CSE574 Assignment 3

PA Group 11
Deepti Bharadwaj(50363309)
Kaumudi Moholkar (50388592)
Alexander Ma (37136560)

May 5th, 2021

REPORT

Model : Naive-Bayes
Post-Processing Method : Equal Opportunity Results
Secondary Optimization : Accuracy
Total cost: $-760,047,138
Total accuracy: 62.72%

1. What is the motivation for creating a new model to replace COMPAS? What problem are you trying to address?

Any ML algorithm makes a model based on the patterns and numerical encodings present in the data. They do not look into the biases and inherent issues existing in them. Our inspiration as volunteers of a NGO is to kill those intrinsic inclinations by making a model that examines and eradicates these inadequacies. The COMPAS is biased towards certain groups especially the African-American. It did not include race as a factor in the model. The white defendants were classified as the lower risk as compared to the black defendants.

We are focused on making the model as fair as possible by making the predictions such that they represent all groups equally. This is done by setting the appropriate threshold values such that it satisfies the requirement. The secondary optimization parameter is set as accuracy as we are an NGO and we strive to remove the bias from our model.

2. Who are the stakeholders in this situation?

As mentioned in the beginning of the report our primary concern is to ensure equality amongst all communities, so our primary stakeholder are the defendants, their families, and the communities where this software is implemented.

Depending on the accuracy of our software, the suspect and families lives can be impacted. If our model produces a lot of false positives many of the defendants would have longer sentences. On the other hand a higher false negative would result in offenders being back in the streets early causing a higher crime rate in the community.

Another stakeholder is the local and state government looking to adopt our software. They are the one investing into this project, so we would need to make sure the software can minimize their financial loss and bring confidence to their governing body for selecting our software as being a fair model.

3. What biases might exist in this situation? Are there biases present in the data? Are there biases present in the algorithms?

We can see from demographic data that there are racial biases in the model. There is disparity in the people classified as recidivistic or not recidivistic. The black defendants have more people classified as recidivistic and the larger number of white counterparts were classified as non-recidivistic. Since the data is skewed towards certain races differently for groups the predictions were biased towards the same result.

Similarity in demographic characteristics might be due to the fact they are common characteristics of one race. This proves a strong case for why race as a metric should be examined more thoroughly. Since certain races might have factors like joblessness and poverty that are more dominant than amongst other races.By not factoring in race in its initial design ,the biases in the data have seeped into the algorithm and resulted in inaccurate scores.[4][5]

Another possible bias is that we are bucketing races into four categories. Anything outside of African-America, Caucasian and Hispanics races are bucketed into Other. Which can lead to unfair representation for the different races that fall under the Other category.

4. What is the impact of your proposed solution?

Equal Opportunity states that each group should get positive outcomes at equal rates. Hence our model aims to solve issues in the equal opportunity model. Since we are volunteers of a humanitarian NGO, we would like to work on a model that is sensitive to different groups, giving them an equal chance to be labeled as recidivistic or not recidivistic.

Equal opportunity strives to create a good classifier that is sensitive towards the groups. This helps in eradicating any type of biases in the data and our model can be as fair as possible.This can result in an overall lower accuracy but helps in giving a fair model. It provides incentive to reduce errors uniformly in all groups.

5. Why do you believe that your proposed solution is a better choice than the alternatives? Are there any metrics (TPR, FPR, PPV, etc?) where your model shows significant disparity across racial lines? How do you justify this?

Our proposed solution is better than other alternatives as our main focus is on predicting the positive outcomes correctly and providing equal chances for people to be labeled properly. In this model we obtain the True Positive Rate(TPR) which has a difference of 0.02 at the maximum for different races of people. Since there is no major disparity in the accuracies we emphasize on focusing on the TPR over other metrics. The disparity between the FPR amidst the different racial groups is also significantly lesser which implies the proposed model is more accurate.

TPR for African-American: 0.6964755391899
TPR for Caucasian: 0.7025530605967394
TPR for Hispanic: 0.6939655172413793
TPR for Other: 0.6992481203007519

Extra Credits Report

How do you justify valuing one metric over the other as constituting “fairness”?

Equal opportunity gives an equal chance of labeling people as recidivistic. It gives an equal representation over the privileged outcomes which by far is the one means to provide fairness. The demographic parity strives to get an equal proportion of people labeled as recidivistic, which can result in wrong people being labeled as recidivistic. Maximum accuracy focuses majorly on increasing the accuracy without considering other necessary factors and can lead to wrongful representation of groups. As single threshold considers only one threshold value for all the groups, it does not seem to be fair based on our role and requirements. Equal opportunity provides the fairness we are looking for in building a software which will give better predictions.

What assumptions are made in the way we have presented the assignment? Are certain answers presupposed by the way we have phrased the questions?

An assumption made about this assignment is that there are five methods of fairness that can be implemented to improve the COMPAS model but there are several more like equalized odds, individual fairness that can be explored. Equalized odds is a stronger version of equal opportunity, which considers minimizing the false positive rate, which can be an important factor considering our secondary optimization is ‘accuracy’ [1].

As mentioned in an article from Cortez [2], Equalised Odds are preferred when there is a “strong emphasis on predicting the positive outcome correctly” as well as “strongly care about minimising costly False Positives”. False Positives should be heavily focused in our scenario as we do not want to wrongly increase the sentence/bail of the defendants.

We calculated Equalised Odds to see the difference between the model we selected (Equal Opportunity) and saw a big drop in False Positive Rate. From 60% to under 30%. The only drawback for this fairness method is that the True Positive Rate also drops from 83% to 40~60%. We think the tradeoff is fair, since having a higher false positive will negatively affect a person’s life.

We’ve implemented the Equalised odds methods based on the rule :

$$P(\hat{Y} = 1 \mid A = 0,\ Y = y) \ = \ P(\hat{Y} \ = 1 \mid A = 1,\ Y = y),\ y \in \{0,1\}$$

EQUAL OPPORTUNITY RESULTS	EQUALISED ODDS RESULTS
Accuracy for African-American: 0.6365521327014217	Accuracy for African-American: 0.5364336492890995
Accuracy for Caucasian: 0.6091971701015072	Accuracy for Caucasian: 0.5932020916641033
Accuracy for Hispanic: 0.5488454706927176	Accuracy for Hispanic: 0.6465364120781527
Accuracy for Other: 0.5443786982248521	Accuracy for Other: 0.650887573964497
FPR for African-American: 0.6250847457627119	FPR for African-American: 0.27322033898305087
FPR for Caucasian: 0.6124269455552138	FPR for Caucasian: 0.27529990772070134
FPR for Hispanic: 0.649546827794562	FPR for Hispanic: 0.2930513595166163
FPR for Other: 0.6439024390243903	FPR for Other: 0.2975609756097561
TPR for African-American: 0.8395581273014203	TPR for African-American: 0.3887427669647554
TPR for Caucasian: 0.8308212857582282	TPR for Caucasian: 0.4617040910489081
TPR for Hispanic: 0.8318965517241379	TPR for Hispanic: 0.5603448275862069
TPR for Other: 0.8345864661654135	TPR for Other: 0.5714285714285714
Total cost:	Total cost:
$-774,369,540	$-840,699,054
Total accuracy: 0.6125800166991372	Total accuracy: 0.5800166991372112

Table comparing Equal Opportunity Result vs Equalised Odds Result

In what ways do these simplifications not accurately reflect the real world?

The simplification from assumptions made in this assignment provides a finite list of metrics.In the real world certain states in the US use different decision making models that take into account static and dynamic factors that can improve the model. These factors come from different criminological approaches and could help create a more nuanced understanding of risk scoring.[6]

To what extent should base rates of criminality / recidivism among different groups be factored into your decision?

The rates of criminality/recidivism should be taken as an insight and not a final drawn conclusion for individuals. The aim of this model is to be a tool and not to replace judicial decision making. This model is binded by only a few particular metrics when there could be a better solution for the model we have made.Since there is always room for improvement,with more studies,additional metrics (possibly more than five) could replace the current model.

The tools we provide can split the predictions into different protected categories, such as by age or gender. What disparities arise in these groups? How do these disparities compare to those shown when the predictions are split by race?

If age is used as a protected category then it could cause disparities in the way individuals are scored. A first time offender should not get a worse score in comparison to a repeat offender based on how young or how old they are. When it comes to gender there are studies that discuss the nature of gender bias in algorithms like the COMPAS model. Keeping this in mind,it has been found that women are overclassified in higher risk groupings. Which goes to show it is important to question whether gender should be included as a protected category or not because this may play into biases that arise from societal gender norms. [7]

References

1. Colyer, A. (2018, May 7). *Equality of opportunity in supervised learning | the morning paper*. Blog. https://blog.acolyer.org/2018/05/07/equality-of-opportunity-in-supervised-learning/
2. Cortez, V. (2020, June 8). *How to define fairness to detect and prevent discriminatory outcomes in Machine Learning*. Medium. https://towardsdatascience.com/how-to-define-fairness-to-detect-and-prevent-discriminatory-outcomes-in-machine-learning-ef23fd408ef2
3. Hao, K. (2020, April 2). *AI is sending people to jail—and getting it wrong*. MIT Technology Review. https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/
4. ProPublica. (2020a, February 29). *How We Analyzed the COMPAS Recidivism Algorithm*. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
5. ProPublica. (2020b, February 29). *Machine Bias*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
6. Jackson, E. (2020, March 31). *Setting the Record Straight: What the COMPAS Core Risk and Need Assessment Is and Is Not · Issue 2.1, Winter 2020*. Harvard Data Science Review. https://hdsr.mitpress.mit.edu/pub/hzwo7ax4/release/4
7. Hamilton, M. (2019, March 1). *The sexist algorithm*. Wiley Online Library. https://onlinelibrary.wiley.com/doi/abs/10.1002/bsl.2406