**CSE574 Programming Assignment 3 Report**

**Role in the Project: Machine Learning Engineers (Volunteer for NGO)**

May 8, 2020

**Group Information**

Group Number: 29

Group Members:

Xiyun Xie (51062104)

Mollie Wugetemole (50165532)

Yutong Yang (50321338)

**Proposed Model Information(Detailed result in last page)**

Model Choice: Neural Network

Algorithm Choice: Maximum profit

Secondary Optimization Criteria: Accuracy

Overall Accuracy:

Training data: 0.6410850915835845

Test data: 0.6576869484440316

Whole data: 0.6444011503850079

Overall Cost:

Training data: $-593,855,222

Test data: $-140,929,366

Whole data: $-734,784,588

This project is motivated by the potential fairness violation issues presented in COMPAS. Our goal is to propose a new model that maintains a reasonable level of prediction accuracy, while, at the same time, properly apply the concept of fairness.

The stakeholders in this situation includes: 1) the defendants, as the decision made by the model will directly affect their future life; 2) the judges, as the model can either lead them into right decision or wrong decision; 3) the government, as the model will determine potential financial cost and ethic reputy; 4) The whole society, as discrimination of different races may cause many reciprocating behavior. A good model will not only make appropriate decisions, but also has to be persuasive to avoid ethnic conflict.

According to the evaluation of ProPublica, COMPAS shows apparent discrimination among different racial groups. Black defendants are more likely to receive longer sentences if their actual criminal magnitude is close to other races. This can be caused by the inherent bias in the raw data and viewed by the ROC curve. When we draw a horizontal line to demonstrate true positive values are the same for curves, African-American’s false positive value is the highest. Besides the Neural-Network model, this phenomenon happens on other models as well.

Some of the biases are inevitable because they come from many other invisible factors that might be related to the sensitive category in our data even a little bit, which we can’t control. But we can adjust the model and algorithm to make relatively fair decisions on the unfair data. To do this, we explore 3 types of prediction model, namely SVM, Naïve Bayes and Neural Network; and we apply 5 different post processing methods, Maximum Accuracy, Single Threshold, Predictive Parity, Demographic Parity and Equal opportunity. While trying to perform the algorithms, we found that algorithms might also have bias. For example, African-American is a group of people whose true positive rate is potentially higher than other races, equal opportunity algorithm will decrease the thresholds for some races to make more people predicted positive so that overall true positive rate will be close, which will virtually make more innocents predicted as potential recidivism. If we perform equal opportunity, then our algorithm will try to protect races like African-Americans whose positive prediction is high. Therefore, for each of the algorithms, the criteria we are using to choose our model is mainly accuracy, but at the same time with persuasive and proper fairness.

Our solution can also appeal to ordinarys’ review of some incorrect political correctness that some races whose behaviors are notorious but blindly asking for more opportunities and forgiveness. However, we can neither stop suspends who have low prediction scores from recidivating, nor stop suspends who have high scores from engaging in charity after they are discharged from prison once.

We believe that our solution is a better choice because we interpret fairness as not doing things that are unfair instead of forcing everything fair. This can be referred to the equal opportunity we mentioned. According to Counterfactual Fairness, “Depending on the relationship between a protected attribute and the data, certain definitions of fairness can actually increase discrimination”. One disparity is False Negative rate of race of Other. This disparity means that many Other race people whose predictions are low but they will recidivate. This phenomenon might be misleading because compared to recidivism of low scoring people, much more Other race people whose scores were high did not recidivate, which causes the threshold to be higher than other races’ thresholds. Since we are using maximum profit(accuracy), we want the accuracies of each race to be the highest. In our model’s result, most of the thresholds are close to each other, which means we are applying relatively equal treatment of those races. The ROC graph shows that our Neural Network model works because all curves are above the True-positive equals False-Negative line. Again, even our data and model are very detailed and complicated, they still cannot tell us human nature precisely, nor predict if people will amend. So we are trying to understand the situations on those disparities and let our model and algorithm reflect the truth of the data only by choosing the best accuracies.

**Reference**

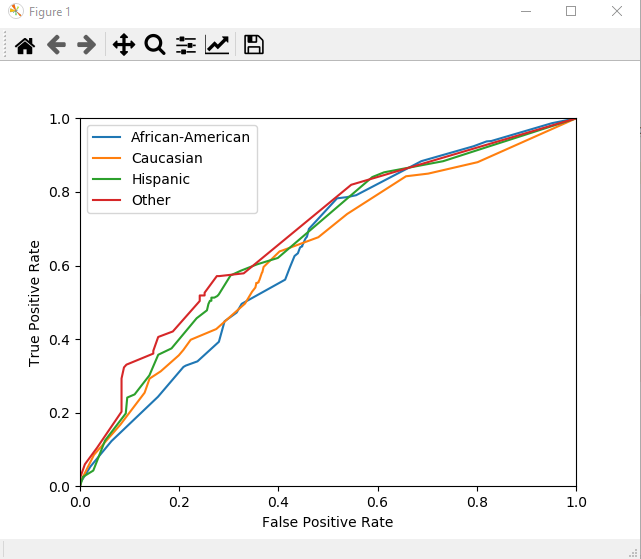
[1] Primier

[1] ProPublica – Machine Bias

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

[2] Model Comparison

[3] ROC



[4] (Counterfactual fairness ,Matt J. Kusner and Joshua R. Loftus and Chris Russell and Ricardo Silva,

2017,arXiv:1703.06856v1)

Market Model results:

