CSE 574 Intro to Machine Learning, PA3, Group No. 36

Names: Sreya Dhar, Vaishnavi Vukku, Sai Lakshmi Navya Maddu

1. What is the motivation for creating a new model to replace COMPAS? What problem are you trying to address?
Ans: From the Broward county dataset, it has been quite evident that the COMPAS system created by Northpointe is biased towards a certain race, i.e., African-Americans, as pointed out by ProPublica. Being at corporate, our goal is to make a fair model that would be financially more rewarding than the COMPAS model and comprise societal ethics. Five different statistical metrics have been implemented to evaluate the model to its best fairness form. Individual races have been assigned with the appropriate threshold for each metric that has been decided from the line search method. Moreover, being associated with a publicly traded corporation, apart from ethical consideration, cost optimization has become our secondary objective.

2. Who are the stakeholders in this situation?
Ans: The people who will directly benefit from our model would be the ***defendants***, who have already been prosecuted in criminal trials, given they are the subject of prediction (/suspect) of our models. It would be easier for the Department of Justice, including ***Judge and Juris,*** to decide on the defendant. ***Northpointe*** could be benefitting in a better way of knowing how to alleviate racial biases in their model. ***Other corporates*** in the business could follow and analyze their predictive model based on a similar approach. Again, our model aims to optimize cost expenditure for a prisoner, which immensely benefits society, as lots of ***tax payers'*** money is at stake.

3. What biases might exist in this situation? Are there biases present in the data? Are there biases present in the algorithms?
Ans: As pointed out by ProPublica, racial biases exist in the Broward county dataset. That means more #of defendants are 'African-Americans' than from any other communities, *which implies there are biases present in the data.* It has been already claimed that the COMPAS model does not consider 'race' as a feature for prediction and fails to function somewhat on the unbalanced dataset. This is an example of *Sampling bias* [1]. It's pretty surprising that out of 137 features, only seven have represented the people in the study [2]. It instates the problem that *there is bias present in their algorithm*, and the COMPAS model performs as simple as logistic regression during decision making. Thus, when the dataset is skewed towards a specific race like 'African-Americans', the COMPAS model underestimates the probability of getting recidivated of Caucasian twice. The black defendants who did not recidivate are rated as high-risk twice the rate of comparable white defendants. Some trials on the risk assessment frameworks should be carried out before releasing the tools for public usage. Only four significant ethnicities have been considered for our analysis, and the numerical count for each group has been given below. Apart from race, gender and age groups have also been shown as examples of population biases.

Race – African-American: 3696; Caucasian: 2454; Hispanic: 637; Others: 377
Age – 25-45: 4109; Greater than 45: 1576; Less than 25: 1528
Gender – Male: 5819; Female: 1395

Racial and population bias has been present in the dataset as all the groups do not exist in an equal ratio in the Broward dataset.

4. What is the impact of your proposed solution?
Ans: Our proposed solution is the following:

Proposed Model: Support Vector Machine
Proposed post-processing (PP) techniques: Demographic Parity
Secondary Optimization metric: financial cost (corresponding *total cost*: $-757,433,340 and *total accuracy*: 62.79%)

The primary aim of being in corporate business is to maximize profit and winning more developing contracts for the company encompassing societal ethics. After comparing three algorithms (Naïve-Bayes, NN, SVM), we found out that the financial cost involved for the SVM model is the least for the Demographic Parity PP technique. In Demographic Parity, thresholds for each racial group have been achieved to have equal predicted positive rates. This PP technique indulges a positive prediction be independent of the racial bias [3]. Thus, it will enforce fairness awareness in our ML model for recidivism prediction across ethnicities in society. Societal impacts like any race choosing any threshold that is too high or too low can worsen any disparity towards any protected group. This problem can be overcome by considering demographic parity as the proposed solution. A tolerance of 2% has been assigned for PPR to any compromised group. It tends to create a fair classifier by reducing variance, given ground truth for the protected groups is equal. From the business perspective, it delivers higher TPR and FNR for race 'African-Americans' than any other techniques; thus, it is very applicable to overcome the drawbacks that the COMPAS model had.

5. Why do you believe that your proposed solution 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?

Ans: From the business perspective, Demographic Parity delivers higher TPR and FNR for race 'African-Americans', which makes the model fair against racial biases. It helps to achieve the flaws that the COMPAS model had. Racial discrimination comes under conditional demographic disparity and which is what we are trying to solve in this assignment. For additional support to our proposed solution, we have considered cost optimization from the corporate angle. By combining both fairness and profit, we believe that the demographic parity maintains the desired balance in the model we want to achieve for our company. If we compare metrics like TPR and TNR, hovering over the ranges of [71-75.5%] and [48-52.5%], respectively. It is worth mentioning that demographic parity maintains a subtle balance among the accuracy metrics [59.5-63.4%] for different racial classes, which helps to alleviate the problem alleged by ProPublica.

Maximum profit is the best cost optimization technique but ignoring any racial disparity in the dataset. A single threshold has the same margins for all the groups, and no additional threshold has been assigned for racially protected groups. Equal opportunity predicts that all the groups would have equal TPR and could not always replicate the real-world scenario. Moreover, Predictive Parity does not compromise accuracy, though it ignores all the statistical metrics like TPR, FPR, TNR, FNR. Thus, considering predictive parity would make our model statistically more vulnerable. So, our proposed solution would address the problems mentioned above to reduce racial bias in our defined model.

Report Extra Credit:

1.How do you justify valuing one metric over the other as constituting "fairness"?
Ans: We have compared metrics like TPR and TNR, which are hovering over the ranges of [71-75.5%] and [48-52.5%], respectively. It is worth mentioning that demographic parity maintains a subtle balance among the accuracy metrics [59.7-61.7%] for different racial classes, which helps alleviate the problem alleged by ProPublica. TPR and FPR for African-Americans are 49% and 28.5%, respectively. These metrics are comparatively lower than any other post-processing techniques. In contrast, TPR and FPR for Caucasians are 47.5% and 25.8%, respectively. That means the probability of wrongful eviction of African-Americans has reduced in our model, substantiating fairness among classes.

2.What assumptions are made in the way we have presented the assignment? Are certain answers presupposed by the way we have phrased the questions?
Ans: As directed in the assignment, we need to think from a business perspective to get a fair balance between profitable income as well as societal fairness, which addresses different biases like racial and population biases. So, from the business point of view, we can't completely ignore the parameter of maximum turnover by the end of the year. Thus, pondering all these profitable factors, we have assumed minimizing cost would be our secondary optimizing parameter. We felt that a certain answer is presupposed and solely depends on which team you are associated with, either corporate or NGO—saying that there is no predefined answer for the proposed solution. It takes specific skills to balance the fine line between profit and accuracy, which comes from years of practice in the domain of fairness in ML.

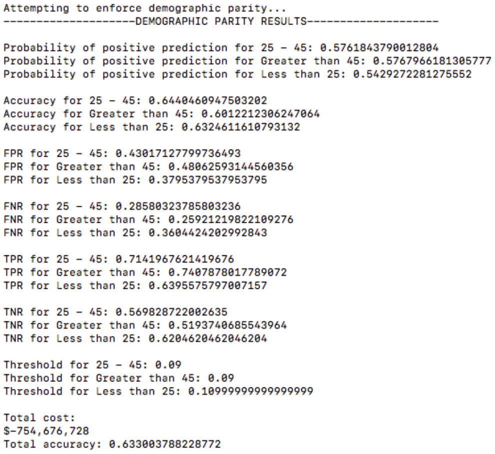
3.In what ways do these simplifications not accurately reflect the real world?
Ans: We acknowledge that the simplifications are for better prediction, but it would not always replicate real-world scenarios. To be more accurate to assumptions, we need more features, samples, and a balanced dataset demographically. More communities or organizations should be willing to contribute their parts so that fairness prediction would not get limited to the State of Florida, instead get explored in other parts of the country too.

4.How do uncertainty and risk tolerance factor into your decision?
Ans: The uncertainty and risk factors are inversely proportional to cost optimization and profit. The higher the risk for a model, the lesser the profit/gain from the prediction.

5.To what extent should base rates of criminality/recidivism among different groups be factored into your decision?
Ans: To a great extent, the base rates of criminality/recidivism among different groups be factored into our proposed solution. The statistical metrics that we achieved from the demographic parity of the SVM model are within a reasonable range so that none of the groups would get penalized wrongfully from the prediction model. Nevertheless, our proposed solution outperforms the COMPAS model, and the imperfection in the COMPAS model has been addressed efficiently in our prediction framework.

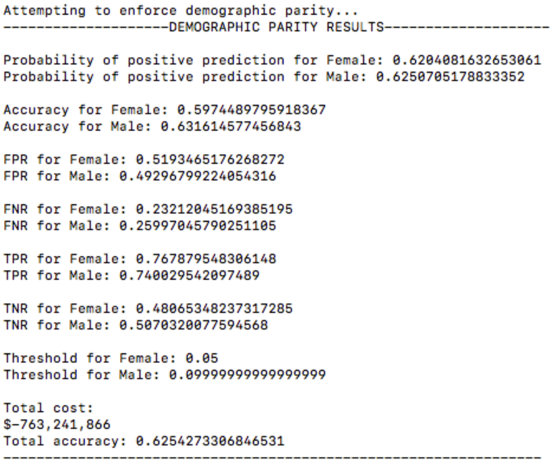
6.The tools we provide can split the predictions into different protected categories, such as by age or gender. What disparities arise in these groups? How do these disparities compare to those shown when the predictions are split by race?
Ans: If we split the predictions into different gender or age categories, we would have demographic disparity in our result [4, 5]. As it has been already discussed in Q3, there are unbalanced classes in those categories, so assigning different thresholds for different classes would be necessary to tackle the population bias. The demographic parity of age and gender features have been analyzed for the SVM model to get the statistical metrics. It has been seen from Tables 1-2. that demographic disparity exists for age *less than 25* with higher FNR. Statistical metrics (TPR, TNR) for males and females are pretty comparable for gender. Hence, disparities in accuracies [62.5 − 63.3 %] and cost [$754mil − $763mil] are in a close range for the race, age, and gender attributes.

Table1 Demographic disparity for *age*



```
Attempting to enforce demographic parity...
--------------------DEMOGRAPHIC PARITY RESULTS--------------------

Probability of positive prediction for 25 - 45: 0.5761843790012804
Probability of positive prediction for Greater than 45: 0.5767966181305777
Probability of positive prediction for Less than 25: 0.5429272281275552

Accuracy for 25 - 45: 0.6440460947503202
Accuracy for Greater than 45: 0.6012212306247064
Accuracy for Less than 25: 0.6324611610793132

FPR for 25 - 45: 0.43017127799736493
FPR for Greater than 45: 0.48062593144560356
FPR for Less than 25: 0.3795379537953795

FNR for 25 - 45: 0.28580323785803236
FNR for Greater than 45: 0.25921219822109276
FNR for Less than 25: 0.3604424202992843

TPR for 25 - 45: 0.7141967621419676
TPR for Greater than 45: 0.7407878017789072
TPR for Less than 25: 0.6395575797007157

TNR for 25 - 45: 0.569828722002635
TNR for Greater than 45: 0.5193740685543964
TNR for Less than 25: 0.6204620462046204

Threshold for 25 - 45: 0.09
Threshold for Greater than 45: 0.09
Threshold for Less than 25: 0.10999999999999999

Total cost:
$-754,676,728
Total accuracy: 0.633003788228772
-----------------------------------------------------------------
```

Table2 Demographic disparity for *gender*



```
Attempting to enforce demographic parity...
--------------------DEMOGRAPHIC PARITY RESULTS--------------------

Probability of positive prediction for Female: 0.6204081632653061
Probability of positive prediction for Male: 0.6250705178833352

Accuracy for Female: 0.5974489795918367
Accuracy for Male: 0.631614577456843

FPR for Female: 0.5193465176268272
FPR for Male: 0.49296799224054316

FNR for Female: 0.23212045169385195
FNR for Male: 0.25997045790251105

TPR for Female: 0.767879548306148
TPR for Male: 0.740029542097489

TNR for Female: 0.48065348237317285
TNR for Male: 0.5070320077594568

Threshold for Female: 0.05
Threshold for Male: 0.09999999999999999

Total cost:
$-763,241,866
Total accuracy: 0.6254273306846531
-----------------------------------------------------------------
```

References:

1. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A Survey on Bias and Fairness in Machine Learning. ArXiv, abs/1908.09635.
2. Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. Science Advances 4, 1 (2018). https://doi.org/10.1126/sciadv.aao5580 arXiv:https://advances.sciencemag.org/content/4/1/eaao580.full.pdf
3. Machine Learning Fairness Primer (2020). https://mlcourse-ub.readthedocs.io/en/latest/docs.html
4. Das, S., Donini, M., Gelman, J., Haas, K., Hardt, M., Katzman, J., ... & Zafar, B. Fairness Measures for Machine Learning in Finance.
5. Zhang, Y., & Zhou, L. (2019). Fairness assessment for artificial intelligence in financial industry. arXiv preprint arXiv:1912.07211.