Analyzing the Biases in Machine Learning: How Fair Can We Be?

Rutu R. Gandhi[*]
Institute of Artificial Intelligence
University of Georgia
Athens, GA 30602
rutu.gandhi@uga.edu

Dhaval R. Bhanderi[†]
Department of Computer Science
University of Georgia
Athens, GA 30602
dhaval.bhanderi@uga.edu

Shrinidhi S. Adke[‡]
Institute of Artificial Intelligence
University of Georgia
Athens, GA 30602
shrinidhi.adke@uga.edu

Abstract

With the growing use of AI and Machine learning in every field around the world, often times the Fairness of these algorithms is left unconsidered. The focus of this project will be on understanding and mitigating discrimination based on sensitive characteristics such as religious, ethical and social beliefs. Recent years have shown that unintended discrimination and bias arise naturally and frequently in the use of machine learning and algorithmic decision making. The two factors that should be taken into account while evaluating fairness are incomplete information and limited resources during the model building process. Our focus is modeling classifiers in such a way that the final classification process can be considered fair as per the evaluation standards. In this project, we also explore the interaction between multiple decision-makers. The model can choose to say 'I DON'T KNOW!', and pass the decision to a downstream decision-maker. We extend this concept by applying *learning to defer*, a generalization of rejection learning. Our experiments show that these post-processing techniques can make not only less biased but also more accurate predictions. Even working with inconsistent data, our results show that deferring models greatly improve fairness and/or accuracy.

1 Introduction

Have you ever been treated unfairly? How does it feel? Probably not so good. The concept of *fairness* is ingrained in our psyche as a fundamental, essential part of our existence. The growing use of Machine Learning and algorithmic decision-making in complex domains such as criminal justice, loan approvals and medical diagnosis, has raised critical questions about the role of machine learning in high-stakes decision-making. Bias in ML has been almost ubiquitous when the application involves people and it already has hurt the benefit of people in minority groups.

There are many definitions of fairness that have been proposed in literature. The most common definitions of fairness in machine learning are statistical in nature. They proceed by fixing a small

[*]Enrolled for CSCI 8360 Data Science Practicum
[†]Enrolled for CSCI 8360 Data Science Practicum
[‡]Enrolled for ARTI 8950 Machine Learning

CSCI 8360 Data Science Practicum (2019), Athens, GA.

number of "protected subgroups" (such as racial or gender groups) and then ask that some statistic of interest be approximately equalized across groups. These statistics include *positive predictive value (PPV)*, *negative predictive value (NPV)*, *false positive rate (FPR)* and *false negative rates (FNR)*. The task of classification using complex machine learning algorithms considering appropriate fairness becomes difficult under the circumstances of incomplete/partial information and limited resources. In different fields where ML is applied, the notion of fairness changes. Canetti et al. (2019) deal with this problem by processing scores of individual groups from soft classifiers and then taking the binary decision. These hard decisions are obtained by applying various post-processing algorithms. Our aim is to evaluate the possibility of these hard decisions holding true for different datasets.

The task of post-processing a traditional soft classifier calibrated score using a threshold is one of the areas to be discussed in this work. When there are certain fairness constraints and some of the groups are *protected*, it becomes challenging to come up with a binary classifier, as some of the information is incomplete.

Rejection learning also proposes a solution: allow models to *pass* (and not making a prediction) when they are not confidently accurate. We are exploring a two-stage framework containing an automated model and an external decision-maker. Here both the model and the decision-maker act independently of one another. The aspect of this two-stage mechanism is that each stage can be viewed as aimed at a different goal: the first stage is aimed at gathering information and providing the best accuracy possible, with probably minimal regard of fairness. The second stage is aimed to extract a decision from the information collected in the first stage and making sure that "errors are distributed fairly".

To visualize this concept, we can give a simple example of bias in a Facial recognition software where the model is giving a high true positive rate for one race, but low for another minor race. In this case, the proposed method will consider this task in two stages, first, the scores for each group- including race- is assigned by a soft classifier. Then, the second stage takes the decision by post-processing the soft classifier and also considering the group to which the face belongs. Based on the input facial features and given circumstances, we evaluate whether this hard decision is "fair" or not.

Briefly, our motivation for this project is to find the answer to this broad question:

Under what conditions should we post-process a calibrated soft classifier's outputs so that the resulting hard classifier equates a subset of PPV, NPV, FPR, FNR across the set of protected groups?

In this project, we conduct an extensive set of experiments to answer these questions. We study the algorithm from Canetti et al. (2019). On three real datasets, we characterize:

(1) The trade-off between group fairness and accuracy. We find that for each dataset, thus achieving rich group fairness is possible in practice without a severe loss in prediction accuracy.

(2) A threshold post-processing technique usually returns 1 for individual i whenever the score is above some fixed threshold, and return 0 otherwise. We find that while using different thresholds for the different groups, we can equalize PPV or NPV across the two groups.

(3) We consider the post-processing strategies that do not always output a decision. It means, with some probability the output is "I don't know"(), which means that the decision is deferred to another downstream decision-maker. Our first strategy is a natural extension of the per-group threshold: we used two thresholds per group, which returns 1 above the right threshold, 0 below the left threshold, and between the two thresholds. After applying this technique on various datasets, we find that there exists a way to select the values of the thresholds such that, both the PPV and NPV are equal across each group.

2 Related Work

This work is focused on group fairness notions and the formalization of PPV, NPV, FPR, FNR which are explained by Chouldechova (2017) and Kleinberg et al. (2016). The concepts of soft and hard calibration were formalized by Pleiss et al. (2017) which combined with above notions, lays the formal ground for this work.

Corbett-Davies et al. (2017) deals with the post-processing in decision making by considering the notions of statistical parities, FPR and PPV. Our work also deals with the use of deferrals in post-processing as discussed by Madras et al. (2018) in which they use two-threshold deferring

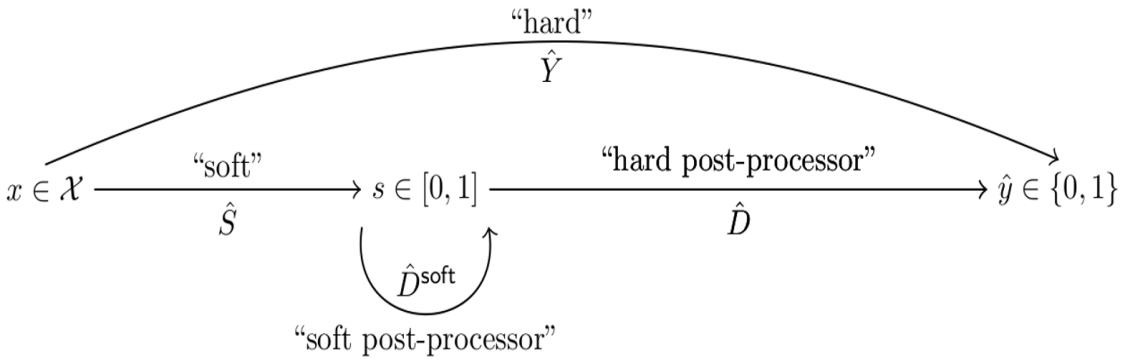

Figure 1: Hard classifier returns decision in {0,1} while a Soft classifier gives results in range [0,1]. *Post-processors* are classifiers that take the output of soft classifier as input. Canetti et al. (2019)

post-processors for promoting a combination of accuracy and fairness. We have built this work on top of the above literature and extended the algorithm for multiple groups of various datasets as discussed in further sections.

3 Methods and definitions

To begin our analysis of fairness, we first need to briefly understand the notions of various terminologies we have adopted from previous works.

3.1 Post-Processing Approach

The ultimate aim of this work is to study binary classifiers which classifies an instance of data $x \in \mathcal{X}$ with a *true type* $\mathrm{Y}(x) \in \{0,1\}$. Also, each instance x is associated with a *group* $\mathrm{G(x)} \in \mathcal{G}$, where $\mathcal{G}$ is the set of groups.

A *classifier* is a randomized function with domain $\mathcal{X} \times \mathcal{G}$. As explained in 1, a *hard classifier*, denoted $\hat{Y}$, outputs a *prediction* in {0,1} while a *soft classifier*, denoted $\hat{S}$, outputs a *prediction* in [0,1].

A *Post-processor* is a randomized function with domain $[0,1] \times \mathcal{G}$. We consider both post-processors, Hard $\hat{D}$ and Soft $\hat{D}^{\text{soft}}$. As with classifiers, we call a post-processor group blind if its output is independent of the group g.

3.2 Accuracy Profile

The *Accuracy Profile* (AP) of a calibrated soft classifier $\hat{S}$ for a group g, denoted by $\hat{\mathcal{P}}_{\mathrm{g}}$, is the Probability Mass Function (PMF) of $\hat{S}(X_{\mathrm{g}})$. That is, for $\mathrm{s} \in [0,1]$, $\hat{\mathcal{P}}_{\mathrm{g}}(\mathrm{s}) = \Pr[\hat{S}(X_{\mathrm{g}}) = \mathrm{s}]$. The AP is a distribution of scores for a calibrated classifier $\hat{S}$, which conveys the information about the performance of $\hat{S}$, and is constrained by underlying distribution on X.

3.3 Group Fairness Measures

The notions of group fairness is formally defined by following set of definitions:

false positive rate of $\hat{Y}$ for g: $\mathrm{FPR}_{\hat{Y},g} = \Pr[\hat{Y}(X_{\mathrm{g}}) = 1 \mid \mathrm{Y}(X_{\mathrm{g}}) = 0]$;

false negative rate of $\hat{Y}$ for g: $\mathrm{FNR}_{\hat{Y},g} = \Pr[\hat{Y}(X_{\mathrm{g}}) = 0 \mid \mathrm{Y}(X_{\mathrm{g}}) = 1]$;

positive predictive value of $\hat{Y}$ for g: $\mathrm{PPV}_{\hat{Y},g} = \Pr[\mathrm{Y}(X_{\mathrm{g}}) = 1 \mid \hat{Y}(X_{\mathrm{g}}) = 1]$;

negative predictive value of $\hat{Y}$ for g: $\mathrm{NPV}_{\hat{Y},g} = \Pr[\mathrm{Y}(X_{\mathrm{g}}) = 0 \mid \hat{Y}(X_{\mathrm{g}}) = 0]$;

All the previous works discusses equalizing only one or both of the FPR and FNR, which is known as *balance* for negative and positive classes respectively. Notion of *Predictive Parity* discusses equalizing PPV and NPV across groups. Here we split these values for the two classes. To achieve this, we try to equalize PPV or NPV as well as AP by considering a threshold and a threshold with a deferral condition.

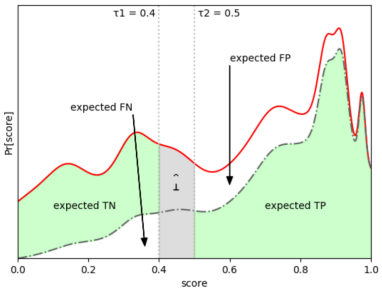

Figure 2: Illustration of Threshold post-processor with deferrals defer between the thresholds. Canetti et al. (2019)

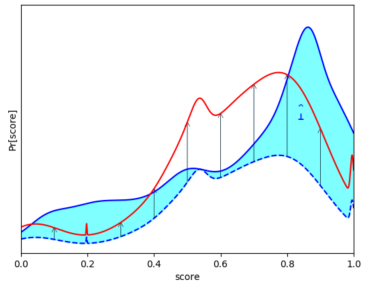

Figure 3: Equalizing AP with deferral: Here, the red line is the original AP. By deferring at the rates indicated by the shaded region, the resulting conditional AP is represented by the dark blue line. Canetti et al. (2019)

3.4 Equalizing with Threshold and deferral

PPV and NPV cannot both be equalized across groups in general when using only a single threshold per group. By using two thresholds per groups and deferring on some inputs, PPV and NPV can always be equalized across groups.

A simple threshold post-processor $\hat{\mathcal{D}}_{(\tau,\mathcal{R})}$: [0,1] $\times\mathcal{G} \rightarrow$ 0,1 is a function from a score s $\in$ [0,1] and a group $g \in \mathcal{G}$, parameterized by $\tau, \mathcal{R}$. The threshold parameter $\tau : \mathcal{G} \rightarrow [0,1]$ specifies the threshold for the group g, and $\mathcal{R} : \mathcal{G} \rightarrow [0,1]$ is the probability of returning 1 when the input score s is on the threshold $\tau(g)$.

While we consider the notion of *deferral* illustrated in figure 2 in which the inputs that falls between two thresholds are assigned the deferral value. A deferring threshold post-processor $\hat{\mathcal{D}}_{(\tau_0,\tau_1,\mathcal{R}_0,\mathcal{R}_1)}$ assigns to each group g two thresholds $\tau_0(g)$, $\tau_1(g) \in$ Supp($\hat{\mathcal{P}}_g$), and two probabilities $\mathcal{R}_0$(g), $\mathcal{R}_1$(g) $\in$ [0,1], with the following requirements,

1. for all $g \in \mathcal{G}$, $\tau_0(g) \leq \tau_1(g)$
2. for all $g \in \mathcal{G}$ for which $\tau_0(g) = \tau_1(g)$, $\mathcal{R}_0$(g) + $\mathcal{R}_1(g) \leq 1$.
   This corresponds to the case where the two thresholds are the same and therefore individuals with that score must be mapped to 1 with probability $\mathcal{R}_1(g)$, and to 0 with probability $\mathcal{R}_0(g)$, with the remainder mapped to $\perp$.

3.5 Equalizing AP with deferrals

While thresholding is a conceptually simple approach to post-processing a soft classifier, its power is limited. We now consider a very different approach using soft post-processors to equalize the APs across groups of a soft classifier. The intuition is that if the APs are equal across groups, then any hard post-processor that is group blind should result in equal PPV, NPV, FPR, and FNR. Equalizing the conditional APs between groups renders trivial the task of downstream decision-making subject to equality of PPV, NPV, cFPR, and cFNR. Figure3 shows this concept with sample example. Choosing the deferrals appropriately allows transforming one AP into another (conditional) AP.

4 Experiments

We ran experiments on 3 datasets: **COMPAS** data made available by Propublica, **Loan Status** and **Student Performance** dataset from UCI Machine Learning RepositoryDua & Graff (2017). We

Dataset	Size	Prediction	#Features	#Protected	Protected Feature	Baseline
COMPAS	60.8K	Risk Status	28	2	Gender, Race	error rate
Loan Status	185	Loan Repaid?	13	1	Gender	error rate
Student Grades	396	Pass/Fail	17	2	Gender	error rate

Table 1: Description of Datasets studied

selected these datasets due to their potential fairness concerns. The properties of these datasets are summarized in Table 1, including the predictions being made, the number of features, number of protected features.

4.1 COMPAS Dataset

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is a popular commercial algorithm used by judges and parole officers for scoring criminal defendant's likelihood of re-offending (recidivism). It gives an insight on who actually committed violent crimes after 2 year of receiving bail. The algorithm has been shown to be biased in favor of white defendants, and against black inmates. The pattern of mistakes, as measured by precision/sensitivity is notable. The COMPAS scoring mechanism is an approximately calibrated soft classifier with 10 possible outcomes. We note here that the distribution of the COMPAS scores differs significantly across groups. In particular, the scores for Caucasians are skewed as opposed to the even distribution seen with African-Americans. Similarly, scores for Males are more evenly distributed as opposed to the skewed distribution of Female scores.

Black defendants were often predicted to be at a higher risk of recidivism than they actually were. The analysis shows that white defendants were often predicted to be less risky than they were. Black defendants who did not recidivate over a two-year period were nearly twice as likely to be misclassified as higher risk compared to their white counterparts (45 percent vs. 23 percent). The analysis found that white defendants who re-offended within the next two years were mistakenly labeled low risk almost twice as often as black re-offenders (48 percent vs. 28 percent). The analysis also showed that even when controlling for prior crimes, future recidivism, age, and gender, black defendants were 45 percent more likely to be assigned higher risk scores than white defendants.

4.1.1 Thresholding with Deferrals

We first used this mechanism for post-processing that uses two thresholds i.e. a range to defer. The deferrals equalize PPV and NPV across Caucasians and African-Americans. For simplicity, an approximate equalization is performed. The number of deferrals is smaller than 20% which shows that a large number can be classified without a downstream decision-maker. The resulting deferring hard classifier is stricter towards the African-American group as shown in Figure 4 and Figure 5.

Similarly, the deferrals equalize PPV and NPV across Male and Female inmates. The number of deferrals is smaller than 22% which shows again that a large number can be classified without a downstream decision maker.

Next we look at two methods to equalize all four quantities PPV, NPV, FPR, NPR. Equalizing APs of the two groups post-deferral achieves this goal. Following are the methods:

4.1.2 Converting one AP into another

In this method, first we use deferrals to create conditional AP for African-Americans that matches the APs of Caucasians (Figure 8). Then we create conditional AP for Caucasians that matches the AP of African-Americans. (Figure 9)

Similarly, we use deferrals to create conditional AP for Females that match the APs of Males (Figure 10). Then we create conditional AP for Males that match the APs of Females (Figure 11).

4.1.3 Equalizing APs using min PDF method

In this method we take an equal fraction of African-Americans and Caucasians. This fraction is the total variation distance called the total deferral rate between the two APs which is equalized across

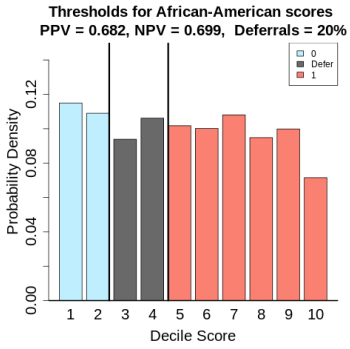

Figure 4: Equalizing PPV and NPV using two thresholds for African-Americans

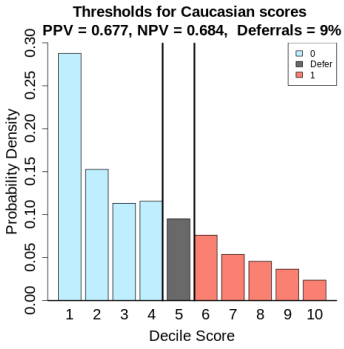

Figure 5: Equalizing PPV and NPV using two thresholds for Caucasians

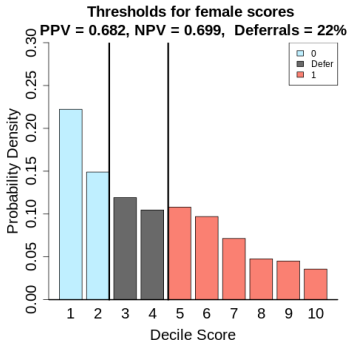

Figure 6: Equalizing PPV and NPV using two thresholds for Female inmates

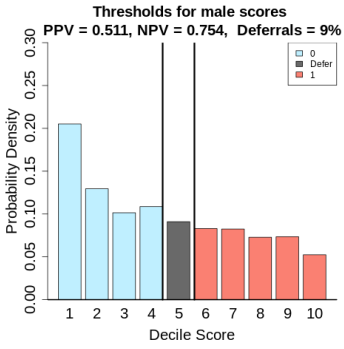

Figure 7: Equalizing PPV and NPV using two thresholds for Male inmates

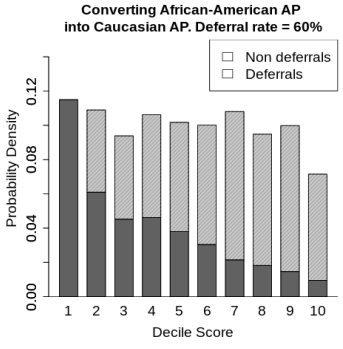

Figure 8: African-Americans to Caucasians

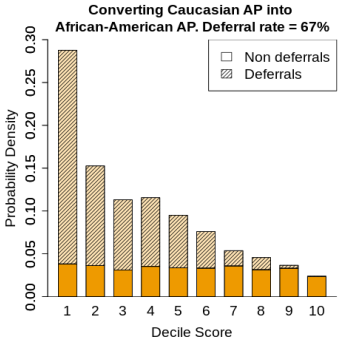

Figure 9: Caucasians to African-Americans

the two groups. The result on the racial groups is shown in Figure 12. Similarly, this method is performed on the gender groups as shown in Figure 13.

4.2 Loan Status Dataset

This dataset looks at various attributes of loan-applicants and assigns a score to the individual. The scores assigned to each individual indicates the risk of them not repaying within the first payback period. Here, we study the gender-based bias in this scoring mechanism.

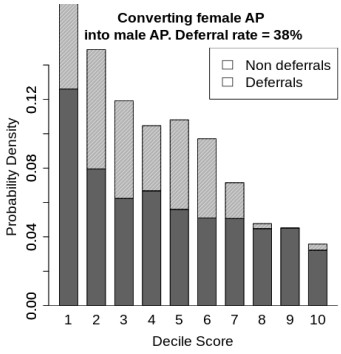

Figure 10: Females to Males

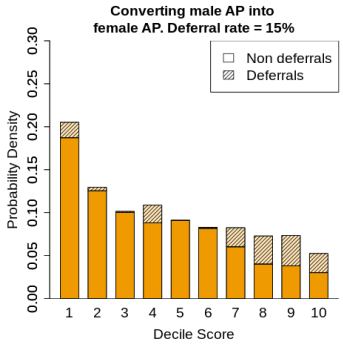

Figure 11: Males to Females

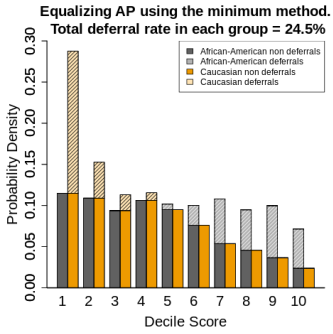

Figure 12: Equalizing AP of African-American and Caucasian defendants for COMPAS

4.2.1 Thresholding with Deferrals

We first used this mechanism for postprocessing that uses two thresholds i.e. a range to defer. The deferrals equalize PPV and NPV across Males and Females. For simplicity, an approximate equalization is performed. The number of deferrals is smaller than 2% which shows that a large number can be classified without a downstream decision-maker.

4.2.2 Converting one AP into another

In this method, first we use deferrals to create conditional AP for Females that matches the APs of Males (Figure 14). Then we create conditional AP for Males that matches the AP of Females (Figure 15).

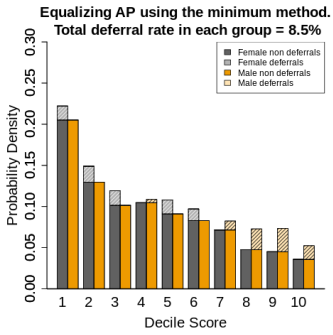

Figure 13: Equalizing AP of Female and Male defendants for COMPAS

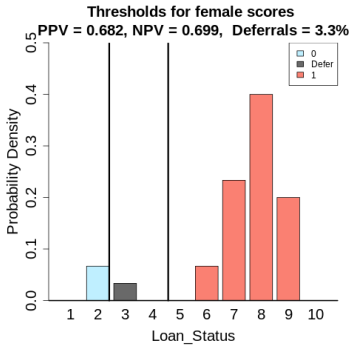

Figure 14: Equalizing PPV and NPV using two thresholds for Females for the loan dataset

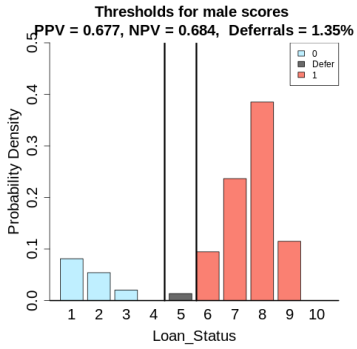

Figure 15: Equalizing PPV and NPV using two thresholds for Males for the loan dataset

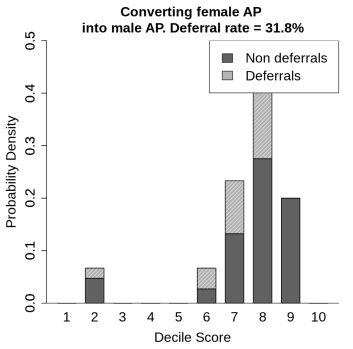

Figure 16: Female to Male for LOAN

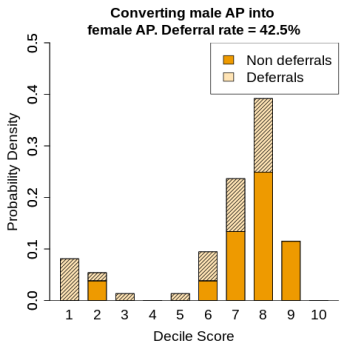

Figure 17: Male to Female for LOAN

4.2.3 Equalizing APs using min PDF method

In this method we take an equal fraction of Males and Females. This fraction is the total variation distance called the total deferral rate between the two APs which is equalized across the two groups. The result on the gender groups is shown in Figure 16.

4.3 Student Dataset

This dataset depicts student achievement in secondary education of two Portuguese schools. The data attributes include student demographic, grades, social and school related features. The dataset is provided regarding the performance in two distinct subjects: Mathematics(G1) and Portuguese language(G2).

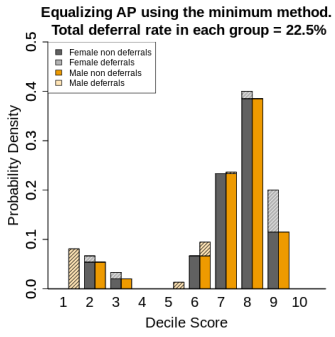

Figure 18: Equalizing AP of Females and Males in the Loan dataset

We first ran two-threshold post-processing technique and obtained a binary decision algorithm with deferrals which equalizes both PPV and NPV across Males and Females(Figure 19 20).

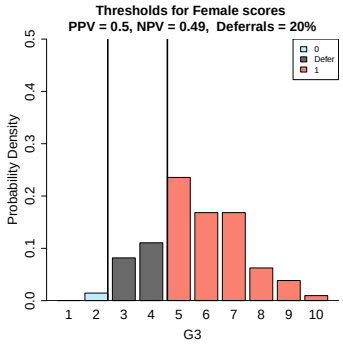

Figure 19: Equalizing PPV and NPV using two thresholds for Females for the Student dataset

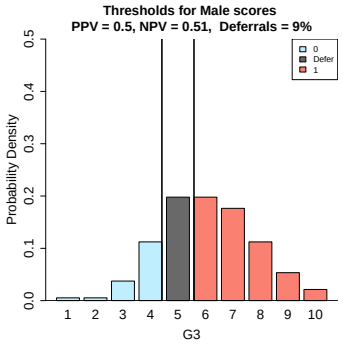

Figure 20: Equalizing PPV and NPV using two thresholds for Males for the Student dataset

4.3.1 Converting one AP into another

First we use deferrals to create conditional AP for Females that matches the APs of Males(Figure 21). Then we create conditional AP for Males that matches the AP of Females(Figure 22).

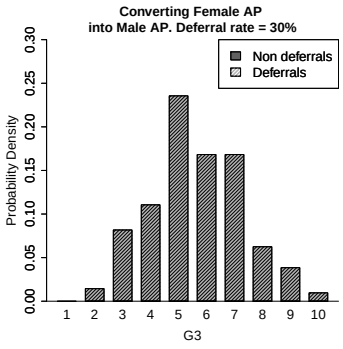

Figure 21: Female to Male

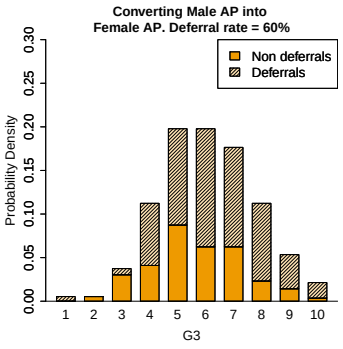

Figure 22: Male to Female

4.3.2 Equalizing APs using min PDF method

First, using the method of deferring only on Females we defer on roughly 36% of the total decisions. when we defer only on Males, this number goes down to roughly 25%(Figure 21 22).

5 Results

As seen in Figure 12, when the decile score is towards the lower end of the spectrum, the probability density of the Caucasians is higher which means the deferral rate is higher. Towards the higher end of the decile score spectrum, the probability distribution of the African-Americans is higher which means their deferral rate is higher. This tells us that, at higher decile risk scores, the confidence of the classifier in the African-American predictions is lower compared to Caucasians with the same score.

In the gender-based study of the COMPAS dataset, Figure 13 shows that at higher decile risk scores, the confidence of the classifier in the Male predictions is lower compared to Females with the same score.

In the analysis of the Loan dataset, Figure 18 shows that at higher decile risk scores, the confidence of the classifier in the Female predictions is lower compared to Males with the same score. The total deferral fraction is 22.5% for both groups after equalizing the APs using the min PDF method.

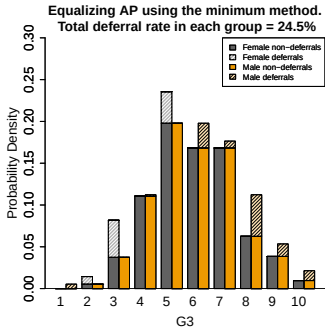

Figure 23: Equalizing AP of African-American and Caucasian defendants

For the Student performance dataset, we observed that the deferral rate in total is smaller 20% of the decisions made which means that a large number of the defendants can be classified without having to defer. The resulting deferring hard classifier is fair for both the groups. We observe total deferral fraction is around 10% when we are deferring on an equal fraction of Males and Females.

Secondly, in all AP equalization methods, the deferral happens where the confidence is high, i.e. towards the extremes (close to 0 or close to 1). This is contrasting to the two-threshold method where the deferral happens in the low confidence regions (close to 3,4 and 5). While this seems counter-productive, any method that seeks to first equalize the accuracy profiles will have to defer most on the scores which appear in different probabilities across the two groups (which in case of the COMPAS data is the extremes). Deferring on these extremes makes sense from a social point of view because when a score appears at drastically different rates across two groups, the decision-making should be given to a moderator so that systemic bias can be kept in check.

On the other hand, if one wants to have a classifier that only defer on low confidence instances, and still guarantee equal APs for the protected groups, then the soft classifier should be designed such that the APs are the same (either close to 0 or 1).

6 Conclusions and Future Work

In this project, we applied a framework for two-stage decision-making on different socially sensitive datasets. We implemented a method, learning to defer, which generalizes adaptive rejection learning. Whether a classifier will promote fairness or not, depends on the context, and this is for both deferring and non-deferring classifies. Here our goal is to minimize the overall rate of deferrals while maintaining the rate of PPV, NPV, FPR and FNR. However, more deep work is needed to better understand the full span of the deferral strategies.

Our project explores the binary classification post-processing techniques. The logical future scope would be to expand it to multi-class classification. Intuitively, if there are n classes, n-1 deferral ranges will have to be assigned.

Another application of these techniques would be in image classification. IBM's face recognition dataset is one such relevant dataset that can be explored. The two possible classes are matched and not-matched. Upon applying these techniques, if the model is less confident about the recognition, it will defer the decision thus avoiding security breaches up to a certain extent.

References

Ran Canetti, Aloni Cohen, Nishanth Dikkala, Govind Ramnarayan, Sarah Scheffler, and Adam Smith. From soft classifiers to hard decisions: How fair can we be? In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 309–318, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287561. URL `http://doi.acm.org/10.1145/3287560.3287561`.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797–806. ACM, 2017.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 100–109. ACM, 2019.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, pp. 6147–6157, 2018.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pp. 5680–5689, 2017.