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 persuasive fairness.

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[1]. White defendants are more likely to be assigned a lower risk score despite the fact that some of them are actually within the high risk of recidivation [1]. These kinds of biases can be caused by many reasons such as original data and machine learning algorithms [2]. This can be caused by the inherent bias in the raw data and viewed by the ROC graph[3]. 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 we use, this phenomenon happens on other models as well.

These issues should be taken seriously because it can affect many aspects of our society. 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 reputation; 4) the ethnic minority groups, namely African-American, Hispanic, Asian, Native American and others, as they are the people whose life will be influenced by machine learning algorithms; 5) The whole society, as discrimination of different races may cause many retaliatory behaviors. With these things taken into consideration, we need to develop a good model which not only help us to make accurate decisions, but also have everyone being treated in a fair manner, and hence avoid causing discrimination and further ethnic conflicts.

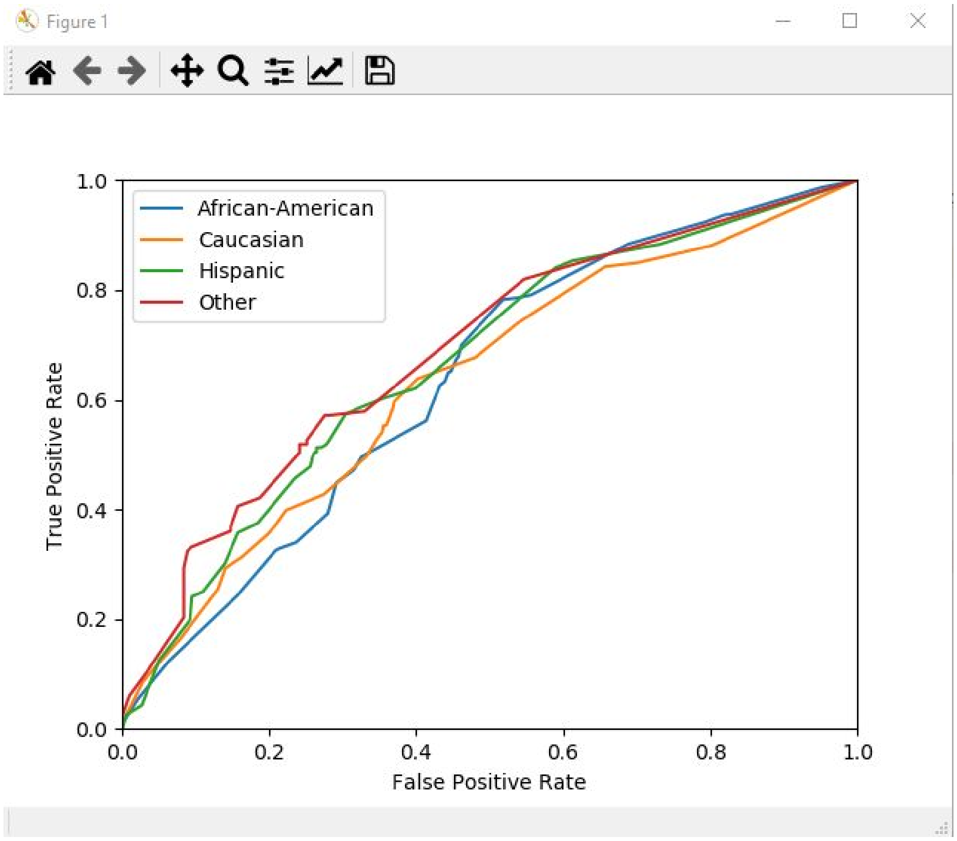
Some of the biases are inevitable because they come from many other invisible factors that might be related to the sensitive category in our data even a little bit, which we can't control. But we can adjust the model and algorithm to make relatively fair decisions. To do this, we explore 3 types of prediction 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 rate of positive label is higher than other races. Equal opportunity algorithm will potentially decrease the thresholds for some other races to make more people of those races 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 rate of positive label is high. Therefore, the algorithm we choose is maximum profit(accuracy), and the secondary optimization we are using in our model is accuracy, but at the same time with persuasive 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 to become fair. For example we don't want to choose some fairness methods like equal opportunity to make things look fair but will actually hurt more people. According to Counterfactual Fairness, "Depending on the relationship between a protected attribute and the data, certain definitions of fairness can actually increase discrimination"[4]. One disparity in our model is that the False Negative rate of race of Other is much higher than other races. 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, many Other race people whose scores were high did not recidivate, which causes the threshold to be higher than other races' thresholds. This can also show that our model and algorithm can show some fundamental information of data. 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 treating those races in a fair manner. The ROC graph shows that our Neural Network model works because all curves are above the True-positive equals False-Negative line. In conclusion, even our data and model are very detailed and complicated, they still cannot tell us a person's human nature precisely, nor predict if he or she 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.

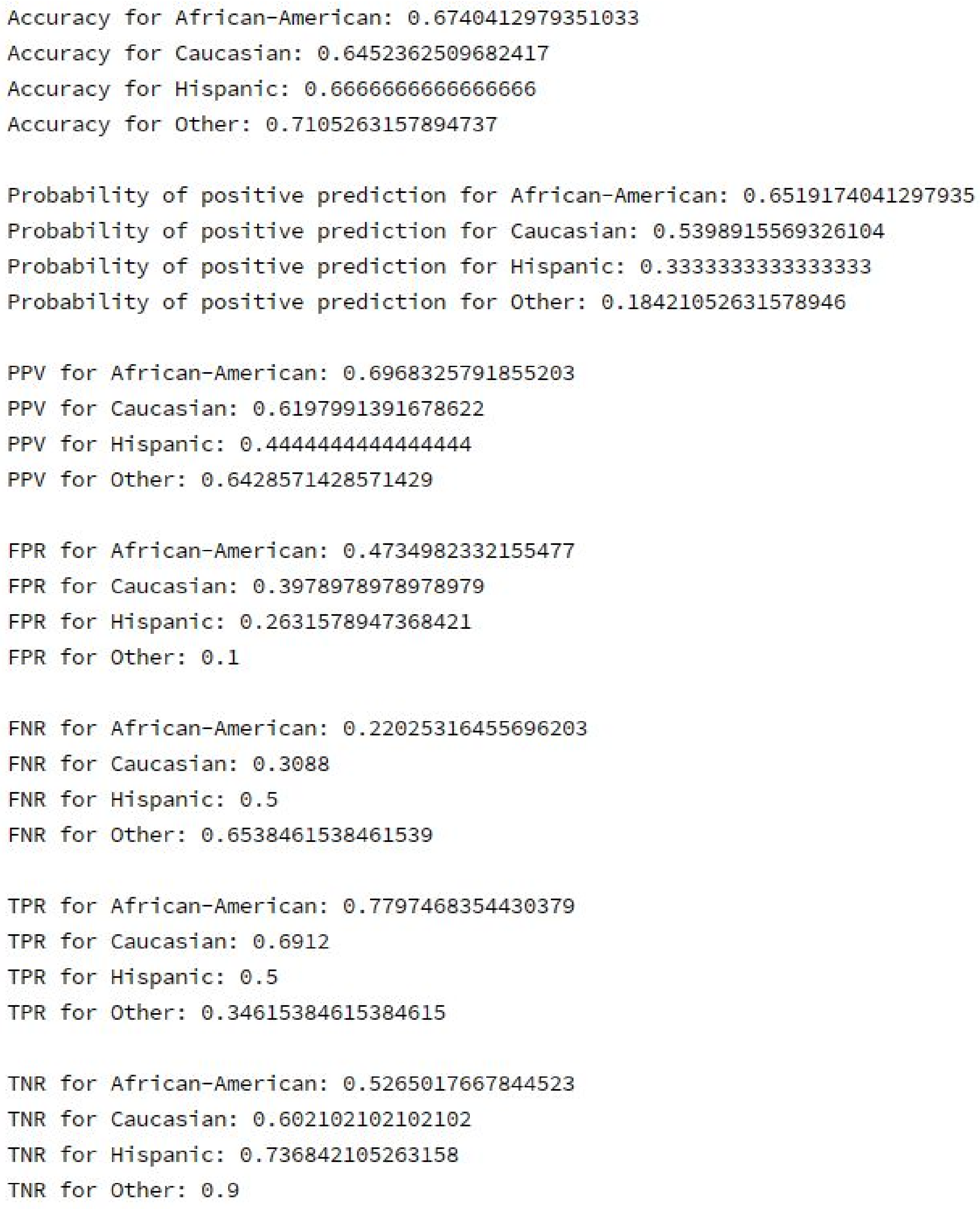
Reference

[1] ProPublica – Machine Bias
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
[2] CSE574 Course Material (fairness-ml-handout.pdf)
[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



```
Accuracy for African-American: 0.6740412979351033
Accuracy for Caucasian: 0.6452362509682417
Accuracy for Hispanic: 0.6666666666666666
Accuracy for Other: 0.7105263157894737

Probability of positive prediction for African-American: 0.6519174041297935
Probability of positive prediction for Caucasian: 0.5398915569326104
Probability of positive prediction for Hispanic: 0.3333333333333333
Probability of positive prediction for Other: 0.18421052631578946

PPV for African-American: 0.6968325791855203
PPV for Caucasian: 0.6197991391678622
PPV for Hispanic: 0.4444444444444444
PPV for Other: 0.6428571428571429

FPR for African-American: 0.4734982332155477
FPR for Caucasian: 0.3978978978978979
FPR for Hispanic: 0.2631578947368421
FPR for Other: 0.1

FNR for African-American: 0.22025316455696203
FNR for Caucasian: 0.3088
FNR for Hispanic: 0.5
FNR for Other: 0.6538461538461539

TPR for African-American: 0.7797468354430379
TPR for Caucasian: 0.6912
TPR for Hispanic: 0.5
TPR for Other: 0.34615384615384615

TNR for African-American: 0.5265017667844523
TNR for Caucasian: 0.602102102102102
TNR for Hispanic: 0.736842105263158
TNR for Other: 0.9
```

```
Threshold for African-American: 0.43214999999996057
Threshold for Caucasian: 0.4328499999999605
Threshold for Hispanic: 0.44764999999995886
Threshold for Other: 0.6214499999999398

Score for African-American: 0.7359617682198326
Score for Caucasian: 0.653555219364599
Score for Hispanic: 0.47058823529411764
Score for Other: 0.45

Accuracy on training data:
0.6410850915835845


Accuracy on test data:
0.6576869484440316

Accuracy on whole data:
0.6444011503850079

Cost on training data:
$-593,855,222


Cost on test data:
$-140,929,366


Cost on all data:
$-734,784,588
```