ORIE 4741 Midterm Report

Binxin Liu (bl642) Mengjia Xia (mx233)

November 2019

1 Introduction

The goal of our project is to use Stop, Question and Frisk Data, which records every stop, question and frisk effected in NYC ranging from 01/01/2003 to 12/31/2018, to predict whether the suspect will be arrested based on information of every stop, question or frisk. For more details, please read our proposal.

2 Exploratory Data Analysis

2.1 Overview

Stop, Question and Frisk Data from New York City Police Department records every stop, question and frisk effected in NYC. For each record, it has 112 variables, including dates, times, locations, physical features of the suspect, crime suspected and so on. There are more than 5 million records during 16 years. As a preliminary analysis, here we only investigated the records from 2016 (12404 records), and we will extend the scope later.

2.2 Data Cleaning

Since the data was manually extracted from UF-250 Form Stop, Question and Frisk Report Worksheet (2016 Version), 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 10942 observations in total.

The variable "crimsusp" (crime suspected) is an important indicator of arrestment decision. However, the values are messy due to 1) different abbreviations used by different officials, 2) typo, 3) the use of Penal Law code instead [1]. 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.

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

2.3 Preliminary Data Analysis

Figure 1 shows the number of arrestment and non-arrestment across the time and its corresponding arrest rate. We can see that the arrest rate reaches its peak at 10:00-11:00 and 16:00-17:00. Though most stops happened at midnight, the arrest rate was not that high compared with daytime.

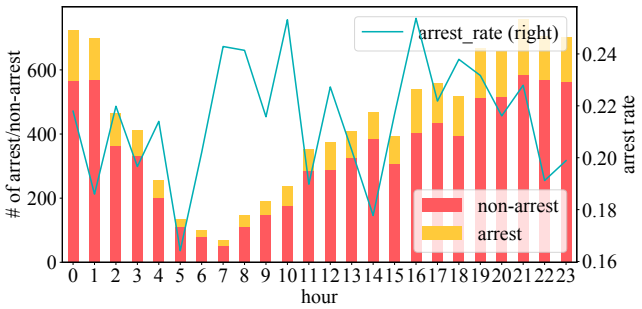

Figure 1: Arrest Across Time

Figure 2 shows the geographical distribution of these observations. Color turning purple to red means higher frequency of stops, questions or frisks while bigger red circles means higher arrest rate. As we can see, arrest rate varies a lot in different precincts.

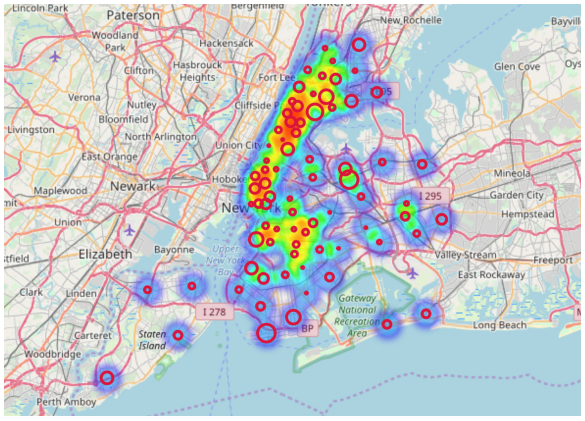

Figure 2: Arrest Across Location

We also investigated arrestments among different genders and races. From Table 1, the arrest rate of female is higher than that of male, but the number of female is far lass than male. Table 2 displays the arrest rate of different races. Though Hispanic had a higher arrest rate,

among those being arrested, Black people accounted for about 50% as shown in Figure 3.

Gender	Arrest Rate	Total
Male	0.212	9940
Female	0.244	598

Table 1: Arrest Across Gender

Label	Race	Arrest Rate	Total
I	American Indian/ Alaskan Native	0.118	34
A	Asian/ Pacific Islander	0.153	639
B	Black	0.202	5558
W	White	0.203	1086
P	Black-Hispanic	0.235	756
Q	White-Hispanic	0.257	2393
U	Other	0.264	72

Table 2: Arrest Across Race

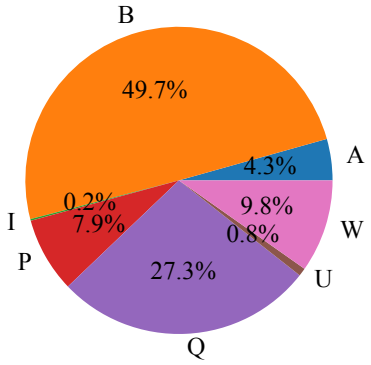

Figure 3: Arrested Suspects Across Race

Figure 4 shows the correlation between different variables. Here, we applied Pearson Correlation for pair of continuous variables and Cramer's V for pair of categorical variables. We can see that the correlations between variables are rather small, which indicates the need of feature selection and engineering.

3 Problem Formulation & Modelling

Prediction of arrestment decision is essentially figuring out what kind of stop, question or frisk and what kind of suspect will have a higher probability to be arrested.

For preliminary analyses, we selected 12 out of 112 variables as features (X) and "arstmade" as Y, then built a classification model. Please refer to Table 3 for an overview of variables we used.

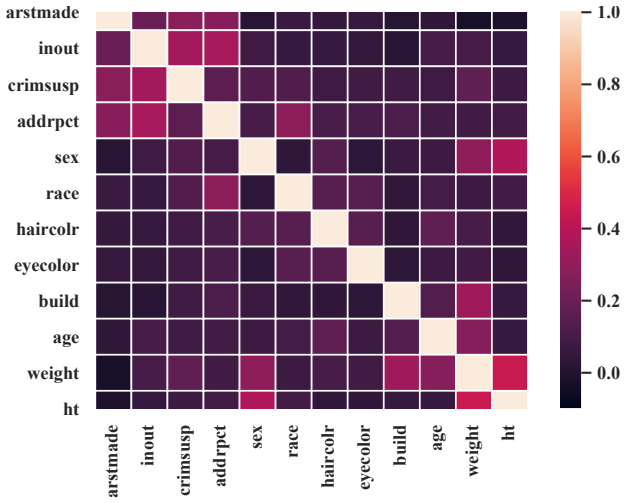

Figure 4: Variables Correlation Heatmap

Variable	Description	Type
arstmade	Was an arrest made ?	Categorical
timestop	Time of stop (hour)	Countinuous
crimsusp	Crime suspected	Categorical
inout	Was stop inside or outside ?	Categorical
sex	Suspect's sex	Categorical
race	Suspect's race	Categorical
age	Suspect's age	Continuous
ht	Suspect's height	Continuous
weight	Suspect's weight	Continuous
haircolr	Suspect's haircolor	Categorical
eyecolor	Suspect's eye color	Categorical
build	Suspect's build	Categorical
addrpct	Location of stop address precinct	Categorical

Table 3: Variables

3.1 Feature Engineering

To begin with, we performed feature engineering. Specifically, we used one-hot encoding on categorical variables (crimsusp, addrpct, sex, race, hair color, eye color, build). More sophisticated feature transformation will be used after we gain an elementary understanding of the prediction model.

3.2 Performance Metrics

Following standard performance metrics are used for our classification task: Accuracy & Precision & Recall (Confusion Matrix), AUC, and F_1-score. Note that AUC and F_1-score are overall metrics for performance of the classification model.

3.3 Modelling

Our first attempt was building a logistic classification model. We randomly chose 80% of the data set as the traing data set and the rest 20% as the test data set.

Firstly, we only used most basic features including suspect features, time and precinct. Figure 5 shows the ROC curve of the model ($model_1$). Then, we added inout, crimsusp into the features and introduced an l_2 penalty to the loss function to prevent over-fitting. Figure 6 shows the ROC curve of the new model ($model_2$). Table 4 displays the comparison between two models.

From Figure 5 and Table 4 we can see that though metrics such as accuracy, precision and recall are relatively high for $model_1$, AUC is low. This means $model_1$ cannot ideally predict the arrestment. Thus, we added two more variables, inout and crimsusp, as features because they are relatively highly correlated with the predicted variable (see Figure 4). However, one-hot encoding increases the number of features (more than 100). To prevent over-fitting and expect a better generalization of our model, we imposed the l_2 penalty on the loss function. From Figure 6 and Table 4 we can conclude that the modified model performs better in a big picture.

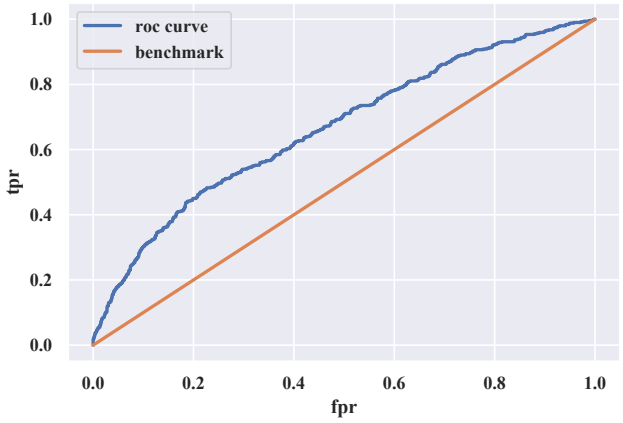

Figure 5: ROC Curve of $model_1$

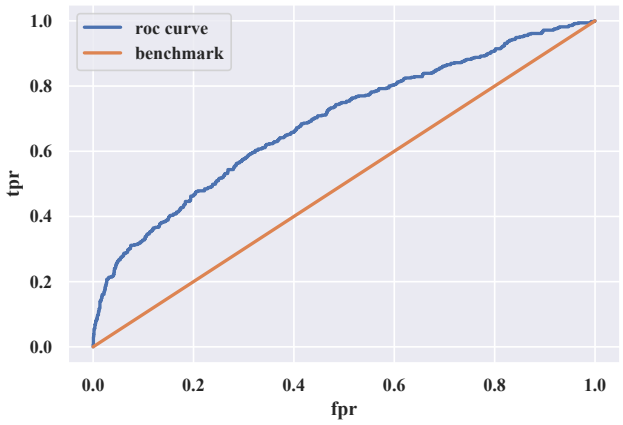

Figure 6: ROC Curve of $model_2$

Metric	$Model_1$	$Model_2$
Accuracy	77.98	79.58
Precision	99.35	98.23
Recall	78.17	80.00
F_1	0.875	0.882
AUC	0.665	0.690

Table 4: Performance of Two Models

4 Future Work

We implemented a simple logistic classification model. Our plan of future work is listed as follow.

4.1 Feature Selection & Engineering

We selected 12 out of 112 variables as features. Others have not been touched on. If we can use more relevant features, the performance can be improved. We will use Random Forests to inspect feature importance, and add more features. Meanwhile, we will explore how to engineer used features to make it more informative.

4.2 Balance Dataset

In our sample, there are about 21.38% suspects being arrested. This means that a naive classification model predicting all suspects will not be arrested can also have an accuracy of 80%. This is a common issue in the scenarios like credit card fraud detection. We will try to balance the data set by 1) adding the arrested samples from nearby years, or 2) using oversampling methods such as SMOTE [1].

4.3 Grid Search & Cross Validation

The hyper-parameters in the model will influence the performance of the model. We will use grid search and cross validation to determine the optimal parameters.

4.4 Model Generalization

Currently, we only used 2016 data to build a classification model. Considering we have a dataset covering 16 years, things can change greatly. We need to evaluate our model on different datasets to see if it can be generalized well. Otherwise, we need to modify our model to incorporate the changing trend.

References

[1] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.