DSC167 Final Paper

Amando Jimenez

June 2022

1 Introduction

For my decision making model, I decided to create a logistic regression model with the purpose of determining whether or not someone should be searched. I decided to use **subject age**, **time of day**, and **reason for stop** as my inputs and a binary variable indicating whether or not someone should be searched as my output variable. I decided to use these input variables as I am attempting to create a model that is as unbiased as possible so I only want the model to rely on variables that should theoretically dictate whether or not someone should be stopped and searched. I used age as an input variable because it has been shown that younger people are responsible for most of the crime which means that it can help determine whether or not someone should be searched [1]. I used time of day because it is more common for crimes to occur at night [2]. I used reason for stop as another input variable because it is extremely important to consider the severity of the initial crime/suspicion in order to make a just decision. For example, it would be completely unfair to search someone if they simply forgot to turn on their turn signal as opposed to searching someone who is driving erratically and is showing signs of driving under the influence. One important thing to note is that the reason why someone is stopped relies heavily on the police officer's interpretation and varies by police officer which could bias the results. I decided to omit the subject's race because as I described in my previous paper, police officers have been shown to unfairly target Black individuals in the past [3]. I also ommitted the location of the stop because it can be heavily correlated with race and as we have seen in the past, the police has oftentimes created feedback loops in which they continuously send more officers into black neighborhoods because they "discover" more crime in those neighborhoods and as a result they continue to send more officers to those areas [4]. All in all, with these variables I hope to create a decision making system that is fair and just.

2 Building a Decision Making System

In this section I will describe my logistic regression model in order to provide some context as to how it works and its intended use case. The model was built

using San Francisco police traffic stop data which was compiled by The Stanford Open Policing Project. The dataset was split into a training and test set which was used to evaluate the performance of the model.

The intended use of this model is to determine whether or not someone should be searched. The intended users of this model are police officers who are in the process of deciding whether or not to go through a person's belongings. For example, let's say a police officer decides to stop a young person who is speeding and driving erratically at midnight, the officer would then stop the car and would proceed to input the person's age, the time of day and the reason for the stop into the model and the model would then output a score suggesting whether or not the person should be searched in order to find evidence that they have committed a crime. Based on the score, police officers would then make the final decision as to whether or not to search them. An important caveat of the model is that it should not be used to determine if someone should be arrested as there are many other factors that should be considered first such as the person's criminal history. A limitation of the model is that it is operating under the assumption that police officers always follow protocol when stopping someone and never show any bias when stopping someone, which is unlikely as it has been seen numerous times in the past [3]. Another limitation of the model is that it is not able to take into account the subject's behavior which could be a good indicator of their actions.

In my analysis I will use false positive rate as a measure of fairness in order to ensure people across racial groups are being treated fairly by the model. I will use false positive parity because as I explained in my previous paper, being stopped and searched by the police can have numerous negative effects among minority communities, so it is extremely important to ensure that they are not being unfairly searched more often than everyone else [3].

In the future, the model should be tested on data from different cities across the US in order to get a better understanding of the model's performance. As previously mentioned, the data relied on information reported by the police which may be biased. Unfortunately, it is very hard to quantify bias which is why it will be extremely important to keep a close eye on the performance of the model on different racial groups in order to ensure bias is minimized.

3 Model Performance

3.1 Graphs

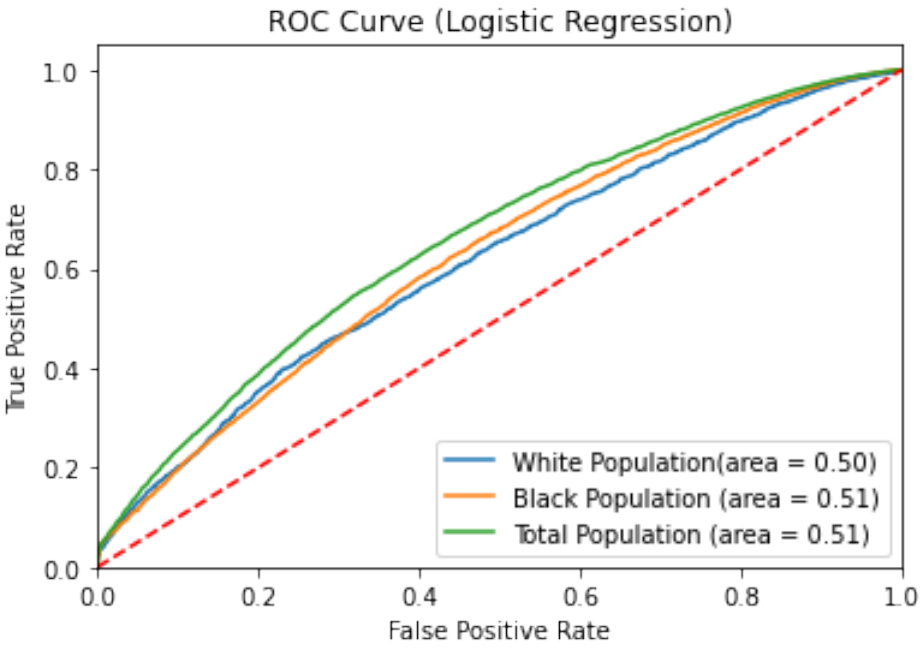
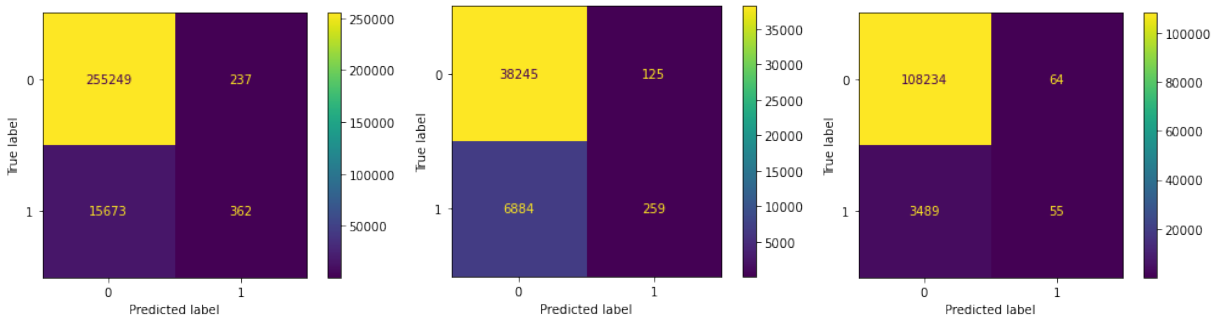


Figure 1: ROC Curve of Logistic Regression Model



(a) Total population (b) Black Population (c) White Population

Figure 2: Confusion Matrices of Logistic Regression Model

Rates	TPR	TNR	FPR	FNR
Total Population	0.0226	0.999	0.0009	0.977
Black Population	0.031	0.9979	0.003	0.967
White Population	0.009	0.9996	0.0004	0.99

Figure 3: Table Containing Rates of Different Populations

3.2 Results

As we can see by the figure 1 and figure 2a the results indicate that the model performed really well on the test data, as it only misclassified about 6% of the data. The high accuracy of the model is most likely a direct result of the data consisting of many more negative instances than positive instances, which essentially trains the model to predict a negative instance most of time and causes the true negative rate and false negative rate to be really high as can be seen in figure 3. Additionally, by observing figure 1 we can see that the model performed better than the red-dashed line which means that it was able to perform better than random. However, even though the results of the entire population are good, we must explore how it performs on the Black and White populations.

As we can see by figure 2c and 3, the white population had more data points in total and had a lower true positive rate than the black population which indicates that Black individuals were searched more often by police officers relative to their respective populations. Additionally the white population had a lower false positive rate and a higher false negative rate which indicates that the model was more likely to recommend against searching white individuals. By these results we can deduce that model essentially favored the white population as it gave them the benefit of the doubt and continually recommended against being searched which is probably why it had a larger false negative rate than the black population.

The observed outcomes, figure 2b and 2c, demonstrate that Black individuals were more likely to be searched than white people with about 18% of Black individuals being searched and less than 8% of white individuals being searched. This is most likely the reason why Black individuals experienced a higher false positive rate since the model learned that Black individuals were more likely to be searched by the police. This difference could be due to a number of reasons such as natural differences in crime rates among race groups, police bias and many other things. As it stands, the model does not satisfy luck egalitarianism as it does not ensure that the results solely rely on the individuals choices. In order to satisfy luck egalitarianism I have to find a way to level the playing field for black individuals who are more likely to live in poverty and have less education. By satisfying luck egalitarianism we will be able to maximize the utility of black and white individuals because instead of being unfairly searched and being traumatized as a result, they will be able to focus on other important tasks such as their education.

As I have explained previously, the false positive parity measure is what

should be considered in this case in order to ensure fairness. In this case, the results do not satisfy false positive parity since the black population is currently experiencing a higher false positive rate than the white population, which indicates the model is not fair.

4 Analysis

4.1 Pre-Processing: Label Flipping

4.1.1 Graphs

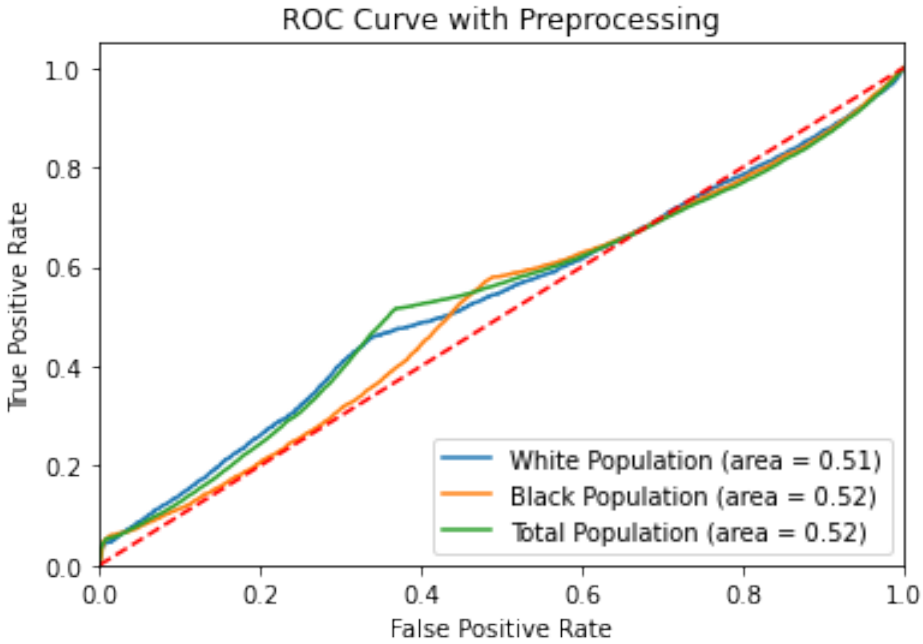
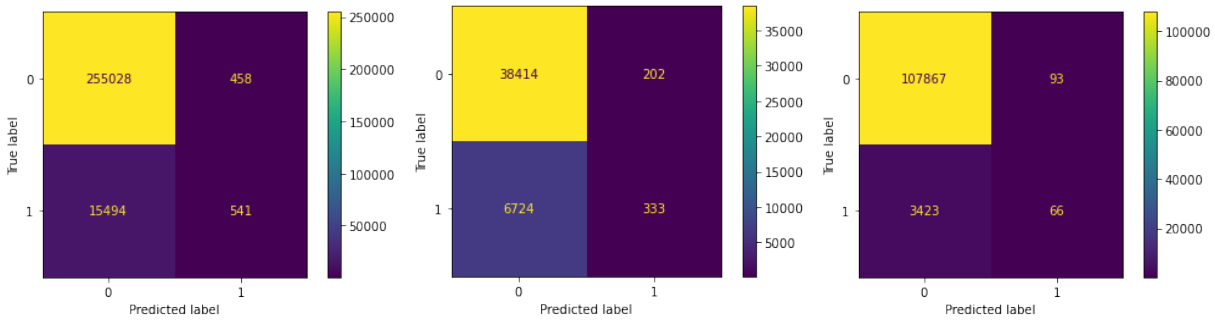


Figure 4: ROC Curve of Logistic Regression Model with Pre-processing



(d) Total population (e) Black Population (f) White Population

Figure 5: Confusion Matrices of Logistic Regression Model with Pre-Processing

Rates	TPR	TNR	FPR	FNR
Total Population	0.0226	0.999	0.0009	0.977
Black Population	0.031	0.9979	0.003	0.967
White Population	0.009	0.9996	0.0004	0.99

Figure 6: Table Containing Rates of Different Populations

4.1.2 Results

For this section of the paper I decided to do label-flipping pre-processing in order to de-correlate the subject's race with the search outcome label by following the procedure outlined in Building Classifiers with Independency Constraints[5]. As I have mentioned previously the data labels may correlated with race so in order to try to make the model fair I attempted to perform label-flipping. As we can see by figures 4, 5 and 6 the results did not change much. The biggest change can be seen in the true positive rate which makes sense since label-flipping essentially injected more positive instances. I hoped to equalize the false positive rates between the black and white population but I was unsuccessful, which means that I was not able to make the model more fair. I think the main reason for this result is that the majority of the data has a negative label in regard to search conducted which the model builds off of during training and causes it to perform better on the negative instances.
An interesting observation from figure 4 is that the ROC curve for the total population has a similar peak as the base model, however after it reaches a false positive rate of about 0.7 the true positive rate starts increasing at a decreasing rate which indicates that it performs worse than random. The black and white population ROC curves follow a similar pattern, but both have different peaks which means that two different models may be needed in order to ensure the trade-off between true positive rate and false positive rate is minimized and both groups are treated fairly.

4.2 Selective Labelling - Contraction

4.2.1 Graphs

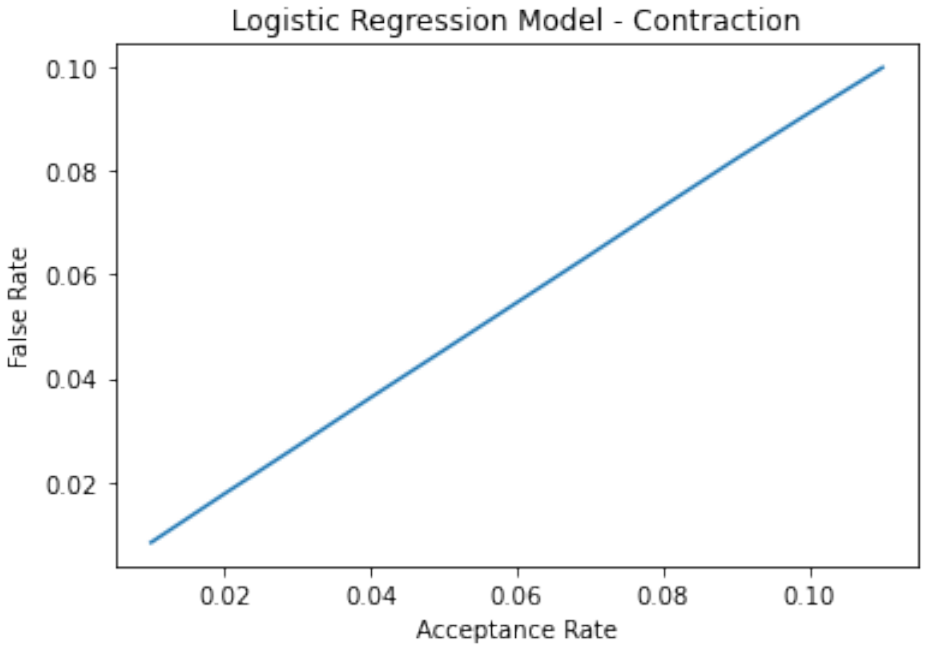


Figure 7: Model Evaluation with Contraction

4.2.2 Results

In this section of the paper I decided to evaluate the model using contraction. The purpose of using contraction is to evaluate models trained with selective labels by only using available data which in this case means data of people that were searched [6]. Since I did not have any information on the decision makers i.e. police officers I decided to use districts as a proxy for police officers since police officers usually patrol the same districts. For the purposes of this analysis I considered whether or not contraband was found as the true outcome since a search is typically done in order to find contraband.

In order to perform contraction on the model I followed the algorithm found in The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables [6]. I found that district C had the highest acceptance rate at 0.11, which means that 11% of people were searched in district C. After performing contraction on the dataset composed of subjects in district C I found that the model had a failure rate of 0.10 which means that 10% of people that it predicted should be searched should not have been.

Additionally as we can see in figure 7 there seems to be a linear relationship between acceptance rate and failure rate in the model, which means that as

we increase acceptance rate the failure rate will also increase. If this trend continues more and more people will be unfairly searched as the acceptance rate increases. All in all, I cannot determine what acceptance rate is fair by simply looking at this trend, further analysis must be done in order to determine an acceptable acceptance and failure rate that is fair for every group involved while also ensuring utility is maximized and the right people are being searched.

5 References

1. "Comparing Offending by Adults amp; Juveniles." Office of Juvenile Justice and Delinquency Prevention, https://www.ojjdp.gov/ojstatbb/offenders/qa03401.asp.
2. "Law Enforcement and Juvenile Crime." Office of Juvenile Justice and Delinquency Prevention, https://www.ojjdp.gov/ojstatbb/crime/ucr.asp?table_in=1amp;selYrs=2019amp;rdoGroups=1amp;rdoData=rp.
3. Jimenez, Amando. "DSC167 Paper 1" May 2022. DSC 167, University of California San Diego.
4. Fraenkel, Aaron. "Lecture 16: Feedback Loops" DSC 167, 19 5 2022, University of California San Diego. Lecture.
5. Calders, Toon, et al. "Building Classifiers with Independency Constraints." 2009 IEEE International Conference on Data Mining Workshops, 2009. 6. Lakkaraju, Himabindu, et al. "The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables." 2017, KDD 2017.