
timestop	0.023485

Table 12. The Top 5 Important Features

Feature	Importance
crimsusp_32	0.012939
crimsusp_50	0.010781
crimsusp_16	0.009015
crimsusp_12	0.004794
crimsusp_34	0.002275

Table 13. The Top 5 Important Crime Types

The results from l_1-logistic regression and random forest agreed with each other. We could conclude that

[4] https://alexisperrier.com/datascience/2015/08/27/feature-importance-random-forests-gini-accuracy.html

among being stopped, searched and frisked, suspects being searched were most likely to be arrested, then was being frisked. Besides, suspects with weapons or contrabands were very likely to be arrested. At what time and in which place also mattered a lot. Stops inside would lead to arrestment with a higher probability. Among features of the economic situation, CPI_M, CPI on medical care, had a positive coefficient of 0.18, while S&P 500 had a coefficient of -0.04 and its square's coefficient was -0.07. This implies that suspects tended to be arrested in times of inflation, especially when people had to spend more on medical care while suspects would be less likely to be arrested in times of bullish market when people earned money from the stock market.

Suspects' race and sex were not determinant features, but we also found disparate treatment over sexes and races, which was discussed before.

7. Conclusion

In this report, we merged data from two sources, NYPD website and Wind Financial Terminal, to build classification models to predict police officers' arrestment decisions. We found that the logistic regression model trained on the balanced data set performed well in terms of F_1 (0.825) and AUC (0.915), but this model was unfair across sex and race.

We also tried Kernel SVM models, among which RBF was the best kernel with F_1 (0.861) and AUC (0.852). The non-linear model featured some advantages such as a higher F_1, but we cannot claim it was better because its AUC decreased sharply. In terms of fairness, it was unfair in the same way.

In the end, we interpreted our models combined with results of random forest. We found that suspects being searched were very likely to be arrested. Time and location also mattered greatly.

7 techniques were applied in this project: (1) outlier treatment, (2) feature engineering, (3) stratified sampling, (4) logistic regression, (5) support vector machine (SVM), (6) random forest, and (7) regularization.

The highlights of our work are (1) through improving our model, we obtain models with good performance in terms of metrics like F_1 and AUC, (2) we evaluated our models in terms of group fairness. Through modeling, we do observe the trade-off between accuracy and fairness. The unfair treatment among difference sexes and races, measured in terms of FPR, is the reason that we should be cautious to use the models to make decisions. The unfairness of our models may come from human bias existing in the training dataset due to historical reasons. When we "blindly" pursue accuracy, we also learn the bias embedded in the data set, and the model we build will reinforce such discrimination in one way or another. In the future, a fairer model can be expected via pre-processing, optimization at training time, and post-processing. [5]

References

[1] Pratik Gajane and Mykola Pechenizkiy. On Formalizing Fairness in Prediction with Machine Learning. *arXiv preprint arXiv:1710.03184*, 2017.

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[5] Section 5 of *A Tutorial on Fairness in Machine Learning* introduces these methods in detail.