Fairness in Machine-Learning-Generated Risk Scores via Equitable Thresholding

Jordan Purdy*, Brian Glass, and Fariborz Pakseresht

Oregon Department of Human Services, Office of Reporting, Research, Analytics and Implementation

May 3, 2018

Abstract

When machine learning models are built using historical data that demonstrate bias or discrimination, i.e. unfairness, toward one or more levels of a protected attribute, it is well understood that the subsequent model output will likely perpetuate such unfairness. Approaches for combating this issue have been proposed, though many of these are limited to situations in which the protected attribute can be reduced to two possible levels and the desired model output is restricted to a prediction of the binary outcome of interest. In this paper, we propose an extension of the thresholding scheme introduced in Lipton et. al. [7]. This extension allows the protected attribute to have an arbitrary number of levels and it allows for the use of risk scores associated with the binary outcome of interest. We develop this extension within the context of a salary data set and demonstrate that the proposed fairness correction procedure produces equitable thresholds which both increase fairness and largely maintain predictive performance.

1 Introduction

The development of machine learning-based instruments to aid in agency and cross-agency decision making is a cornerstone of the work performed by the Office of Reporting, Research, Analytics, and Implementation (ORRAI) within Oregon's Department of Human Services. The objective of this paper is neither to provide the rationale behind this approach nor to provide further evidence that machine learning models, when coupled with professional discretion, can lead to improved decision making. Instead, the objective of this paper is to present and demonstrate the methodology of a proposed solution for mitigating an inherent limitation of machine learning methods: the perpetuation of unfairness.

The need for such solutions are irrefutable in light of the growing body of literature on fairness, or lack there of, in algorithmic decision making [e.g. 2, 4]. More to the point, when machine learning models are built using historical data that demonstrate bias or discrimination, i.e. unfairness, the subsequent model output can perpetuate, or even exacerbate, such unfairness [3, 4]. In ORRAI's search for a viable method to overcome this problem, several proposed approaches were assessed [1, 8], and ultimately, for our machine learning objectives, a "thresholding scheme" introduced in Lipton et. al. [7] was identified as the most appropriate. The mathematical details justifying the

*Direct all inquires to jordan.e.purdy@dhsoha.state.or.us

attractiveness of this method are beyond the scope of this paper, but suffice to say the intuitive appeal of the approach lies in its transparency, interpretability, and operationalizability.

In order to meet the objectives of our machine learning instruments, however, an extension of this thresholding scheme was required. More specifically, as initially proposed this method applies to situations in which the protected attribute can be reduced to two possible levels (e.g. race coarsely categorized as white or non-white) and the desired model output is restricted to a prediction of the binary outcome of interest (e.g. whether or not an individual will default on a loan). Our proposed extension of the thresholding scheme allows the protected attribute to have an arbitrary number of levels (e.g. multiple race categories) and it allows for the use of risk scores associated with the binary outcome of interest (e.g. low, some, moderate, or high risk of defaulting on the loan). We will hereafter interchangeably refer to this method as either a risk score thresholding scheme or a fairness correction procedure.

1.1 Document Outline

In presenting an extension of existing statistical methodology, a traditional approach would be to include a section in which the procedure is developed in a general sense, followed by one or more sections in which the procedure is deployed on multiple examples, perhaps even incorporating some form of simulation study as a more rigorous demonstration of viability. This will not be the approach adopted in this paper. To be clear, we do hope academics and other individuals working on reducing unfairness in machine learning algorithms will read this paper, provide scrutiny, and perhaps eventually produce improvements or rigorous simulation-based[1] support for the procedure. We also hope, however, that this paper serves as an accessible how-to guide for those interested in reducing unfairness in their statistical machine learning algorithms. To that end, we will facilitate tractability and limit cumbersome notation by forgoing a general development of the risk score thresholding scheme. Instead, in Section 2 we develop and explain the fairness correction procedure within the context of the salary data set from the UCI Machine Learning Repository. We believe such development will be sufficient to enable potential users to extrapolate the procedure to an arbitrary data set. We also note here that since this fairness correction procedure is applied after a machine learning model has been fit and the subsequent risk scores assigned, it is essentially indifferent to the choice of learner so long as predicted probabilities[2] are an obtainable output of the learner. For this reason, there will be limited discussion of the choice and development of the machine learning model utilized for the salary data. Instead, the focus will be on how the risk scores resulting from the model are fairness corrected. Finally, concluding remarks are provided in Section 3, including a discussion of the primary practical limitation of the presented fairness correction procedure.

2 Fairness Via Equitable Thresholds

To develop and demonstrate our fairness correction procedure, we utilize the *Adult Data Set* from the UCI Machine Learning Repository [6]. The outcome variable of interest in this data set is whether or not a person has an annual salary greater than $\$50,000$. For illustrative purposes, the subset of potential predictors utilized within this example are the following: age (in years), amount of education (in years), and race, where the levels of race are American Indian/Native Alaskan (AN), Asian/Pacific Islander (AP), Black (BL), Other (OT), and White (WT). A routine objective

[1] Two simulation studies of interest are proposed in Section 3.1.

[2] These predicted probabilities should arise from either a test set or a fold of a cross-validation procedure.

with such data is to leverage the available predictors to train a model which ultimately outputs a score indicating the "potential" of a person to earn an annual salary greater than $\$50,000$. Note that it is commonplace to label such model output as "risk", as opposed to "potential", but given the context, we will forgo the standard naming convention in lieu of something more intuitive. As will be demonstrated below, the scores outputted by the trained model are unfair across the levels of race, which is the only protected attribute among the covariates utilized. The objective of this section is to develop and demonstrate our fairness correction procedure for adjusting these scores so that 1) fairness is increased across the levels of the protected attribute and 2) the predictive ability of the model is not substantially compromised.

2.1 Defining Fairness

There are multiple mathematical definitions of fairness in the literature. Chouldechova [5] defines, discusses, and compares four commonly utilized definitions of fairness: calibration, predictive parity, error rate balance, and statistical parity. Unfortunately, these and other definitions of fairness are often in competition with one another [4, 5]. In other words, by meeting criteria articulated by one definition of fairness, one or more criteria demanded by alternative definitions of fairness are necessarily violated. The selection of a fairness criteria should therefore be rooted in the context of the data [3]. With this in mind, error rate balance is currently the criteria being utilized for assessing fairness in the machine learning endeavors of ORRAI and it will likely continue to be a commonly implemented criteria in future endeavors. The aforementioned machine learning endeavors and corresponding context will be the focus of a companion paper to this manuscript. Consequently, for illustrative purposes, error rate balance is the criteria for fairness utilized in this example, though our fairness correction procedure can also be utilized in an analogous manner with any of the other three definitions named above.

Having selected the fairness criteria, we now restructure the protected attribute in the data set (i.e. race) to consist of three levels rather than five. The rationale behind doing so is two-fold. First, the numbers of individuals whose race is identified as Other or American Indian/Alaskan Native are insufficient[3] to obtain a robust set of fairness-corrected scores. Second, as the procedure is particularly useful when the protected attribute has more than two levels, we can limit the length of this document without compromising the development of the method by using a three-level protected attribute as opposed to a four-level protected attribute. Hence, while the first reason is pragmatic in nature, the second is ultimately for illustrative purposes in considering the overall scope of this document. With that being said, we restructure the race variable to consist of levels ANO, APW, and BLK, where ANO indicates that the individual is American Indian, Native Alaskan, or Other, APW indicates that the individual is Asian, Pacific Islander, or White, and BLK indicates that the individual is Black. The justification for grouping the initial five levels of race in this way is evident in Table 1, where the sample size (i.e. Count), observed prevalence of the outcome variable (i.e. Proportion w/ annual income $> \$50,000$), and disparity index, relative to White, for the outcome variable (i.e. Disparity Index) are provided for each of the original five levels of race in the data set. In particular, AN and OT are combined given that their disparity index values (0.4525 and 0.3606) are similar and that each are too small in overall count to be their own levels. Finally, while AP and BL are both large enough in overall count to be their own levels, because we wanted three levels and the disparity index values for AP and WT are more similar than the disparity index values for BL and OT, we combined AP and WT into one level as opposed to uniting BL with the already combined AN and OT levels.

[3]Insufficient samples sizes across protected attribute levels is related to the primary practical limitation of the method, which is more formally addressed in Section 3.2.

Table 1: *For each of the original five levels of Race in the data set (American Indian/Native Alaskan (AN), Asian/Pacific Islander (AP), Black (BL), Other (OT), and White (WT)), the number of individuals identifying as that race (Count), the observed proportion making more than* $\$50,000$ *annually (Proportion w/ annual income* $> \$50,000$*), and the disparity index, relative to White, for the outcome variable (Disparity Index) are provided.*

Race	Count	Proportion w/ annual income $> \$50,000$	Disparity Index
AN	311	0.1158	0.4524
AP	1039	0.2656	1.0382
BL	3124	0.1239	0.4842
OT	271	0.0923	0.3606
WT	27816	0.2559	1.0000

Before formally defining fairness via error rate balance, we establish relevant notation. Let $Y = \{0, 1\}$ denote annual salary with $Y = 1$ indicating the individual makes more than $\$50,000$ annually and $Y = 0$ indicating the individual makes less than or equal to $\$50,000$ annually. Let $X = \{X_1, X_2, X_3\}$ denote the vector of covariates, where X_1 denotes age, X_2 denotes amount of education, and X_3 denotes race, where the three possible levels of race are denoted as ANO, APW, and BLK. Let $S = S(X) \in \{1, 2, 3, 4\}$ denote the potential score (i.e. risk score), where scores of 1, 2, 3, and 4 indicate, respectively, low potential, some potential, moderate potential, and high potential for making more than $\$50,000$ annually. We utilize a four-level potential (i.e. risk) scoring system in this example because the objectives of the initial models developed in ORRAI corresponded nicely with such a four-level scoring system. However, our fairness correction procedure can theoretically be applied to scoring systems with an arbitrary number of levels, as long as there are sufficiently many individuals within each level of the protected attribute. Such four-level scoring systems have three thresholds: the first separates scores less than or equal to 1 from scores greater than 1, the second separates scores less than or equal to 2 from scores greater than 2, and the third separates scores less than or equal to 3 from scores greater than 3. We use T_1, T_2, and T_3 to denote, respectively, these three thresholds. Within the context of this example, we now formally define fairness via the error rate balance criteria.

Error Rate Balance: The potential (i.e. risk) scoring system satisfies error rate balance at a threshold if both the false positive and false negative error rates are equal across the three levels of the protected attribute (i.e. ANO, APW, and BLK). In other words, for this example, there are three thresholds at which error rate balance is assessed:

1. Error rate balance is satisfied at the first threshold, which separates scores less than or equal to 1 from scores greater than 1, if the following holds:

$$\begin{aligned} \mathbb{P}(S > 1 \mid Y = 0,\ X_3 = \text{ANO}) &= \mathbb{P}(S > 1 \mid Y = 0,\ X_3 = \text{APW}) \\ &= \mathbb{P}(S > 1 \mid Y = 0,\ X_3 = \text{BLK}),\ \text{ and} \end{aligned} \tag{1}$$

$$\begin{aligned} \mathbb{P}(S \leq 1 \mid Y = 1,\ X_3 = \text{ANO}) &= \mathbb{P}(S \leq 11 \mid Y = 1,\ X_3 = \text{APW}) \\ &= \mathbb{P}(S \leq 11 \mid Y = 1\ , X_3 = \text{BLK}) \end{aligned} \tag{2}$$

2. Error rate balance is satisfied at the second threshold, which separates scores less than or

equal to 2 from scores greater than 2, if the following holds:

$$\begin{aligned} \mathbb{P}(S > 2 \mid Y = 0,\ X_3 = \text{ANO}) &= \mathbb{P}(S > 2 \mid Y = 0,\ X_3 = \text{APW}) \\ &= \mathbb{P}(S > 2 \mid Y = 0,\ X_3 = \text{BLK}), \text{ and} \end{aligned} \tag{3}$$

$$\begin{aligned} \mathbb{P}(S \leq 2 \mid Y = 1,\ X_3 = \text{ANO}) &= \mathbb{P}(S \leq 2 \mid Y = 1,\ X_3 = \text{APW}) \\ &= \mathbb{P}(S \leq 2 \mid Y = 1\ , X_3 = \text{BLK}) \end{aligned} \tag{4}$$

3. Error rate balance is satisfied at the third threshold, which separates scores less than or equal to 3 from scores greater than 3, if the following holds:

$$\begin{aligned} \mathbb{P}(S > 3 \mid Y = 0,\ X_3 = \text{ANO}) &= \mathbb{P}(S > 3 \mid Y = 0,\ X_3 = \text{APW}) \\ &= \mathbb{P}(S > 3 \mid Y = 0,\ X_3 = \text{BLK}), \text{ and} \end{aligned} \tag{5}$$

$$\begin{aligned} \mathbb{P}(S \leq 3 \mid Y = 1,\ X_3 = \text{ANO}) &= \mathbb{P}(S \leq 3 \mid Y = 1,\ X_3 = \text{APW}) \\ &= \mathbb{P}(S \leq 3 \mid Y = 1\ , X_3 = \text{BLK}) \end{aligned} \tag{6}$$

To help clarify the notion of error rate balance, consider equations 3 and 4 from the second possibility articulated above. In particular, the far left-hand side of equation 3 (i.e. $\mathbb{P}(S > 2 \mid Y = 0,\ X_3 = \text{ANO})$) specifies the rate at which individuals in the ANO group are mislabeled as either moderate or high potential for making more than $\$50,000$ annually; this is essentially the false positive error rate among the ANO group at the threshold separating scores less than or equal to 2 from scores greater than 2. Similarly, the far left-hand side of equation 4 (i.e. $\mathbb{P}(S \leq 2 \mid Y = 1,\ X_3 = \text{ANO})$) specifies the rate at which individuals in the ANO group are mislabeled as either low or some potential for making more than $\$50,000$ annually; this is essentially the false negative error rate among the ANO group at the threshold separating scores less than or equal to 2 from scores greater than 2. Hence, equation 3 ultimately specifies that the rate at which individuals are mislabeled as either moderate or high potential for making more than $\$50,000$ annually should be the same across all three levels of the protected attribute. Similarly, equation 4 ultimately specifies that the rate at which individuals are mislabeled as either low or some potential for making more than $\$50,000$ annually should be the same across all three levels of the protected attribute.

2.2 Quantifying Fairness

Equations 1 and 2, 3 and 4, and 5 and 6 ultimately provide a method for assessing whether or not the resulting set of scores are categorically fair at any one of the three thresholds. A fairness correction procedure, however, additionally requires a measure which quantifies how close these scores are at each threshold to meeting the criteria for fairness. To exemplify the measurement we are utilizing for such purposes, we focus on the second threshold, which separates scores less than or equal to 2 from scores greater than 2. The corresponding measures for the first and third thresholds are calculated in an analogous manner.

Our measure for quantifying the fairness of the scores at the second threshold is based on the $\binom{3}{2} = 3$ pairings of levels of the protected attribute. In particular, in light of equations 3 and 4, first calculate the ratio of false positive error rates and the ratio of false negative error rates for each of the three pairings of attribute levels, where the numerator and denominator of the ratio are chosen such that the subsequent value is between zero and one[4]:

[4]The minimum function in equations 7 through 12 ensures that the respective ratio is between 0 and 1.

- Pairwise ratio of false positive error rates between ANO and APW levels at second threshold:

$$\text{FPR}_{\text{ANO,APW}}(T_2) \;:=\; \min\left\{\frac{\mathbb{P}(S>2 \mid Y=0,\ X_3=\text{ANO})}{\mathbb{P}(S>2 \mid Y=0,\ X_3=\text{APW})}\ ,\ \frac{\mathbb{P}(S>2 \mid Y=0,\ X_3=\text{APW})}{\mathbb{P}(S>2 \mid Y=0,\ X_3=\text{ANO})}\right\} \tag{7}$$

- Pairwise ratio of false positive error rates between ANO and BLK levels at second threshold:

$$\text{FPR}_{\text{ANO,BLK}}(T_2) \;:=\; \min\left\{\frac{\mathbb{P}(S>2 \mid Y=0,\ X_3=\text{ANO})}{\mathbb{P}(S>2 \mid Y=0,\ X_3=\text{BLK})}\ ,\ \frac{\mathbb{P}(S>2 \mid Y=0,\ X_3=\text{BLK})}{\mathbb{P}(S>2 \mid Y=0,\ X_3=\text{ANO})}\right\} \tag{8}$$

- Pairwise ratio of false positive error rates between APW and BLK levels at second threshold:

$$\text{FPR}_{\text{APW,BLK}}(T_2) \;:=\; \min\left\{\frac{\mathbb{P}(S>2 \mid Y=0,\ X_3=\text{APW})}{\mathbb{P}(S>2 \mid Y=0,\ X_3=\text{BLK})}\ ,\ \frac{\mathbb{P}(S>2 \mid Y=0,\ X_3=\text{BLK})}{\mathbb{P}(S>2 \mid Y=0,\ X_3=\text{APW})}\right\} \tag{9}$$

- Pairwise ratio of false negative error rates between ANO and APW levels at second threshold:

$$\text{FNR}_{\text{ANO,APW}}(T_2) \;:=\; \min\left\{\frac{\mathbb{P}(S\leq 2 \mid Y=1,\ X_3=\text{ANO})}{\mathbb{P}(S\leq 2 \mid Y=1,\ X_3=\text{APW})}\ ,\ \frac{\mathbb{P}(S\leq 2 \mid Y=1,\ X_3=\text{APW})}{\mathbb{P}(S\leq 2 \mid Y=1,\ X_3=\text{ANO})}\right\} \tag{10}$$

- Pairwise ratio of false negative error rates between ANO and BLK levels at second threshold:

$$\text{FNR}_{\text{ANO,BLK}}(T_2) \;:=\; \min\left\{\frac{\mathbb{P}(S\leq 2 \mid Y=1,\ X_3=\text{ANO})}{\mathbb{P}(S\leq 2 \mid Y=1,\ X_3=\text{BLK})}\ ,\ \frac{\mathbb{P}(S\leq 2 \mid Y=1,\ X_3=\text{BLK})}{\mathbb{P}(S\leq 2 \mid Y=1,\ X_3=\text{ANO})}\right\} \tag{11}$$

- Pairwise ratio of false negative error rates between APW and BLK levels at second threshold:

$$\text{FNR}_{\text{APW,BLK}}(T_2) \;:=\; \min\left\{\frac{\mathbb{P}(S\leq 2 \mid Y=1,\ X_3=\text{APW})}{\mathbb{P}(S\leq 2 \mid Y=1,\ X_3=\text{BLK})}\ ,\ \frac{\mathbb{P}(S\leq 2 \mid Y=1,\ X_3=\text{BLK})}{\mathbb{P}(S\leq 2 \mid Y=1,\ X_3=\text{APW})}\right\} \tag{12}$$

If all six of these pairwise ratios of error rates are equal to 1, then the corresponding scoring system is fair at this threshold according to the error rate balance criteria. Inversely, if any one of these six pairwise ratios in not equal to 1, then the scoring system is not fair at this threshold. Furthermore, a ratio value which is closer to zero indicates greater unfairness within the scoring system than a ratio value which is closer to one[5]. Hence, as a single quantitative measure of the overall fairness of the scoring system at a threshold, we propose utilizing the biggest disparity in pairwise mislabeling rates via the minimum of the six pairwise ratios of error rates:

$$\begin{aligned}\text{ERB}(T_2) := \min\{&\text{FPR}_{\text{ANO,APW}}(T_2)\ ,\ \text{FPR}_{\text{ANO,BLK}}(T_2)\ ,\ \text{FPR}_{\text{APW,BLK}}(T_2)\ ,\\ &\text{FNR}_{\text{ANO,APW}}(T_2)\ ,\ \text{FNR}_{\text{ANO,BLK}}(T_2)\ ,\ \text{FNR}_{\text{APW,BLK}}(T_2)\}\end{aligned} \tag{13}$$

Equation 13, stemming from equations 7 through 12 via equations 3 and 4, quantifies the fairness of the scoring system at the threshold separating scores less than or equal to 2 from scores

[5]A ratio value of 0 can only be obtained if one of the protected attribute levels has an error rate of 0, which is likely a consequence of having too few individuals identifying within that level of the protected attribute. In such situations, we suggest restructuring the levels of the protected attribute such that the level in question is combined with another level of the protected attribute.

greater than 2. To measure the fairness of the scoring system at the threshold which separates a score less than or equal to 1 from scores greater than 1, as well as at the threshold which separates scores less than or equal to 3 from scores greater than 3, calculate the analogue of equation 13 by using equations 1 and 2 or equations 5 and 6, respectively, to create equations 7 through 12. The subsequent values are respectively denoted $\text{ERB}(T_1)$ and $\text{ERB}(T_3)$.

In order to then meet the error rate balance criteria for fairness at all three thresholds, $\text{ERB}(T_1)$, $\text{ERB}(T_2)$, and $\text{ERB}(T_3)$ all need to be within a tolerable distance of one. For this example, we set our distance at 0.20, which is equivalent to tolerating a value for the fairness measure as low as 0.80; in other words, the scores are "fair" at all three thresholds if the following holds: $\text{ERB}(T_1) \geq 0.80$, $\text{ERB}(T_2) \geq 0.80$, and $\text{ERB}(T_3) \geq 0.80$. We selected such a tolerance for two reasons: 1) as will be demonstrated below, it represents a substantial improvement in fairness relative to the baseline and 2) through trial and error it was determined that any tolerance demanding greater fairness lead to unstable performance in our proposed procedure[6].

The pre- and post-fairness corrected values for $\text{ERB}(T_1)$, $\text{ERB}(T_2)$ and $\text{ERB}(T_3)$, as well as the six pairwise ratios of error rates at each of the three thresholds (e.g. equations 7 through 12 for second threshold), can then be compared to assess the effectiveness of a fairness correction procedure. Such comparisons will convey the amount of unfairness that is removed from the initial scoring system and ultimately indicate the viability of a proposed method for increasing fairness. Additionally, the cost of a fairness correction procedure, in terms of predictive performance, can be assessed via a number of performance related measures at each threshold (i.e. accuracy, negative predictive value, positive predicted value, sensitivity, and specificity). As long as these performance measures have not substantially changed for the worse, the cost would be deemed acceptable.

2.3 Pre-Fairness Corrected Results

For each of the 32,561 individuals comprising the training set provided with the *Adult Data Set*, their predicted probability of making more than $\$50,000$ annually was obtained via a logistic regression model utilizing the covariates specified by X. The three thresholds utilized to assign the initial,"pre-fairness corrected", potential (i.e. risk) scores were then obtained via the following bootstrap aggregation (i.e. bagging) procedure:

0. Via the fitted logistic regression model, generate the predicted probability of making more than $\$50,000$ annually for each individual in the training set;
1. Randomly sample, with replacement, 32,561 individuals from the training set;
2. Identify the 10^{th} percentile among the predicted probabilities in the sample. This represents the first threshold for the sample;
3. Determine the prevalence of the outcome variable (i.e. proportion of individuals with $Y = 1$) in the sample. This represents the second threshold for the sample;
4. Identify the 90^{th} percentile among the predicted probabilities in the sample. This represents the third threshold for the sample;
5. Repeat steps 1-4 an additional 999 times;
6. Aggregate the 1000 sets of thresholds via averaging. In other words, the first threshold, for example, is the average of the 1000 bootstrap sample first thresholds.

[6]This issue ultimately foreshadows the need for an iterative use of this procedure, which will be highlighted in Section 3.2.

The three thresholds resulting from the above procedure, which we will henceforth refer to as the low potential, prevalence, and high potential thresholds, are given in Table 2 below. Hence, $T_1 = 0.05679882$, $T_2 = 0.24071816$, and $T_3 = 0.51923901$. While it is standard for the three levels of the protected attribute to have identical values across the three thresholds, we present the thresholds in this way (as opposed to simply providing three values) to enable a more direct comparison with the thresholds that will be obtained from the fairness correction procedure in Section 2.4. It should be noted that, as far as we are aware, it is not standard to use such a bagging algorithm to obtain potential (i.e. risk) score thresholds. We have implemented such procedures to enable direct comparison with the thresholds of our fairness correction procedure, which utilizes an appropriate resampling method (e.g. bagging) to reduce the variability in the effectiveness of the subsequent thresholds on unseen data.

Table 2: *Bootstrap aggregated (i.e. bagged) pre-fairness corrected thresholds obtained from the training set.*

Threshold	Protected Attribute Level		
	ANO	APW	BLK
Low Potential	0.05679882	0.05679882	0.05679882
Prevalence	0.24071816	0.24071816	0.24071816
High Potential	0.51923901	0.51923901	0.51923901

Using the logistic regression model that was fit to the training set data, the predicted probability of making more than $\$50,000$ annually was then obtained for each of the 16,281 individuals comprising the test set that is readily available with the *Adult Data Set*. Via the thresholds of the bagging procedure, which are given in Table 2.4, potential (i.e. risk) scores were then assigned to each individual in the test set. In particular, any individual with a predicted probability less than or equal to the low potential threshold (0.05679882) is assigned a score of 1; any individual with a predicted probability greater than the low potential threshold, but less than or equal to the prevalence threshold (0.24071816), is assigned a score of 2; any individual with a predicted probability greater than the prevalence threshold, but less than or equal to the high potential threshold (0.51923901) is assigned a score of 3; and any individual with a potential score greater than the high potential threshold is assigned a score of 4.

To assess the quality of the scoring system on the test set, we consider a number of performance-related measures at each of the three thresholds. Table 8 provides the accuracy, positive predicted value, negative predicted value, sensitivity, and specificity of the scoring system at each of the three thresholds when applied to the test set. While much could be said about what these values indicate about the scoring system, we will provide three observations based on this table which suggest that the scoring system is performing reasonably well given that it was trained using only three covariates.

Table 3: *The accuracy (ACC), positive predicted value (PPV), negative predicted value (NPV), sensitivity, and specificity of the potential (i.e. risk) scoring system at each of the three pre-fairness corrected thresholds. These performance-related measures are for the test set.*

Threshold	ACC	PPV	NPV	Sensitivity	Specificity
Low Potential	0.3308	0.2597	0.9771	0.9904	0.1267
Prevalence	0.6868	0.4032	0.8741	0.6789	0.6893
High Potential	0.7857	0.6103	0.8051	0.2569	0.9493

First, among the individuals who are labeled as low potential for making greater than $\$50,000$ annually (i.e. $S = 1$) 97.71% ultimately do not make greater than $\$50,000$ annually (i.e. NPV at Low Potential Threshold is 0.9771). Furthermore, among the individuals who do not make greater than $\$50,000$ annually (i.e. $Y = 0$), 12.67% are labeled as low potential for making greater than $\$50,000$ annually (i.e. Specificity at Low Potential Threshold is 0.1267). Hence, a potential score of 1 captures a non-marginal set of individuals who in fact do not make more than $\$50,000$ annually and individuals that receive such a score are rarely mislabeled. Second, among the individuals which are labeled as high potential for making greater than $\$50,000$ annually (i.e. $S = 4$), 61.03% ultimately do make greater than $\$50,000$ annually (i.e. PPV at High Potential Threshold is 0.6103). Additionally, among the individuals who do make greater than $\$50,000$ annually (i.e. $Y = 1$), 25.69% are labeled as high potential for making greater than $\$50,000$ annually (i.e. Sensitivity at High Potential Threshold is 0.2569). In other words, a potential score of 4 captures a non-marginal set of individuals who in fact do make more than $\$50,000$ annually and, relative to the baseline (i.e. test set prevalence) of 23.6%, individuals that receive such a score are much more likely to make more than $\$50,000$ annually. Third, and finally, at the prevalence threshold, where scores of 1 and 2 are separated from scores of 3 and 4, the accuracy of the model is 68.68%; in other words, 68.68% of the time the model will either 1) correctly assign an individual who makes less than or equal to $\$50,000$ annually a score of 1 or 2, or 2) correctly assign an individual who makes greater than $\$50,000$ annually a score of 3 or 4. Hence, at this threshold, the scoring system does a reasonable job of separating those who make more than $\$50,000$ annually from those who do not.

In light of these performance measures, the model and subsequent scoring system would traditionally be deemed ready for implementation. However, in light of the growing awareness surrounding the propensity of such models to perpetuate unfairness, the fairness of such scores should be assessed before any decision regarding their implementation is ultimately made. To that end, the false positive and false negative error rates for each level of the protected attribute at each of the three thresholds are provided in Table 4. If the scoring system were fair according to the error rate balance criteria, then the three values within any row would be roughly equivalent; this is clearly not the case for the current scoring system. For example, at the prevalence threshold, 7.2% of individuals in the ANO level who do not make more than $\$50,000$ are mislabeled with a score of 3 or 4, whereas the same is true for 10.27% of individuals in the BLK level and 34.29% of individuals in the APW level. In other words, via equation 7, the false positive error rate for the ANO level is roughly one-fifth the false positive error rate of the APW level (i.e. $\text{FPR}_{\text{ANO,APW}}(T_2) = 0.2100$). Similarly, also at the prevalence threshold, 68.18% of individuals in the ANO level who do make more than $\$50,000$ are mislabeled with a score of 1 or 2, whereas the same is true for 56.98% of individuals in the BLK level and 30.44% of individuals in the APW level. Hence, via equation 10, individuals in the APW level, relative to individuals in the ANO level, are less than half as likely to be mislabeled with a score of 1 or 2 (i.e. $\text{FNR}_{\text{ANO,APW}}(T_2) = 0.4465$). The remaining 10 pairwise ratios of error rates, encompassing all three thresholds, are provided in Table 5, along with the overall fairness measure at each of the three thresholds (i.e. $\text{ERB}(T_1)$, $\text{ERB}(T_2)$, and $\text{ERB}(T_3)$). As all three overall fairness measures are considerably far from 0.8., let alone 1.0, it is clear that the potential scoring system is not fair according to the error rate balance criteria. In other words, although the predictive performance of the scoring system is reasonably good, the unfairness perpetuated by the scoring system is unacceptable.

2.4 Fairness Correction Procedure

Recall that the objectives of our fairness correction procedure are to update the potential (i.e. risk) scores obtained in Section 2.3 so that 1) the overall fairness measure at each threshold increases

Table 4: *False positive and false negative error rates within the test set for all three levels of the protected attribute (ANO, APW, and BLK) at each of the three pre-fairness corrected thresholds (Low Potential, Prevalence, High Potential).*

Threshold	Error Rate Type	Protected Attribute Level		
		ANO	APW	BLK
Low Potential	False Positive	0.5520	0.9025	0.7026
Low Potential	False Negative	0.0455	0.0086	0.0223
Prevalence	False Positive	0.0720	0.3429	0.1027
Prevalence	False Negative	0.6818	0.3044	0.5698
High Potential	False Positive	0.0120	0.0573	0.0065
High Potential	False Negative	0.9773	0.7292	0.9665

Table 5: *The pairwise ratios of false positive and false negative error rates within the test set for all three pairings of the three levels of the protected attribute (i.e. ANO/APW, ANO/BLK, and APW/BLK) at each of the three pre-fairness corrected thresholds (Low Potential, Prevalence, High Potential), are provided. Additionally, for each of the three thresholds, the subsequent overall (un)fairness measure (i.e. ERB(T_1), ERB(T_2), and ERB(T_3)) is provided.*

Threshold	Error Rates Ratio Type	Pairing of Protected Attribute Levels			Fairness Measure
		ANO/APW	ANO/BLK	APW/BLK	
Low Potential	FPR(T_1)	0.6116	0.7857	0.7785	ERB(T_1) = 0.1882
Low Potential	FNR(T_1)	0.1882	0.4916	0.3829	
Prevalence	FPR(T_2)	0.2100	0.7007	0.2997	ERB(T_2) = 0.2100
Prevalence	FNR(T_2)	0.4465	0.8358	0.5343	
High Potential	FPR(T_3)	0.2094	0.5427	0.1137	ERB(T_3) = 0.1137
High Potential	FNR(T_3)	0.7462	0.9890	0.7545	

toward 0.80 and 2) the predictive performance of the model is not substantially compromised. To accomplish this, our procedure ultimately searches for new thresholds which achieve fairness and are as close as possible to the original (i.e. pre-fairness corrected) thresholds. It should be noted that our procedure does not require the thresholds to be the same across all levels of the protected attribute and therefore each of the three fairness corrected thresholds is really a threshold combination. Utilizing the 32,561 predicted probabilities from the logistic regression model that was fit to the training set, the fairness corrected thresholds are obtained via the following bootstrap procedure[7]:

0. Via the fitted logistic regression model, generate the predicted probability of making more than $\$50,000$ annually for each individual in the training set;

1. Randomly sample, with replacement, 32,561 individuals from the training set;

2. Identify the 10^{th} percentile among the predicted probabilities in the sample. This represents the pre-fairness corrected low potential threshold for the sample and will be denoted by T_1^*;

[7]To facilitate understanding, this procedure is presented as a brute-force approach, which is computationally expensive. In practice, for each bootstrap sample, we employ an iterative approach with the described procedure which sequentially reduces the size of the solution space and subsequently reduces the computational expense. As we do not currently have a general understanding of the convexity, or lack thereof, of the solution space, we have not attempted to employ a general optimization procedure such as gradient descent.

3. Determine the prevalence of the outcome variable (i.e. proportion of individuals with $Y = 1$) in the sample. This represents the pre-fairness corrected prevalence threshold for the sample and will be denoted by T_2^*;

4. Identify the 90^{th} percentile among the predicted probabilities in the sample. This represents the pre-fairness corrected high potential threshold for the sample and will be denoted by T_3^*;

5. Let P_{ANO}, P_{APW}, and P_{BLK} denote, respectively, the set of predicted probabilities in the sample whose corresponding value for the protected attribute is ANO, APW, and BLK. Now construct all possible combinations of three predicted probabilities in which one predicted probability comes from each of P_{ANO}, P_{APW}, and P_{BLK}. We will refer to this collection of predicted probabilities as the initial set of possible fairness corrected threshold combinations. Let $C = (\hat{p}_{ANO},\ \hat{p}_{APW},\ \hat{p}_{BLK})$ denote an arbitrary threshold combination, where $\hat{p}_{ANO} \in P_{\text{ANO}}$, $\hat{p}_{APW} \in P_{\text{APW}}$, and $\hat{p}_{BLK} \in P_{\text{BLK}}$.

6. For each combination, C, in the set of possible fairness corrected threshold combinations, calculate the overall fairness measure, $\text{ERB}(C)$. Discard threshold combinations which do not meet a pre-specified fairness tolerance; e.g. discard any C for which $\text{ERB}(C) < 0.80$. The threshold combinations which remain (i.e. all C for which $\text{ERB}(C) \geq 0.80$) will henceforth be referred to as the set of viable fairness corrected threshold combinations[8].

7. For each combination, C, in the set of viable fairness corrected threshold combinations, calculate the Euclidean distance of that threshold combination from each of T_1^*, T_2^*, and T_3^*:

$$d(C, T_1^*) = \sqrt{(\hat{p}_{ANO} - T_1^*)^2 + (\hat{p}_{APW} - T_1^*)^2 + (\hat{p}_{BLK} - T_1^*)^2} \tag{14}$$

$$d(C, T_2^*) = \sqrt{(\hat{p}_{ANO} - T_2^*)^2 + (\hat{p}_{APW} - T_2^*)^2 + (\hat{p}_{BLK} - T_2^*)^2} \tag{15}$$

$$d(C, T_3^*) = \sqrt{(\hat{p}_{ANO} - T_3^*)^2 + (\hat{p}_{APW} - T_3^*)^2 + (\hat{p}_{BLK} - T_3^*)^2} \tag{16}$$

8. The fairness corrected low potential threshold combination for the sample is then the threshold combination, C, with the smallest value outputted by equation 14. In the unlikely event of a tie between multiple combinations, choose the combination with the smallest negative predicted value.

9. The fairness corrected prevalence threshold combination for the sample is then the threshold combination, C, with the smallest value outputted by equation 15. In the unlikely event of a tie between multiple combinations, choose the combination with the greatest accuracy.

10. The fairness corrected high potential threshold combination for the sample is then the threshold combination, C, with the smallest value outputted by equation 16. In the unlikely event of a tie between multiple combinations, choose the combination with the largest positive predicted value.

11. Repeat steps 1-10 an additional 999 times;

12. Obtain the bagged fairness corrected threshold combinations by aggregating (i.e. averaging) the 1000 sets of fairness corrected thresholds. For example, the fairness corrected low potential threshold combination is the average of the 1000 bootstrap sample fairness corrected low potential threshold combinations.

[8] If the the fairness tolerance is not the same at each of the three thresholds, then the discard set will be different for each threshold.

The three fairness corrected threshold combinations resulting from the bagging procedure described above are given in Table 6. Let the threshold combinations at the low potential, prevalence, and high potential thresholds be denoted by C_1, C_2, and C_3, respectively. In assessing these threshold combinations and in comparing them with the pre-fairness corrected thresholds in Table 2, a clear pattern emerges[9]. Across all three thresholds, the post-fairness corrected ANO and BLK thresholds are lower than the pre-fairness corrected thresholds, while the post fairness corrected APW thresholds are higher than the pre-fairness corrected thresholds. Such shifting of thresholds seems logical in light of the disparity index values reported in Table 1. More specifically, as individuals in the APW level (i.e. AP and WT levels) of the data set are much more likely to make more than $\$50,000$ annually than individuals in the ANO (i.e. AN and OT levels) or BLK (i.e. BL level) levels of the data set, it seems appropriate that the post-fairness corrected thresholds, relative to the pre-fairness corrected thresholds, would require more evidence of APW individuals, and less evidence of ANO and BLK individuals, before increasing their score.

Table 6: *Bootstrap aggregated (i.e. bagged) post-fairness corrected threshold combinations obtained from the training set.*

Threshold	Protected Attribute Level		
	ANO	APW	BLK
Low Potential	0.04388943	0.1006876	0.05047212
Prevalence	0.15580716	0.3225541	0.17930142
High Potential	0.43867172	0.6785131	0.43083013

To obtain the fairness corrected potential (i.e. risk) scores for the test set, first recall that the logistic regression model that was fit to the training set data was already used to obtain the predicted probability of making more than $\$50,000$ annually for each of the 16,281 individuals comprising the test set. Using these predicted probabilities and their corresponding level of the protected attribute, the fairness-corrected threshold combinations can then be employed to assign a potential score to each individual in the test set. For example, among individuals in the ANO level, any person with a predicted probability less than or equal to the corresponding low potential threshold (0.04388943) is assigned a score of 1; any individual with a predicted probability greater than that low potential threshold, but less than or equal to the corresponding prevalence threshold (0.15580716), is assigned a score of 2; any individual with a predicted probability greater than that prevalence threshold, but less than or equal to the corresponding high potential threshold (0.43867172) is assigned a score of 3; and any individual with a predicted probability greater than that high potential threshold is assigned a score of 4. For individuals in the APW and BLK levels, scores were assigned analogously using the respective threshold combinations delineated in Table 6.

To visualize the differences in the assigned scores between the pre- and post-fairness corrected thresholds across the three levels of the protected attribute, consider Figure 1[10]. The three horizontal maroon lines correspond to the pre-fairness corrected thresholds and the four facets correspond to the assignment of scores (i.e. $S = 1, S = 2, S = 3, S = 4$) under the post-fairness corrected threshold combinations. For each score (i.e. within each facet), the predicted probability of every

[9]To enable direct comparison of the two sets of thresholds, the same set of bootstrap samples that was used to obtain the bagged pre-fairness corrected thresholds was also used to obtain the bagged post-fairness corrected threshold combinations.

[10]Jason Wallin, another researcher within our research and analytics team at ORRAI, designed and developed this graphic to accompany the fairness correction procedure.

individual in the test set who received that score under the post-fairness corrected thresholds is provided, broken down by level of the protected attribute. For example, across the four facets, every point below the first maroon line ($T_1 = 0.05679882$) represents an individual who was assigned a score of 1 under the pre-fairness corrected thresholds. In inspecting the first facet ($S = 1$), however, we see that a large number of APW individuals who were assigned a score of 2 under the pre-fairness corrected thresholds are now assigned a 1 under the post-fairness corrected thresholds. Furthermore, in inspecting the second facet ($S = 2$), we see that a small number of ANO and BLK individuals who were assigned a score of 1 under the pre-fairness corrected thresholds are now assigned a score of 2 under the post-fairness corrected thresholds. Analogous patterns exist across the prevalence and high potential scores as well. Hence, from this figure, it is readily apparent that relative to the pre-fairness corrected thresholds, the post-fairness corrected threshold combinations require more evidence from the APW individuals, and less evidence from the ANO and BLK individuals, before increasing the corresponding score.

In addition to understanding how the scores differ between the two systems across each level of the protected attribute, it is also worth assessing how the overall distribution of scores has changed under the post-fairness corrected scoring system. To that end, the frequency and relative frequency of the potential (i.e. risk) scores under both the pre- and post-fairness corrected scoring systems are provided in Table 7. From this table it is evident that the number of individuals assigned a score of 2 or 3 changes relatively little between the pre- and post-fairness corrected scoring systems. However, it is also evident that a score of 4 is much less common, while a score of 1 is much more common, under the post-fairness corrected scoring system. Such changes seem logical given that the APW level of the protected attribute represents the vast majority of individuals in the data set and that the post-fairness corrected scoring system requires more evidence of APW individuals before increasing their corresponding potential score. Now that the differences in assigned scores between the pre- and post-fairness corrected thresholds are evident, the next step is to evaluate whether these post-fairness corrected threshold combinations in fact meet the two objectives of the endeavor.

Table 7: *The frequency and relative frequency of the four potential (i.e. risk) scores under both the pre- and post-fairness corrected scoring systems.*

Threshold Type	$S = 1$	$S = 2$	$S = 3$	$S = 4$
Pre-Fairness Corrected	1613	8193	4856	1619
	0.099	0.503	0.298	0.099
Post-Fairness Corrected	3264	8253	4235	529
	0.200	0.507	0.261	0.033

2.5 Post-Fairness Corrected Results

To assess the performance of the fairness corrected thresholds, the accuracy, negative predicted value, positive predicted value, sensitivity, and specificity of the post-fairness corrected scores at all three threshold combinations are provided in Table 8. In comparing these performance measures with those of the pre-fairness corrected scores (Table 3), several trade-offs are evident. First, at each threshold, the post-fairness corrected scores, relative to the pre-fairness corrected scores, have better specificity but reduced sensitivity. For example, at the prevalence threshold under the pre-fairness corrected scores, 68.96% of individuals who do not make more than $\$50,000$ annually were assigned either a score of 1 or 2 and 67.89% of individuals who do make more than $\$50,000$ annually

were assigned either a score of 3 or 4. However, at the prevalence threshold under the post-fairness corrected scores, these values shift to 79.32% and 56.99%, respectively. Second, at each threshold, the post-fairness corrected scores, relative to the pre-fairness corrected scores, tend to have slightly improved positive predicted values but slightly degraded negative predicted values. For example, at the low potential threshold under the pre-fairness corrected scores, 25.97% of individuals assigned a score of 2, 3, or 4 make more than $\$50,000$ annually and 97.71% of individuals assigned a score of 1 do not make more than $\$50,000$ annually. However, at the low potential threshold under the post-fairness corrected scores, these values shift to 28.64% and 96.39%, respectively. Finally, at each threshold, the post-fairness corrected scores, relative to the pre-fairness corrected scores, tend to have better or only slightly worse accuracy. For example, at the prevalence threshold under the pre-fairness corrected scores, 68.68% of the time the model will either 1) correctly assign an individual who makes less than or equal to $\$50,000$ annually a score of 1 or 2, or 2) correctly assign an individual who makes greater than $\$50,000$ annually a score of 3 or 4. However, at the high potential threshold under the post-fairness corrected scores, this value shifts to 74.04%. Hence, while there are trade-offs in the various performance measures between the pre- and post-fairness corrected scores, we would argue that the predictive performance of the model does not appear to be compromised in any substantive way as a result of the fairness correction procedure.

Table 8: *The accuracy (ACC), positive predicted value (PPV), negative predicted value (NPV), sensitivity, and specificity of the potential (i.e. risk) scoring system at each of the three post-fairness corrected thresholds. These performance-related measures are for the test set.*

Threshold	ACC	PPV	NPV	Sensitivity	Specificity
Low Potential	0.4222	0.2864	0.9639	0.9693	0.2530
Prevalence	0.7404	0.4601	0.8564	0.5699	0.7932
High Potential	0.7715	0.6182	0.7766	0.0850	0.9838

Having addressed the impact of the fairness correction procedure on predictive performance, we now address the corresponding impact on fairness. To that end, the false positive and false negative error rates for each level of the protected attribute at each of the three fairness-corrected thresholds are provided in Table 9. While the error rates across any row are still not equal, they are less disparate. For example, at the fairness-corrected prevalence threshold, 15.2% of individuals in the ANO level who do not make more than $\$50,000$ are mislabeled with a score of 3 or 4, whereas the same is true for 19.54% of individuals in the BLK level and 20.96% of individuals in the APW level. Recall that these mislabeling rates were 7.2%, 10.27%, and 34.29%, respectively, under the pre-fairness corrected thresholds. Hence, the post-fairness corrected prevalence threshold has achieved more comparable false positive error rates across the levels of the protected attribute by increasing the false positive error rates for the ANO and BLK levels and decreasing the false positive error rate for the APW level. Similarly, also at the fairness-corrected prevalence threshold, 52.27% of individuals in the ANO level who do not make more than $\$50,000$ are mislabeled with a score of 3 or 4, whereas the same is true for 48.6% of individuals in the BLK level and 42.62% of individuals in the APW level. Recall that these mislabeling rates were 44.65%, 53.43%, and 83.58%, respectively, under the pre-fairness corrected thresholds. Hence, the post-fairness corrected prevalence threshold has achieved more comparable false negative error rates across the levels of the protected attribute by decreasing the false negative error rates for the ANO and BLK levels and increasing the false negative error rate for the APW level. Analogous differences can be seen in the false positive and false negative error rates between the pre- and post-fairness corrected scores for both the low potential and high potential thresholds.

Table 9: *False positive and false negative error rates of the test set for all three levels of the protected attribute (ANO, APW, and BLK) at each of the three post-fairness corrected thresholds (Low Potential, Prevalence, High Potential).*

Threshold	Error Rate Type	Protected Attribute Level		
		ANO	APW	BLK
Low Potential	False Positive	0.6640	0.7476	0.7576
Low Potential	False Negative	0.0455	0.0309	0.0223
Prevalence	False Positive	0.1520	0.2096	0.1954
Prevalence	False Negative	0.5227	0.4262	0.4860
High Potential	False Positive	0.0200	0.0162	0.0159
High Potential	False Negative	0.8864	0.9161	0.8994

To quantify the fairness of the post-fairness corrected scores for the test set, the pairwise ratios of false positive and false negative error rates for all three thresholds, as well as the overall fairness measure at each threshold, are provided in Table 10. To better compare these results with those of the pre-fairness corrected scores (see Table 5), consider Figure 2, which provides, respectively, the six pairwise ratios of error rates, pre- and post-fairness corrected, at each of the three thresholds. Note that each line connecting a particular pre-fairness corrected pairwise ratio of error rates to the corresponding post-fairness corrected ratio is artificial and is only included to ease the visual assessment of the change in the respective ratio as a result of the fairness-correction procedure. From this figure, two desirable shifts are evident. First, 16 of the 18 "lines" have a positive slope, with the other two lines having a slope essentially equal to zero. In other words, the fairness correction procedure tends to reduce unfairness (i.e. increase fairness) across all possible pairings of protected attribute levels. For example, at the prevalence threshold, the pre-fairness corrected ratio of false positive error rates between the ANO and APW levels is 0.2100, whereas under the post-fairness corrected scoring system this ratio increases to 0.7253. Second, the lowest point for each of the three post-fairness corrected thresholds is higher than the lowest point for each of the three pre-fairness corrected thresholds. In other words, the overall fairness measure at each of the three thresholds is closer to 0.80 under the post- versus pre-fairness corrected scoring system. For example, at the high potential threshold, under the pre-fairness corrected scoring system, the overall fairness measure is 0.1137, whereas under the post-fairness corrected scoring system, the overall fairness measure increases to 0.7960. These two shifts indicate that the post-fairness corrected scoring system, according to the error rate balance criteria of fairness, is unquestionably fairer than the pre-fairness corrected scoring system. In other words, for this data set, the proposed fairness-correction procedure has substantially reduced unfairness in the potential scoring system without substantively impacting its predictive performance.

Before moving on to Section 3, it is worth noting here that the fairness tolerance of 0.80 at each threshold ensures that the fairness-corrected thresholds obtained for each bootstrap sample result in overall fairness measures greater than or equal to 0.80 when these thresholds are applied to the predicted probabilities of that **same** bootstrap sample. However, these tolerances do not ensure that the aggregated (via averaging) thresholds across these bootstrap samples will result in overall fairness measures greater than or equal to 0.80 when these thresholds are applied to the predicted probabilities of the entire (i.e. non bootstrapped sampled) training set, let alone the test set. This then begs the question: why employ a bagging procedure when simply finding the fairness corrected thresholds one time using the entire set of predicted probabilities (as opposed to using 1000 bootstrap samples of these predicted probabilities) would at least ensure that the overall fairness

Table 10: *The pairwise ratios of false positive and false negative error rates for the test set for all three pairings of the three levels of the protected attribute (i.e. ANO/APW, ANO/BLK, and APW/BLK) at each of the three post-fairness corrected thresholds (Low Potential, Prevalence, High Potential) are provided. Additionally, for each of the three thresholds, the subsequent overall fairness measure (i.e. ERB(C_1), ERB(C_2), and ERB(C_3)) is provided.*

Threshold	Error Rates Ratio Type	Pairing of Protected Attribute Levels ANO/APW	ANO/BLK	APW/BLK	Fairness Measure
Low Potential	FPR(C_1)	0.8882	0.8765	0.9868	ERB(C_1) = 0.4916
Low Potential	FNR(C_1)	0.6801	0.4916	0.7229	
Prevalence	FPR(C_2)	0.7253	0.7780	0.9322	ERB(C_2) = 0.7253
Prevalence	FNR(C_2)	0.8153	0.9298	0.8768	
High Potential	FPR(C_3)	0.8100	0.7960	0.9827	ERB(C_3) = 0.7960
High Potential	FNR(C_3)	0.9676	0.9855	0.9818	

measures are greater than or equal to 0.80 for the predicted probabilities of the training set? The answer, in short, is that a more robust set of fairness-corrected thresholds are obtained by employing a bagging procedure. To elaborate, consider Figure 3 in which the three histograms in the top graphic represent the bootstrap distributions of the fairness corrected prevalence thresholds based on the 1000 bootstrap samples, while the histogram on the bottom represents the corresponding distribution of the overall fairness measure when those 1000 prevalence thresholds are applied to the test set. From these distributions it is evident that the fairness-corrected thresholds can vary substantially depending on the set of predicted probabilities used to obtain those thresholds and, consequently, the corresponding value of the overall fairness measure, when those thresholds are applied to unseen data, can also vary widely. Hence, the objective of the bagging procedure is to reduce the variability of the overall fairness measure, when applied to unseen data, by using the average of the fairness corrected thresholds obtained across the 1000 bootstrap samples.

3 Conclusion

Within the context of the *Adult Data Set* from the UCI Machine Learning Repository, we simultaneously developed and demonstrated our proposed extension of a thresholding scheme introduced in Lipton et. al. [7] for reducing unfairness in machine learning algorithms featuring a binary outcome. With this extension, the protected attribute can have an arbitrary number of levels and the resulting fairness corrected risk scores can have two[11] or more levels. When applied to the *Adult Data Set* under a four-level risk scoring system and a three-level protected attribute, our proposed extension both increased fairness and avoided substantively compromising the predicted performance of the scoring system. More specifically, the overall fairness measure, as assessed via the error rate balance criteria, increased from 0.1882 to 0.4916 at the first threshold, from 0.2100 to 0.7253 at the second threshold, and from 0.1137 to 0.7960 at the third threshold. Furthermore, while the pre-fairness corrected scores at each threshold, relative to the post-fairness corrected scores, tend to have greater sensitivity (e.g. 0.26 to 0.08 at prevalence threshold) and marginally better negative predicted value (e.g. 0.98 to 0.96 at low risk threshold), they also tend to have lesser specificity (e.g. 0.13 to 0.25 at low risk threshold) and marginally worse positive predicted value (e.g. 0.40 to 0.46 at prevalence threshold) and accuracy (e.g. 0.68 to 0.74 at prevalence

[11]If the risk scoring system only has two levels, then the subsequent scoring system is analogous to a binary classification task.

threshold). Hence, while there certainly exist differences in predicted performance between the pre- and post-fairness corrected scores, those differences do not add up to a substantially worse predictive model, especially in light of the improvements pertaining to fairness.

3.1 Future Simulation Studies

While the proposed fairness correction procedure has proven to be effective with both the salary data set utilized in this paper and the initial machine learning endeavors of ORRAI, there are a number of simulation studies that, if conducted, could provide substantive information about its performance more generally. We propose two such studies here. First, a simulation study could be used to explore the performance of the proposed fairness correction procedure under varying distributions of the predicted probabilities produced by the learner. More specifically, in our applications of the procedure thus far, the distribution of the predicted probabilities has been unimodal and right skewed. Hence, it would be useful to understand how the procedure performs under different realizations of this distribution and, in particular, when this distribution is bimodal. Second, a simulation study could be used to better understand the distribution of the overall fairness measure under varying magnitudes of unfairness. Of particular interest with such a study would be the distribution of the overall fairness measure when no unfairness exists. Under such a scenario, the framework for using this procedure to test for unfairness in the assigned scores could be developed in which the null hypothesis asserts that the subsequent scoring system is fair at a particular threshold. Regardless of the nature of the simulation study, we hope it is clear that while we have found the procedure to be consistently effective thus far, we do anticipate situations where the procedure will be less than effective and we believe simulation studies are a means to more generally understand the scenarios in which its utility holds.

3.2 A Practical Limitation and a Point of Emphasis

Having demonstrated the effectiveness of the proposed fairness correction procedure and having discussed possible simulation studies, we now address the primary practical limitation of the method. In particular, while in theory the procedure can have an arbitrary number of levels for both the protected attribute and the risk scoring system, the size of the data set will likely limit how many levels of each can practically be considered. More specifically, for each combination of risk score level and protected attribute level, there must be sufficiently many individuals in order for the procedure to work effectively. In our experience, a lower bound for each such combination is around 500 individuals, though the magnitude of the unfairness certainly impacts this tentative lower bound with greater unfairness requiring larger sample sizes. In the event of insufficient sample sizes at any such combination, our approach has been at least one of the following: 1) reduce the number of levels in the scoring system or 2) combine levels of the protected attribute.

To conclude this paper we want to emphasize the iterative nature of the proposed fairness correction procedure. More specifically, once the fairness corrected scoring system is implemented, the historically biased data that was leveraged to initially train the model will grow to include these newer and fairer data points. Hence, the difference in thresholds across the levels of the protected attribute should decrease over time allowing the fairness corrected thresholds to be re-assessed and re-assigned, perhaps in tandem with an update or re-training of the model. In this way, the overall fairness measure could, over time, be edged closer to one at each threshold and the levels of the protected attribute broadened. While both of these eventualities are desirable, two subsequent possibilities of the latter eventuality are worth distinguishing here. First, this could enable the disaggregation of a current protected attribute level into its corresponding component

levels, such as disaggregating the APW level into its Asian, Pacific Islander and White component levels. Second, it could enable the creation of a multidimensional protected attribute which is a combination of two or more protected attributes, such as race and gender. In such a combination, for example, the protected attribute levels could be Female-ANO, Male-ANO, Other-ANO, Female-APW, Male-APW, Other-APW, Female-BLK, Male-BLK, and Other-BLK. Regardless, we hope that it is evident that this fairness correction procedure is not a one-and-done solution, but rather an instrument that is continually employed to eventually find and subsequently maintain equitable thresholds in machine-learning-generated risk scores.

Acknowledgments

We would like to thank Alexandra Chouldechova of Carnegie Mellon University for meeting with us to discuss the thresholding scheme presented in [7]. This conversation was an essential first step in ORRAI's development of the fairness correction procedure presented in this paper. Additionally, we would like to thank the entire research and analytics team within ORRAI, and in particular Kevin Hamler-Dupras and Maria Duryea, for their frequent input throughout the development of this procedure.

References

[1] Alekh Agarwal et al. "A Reductions Approach to Fair Classification". In: *Proceedings of FAT ML, Halifax, Nova Scotia, Canada, 2017*. 2017.

[2] Julia Angwin et al. *Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks.* 2016. URL: `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

[3] Scott Barsotti. *Can We Automate Fairness? Professor Alexandra Chouldechova On Machine Learning and Discrimination.* 2017. URL: `https://www.heinz.cmu.edu/media/2017/january/automate-fairness-machine-learning-discrimination`.

[4] Richard Berk et al. "Fairness in Criminal Justice Risk Assessments: The State of the Art". In: *arXiv preprint arXiv:1703.09207* (2017).

[5] Alexandra Chouldechova. "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments". In: *arXiv preprint arXiv:1703.00056* (2017).

[6] Dua Dheeru and Efi Karra Taniskidou. *UCI Machine Learning Repository*. 2017. URL: `http://archive.ics.uci.edu/ml`.

[7] Zachary C. Lipton, Alexandra Chouldechova, and Julian McAuley. "Does mitigating ML's disparate impact require disparate treatment?" In: *arXiv preprint arXiv:1711.07076* (2017).

[8] Muhammad Bilal Zafar et al. "Fairness Constraints: Mechanisms for Fair Classification". In: *arXiv preprint arXiv:1507.05259* (2017).

Figure 1: *The differences in assigned potential (i.e. risk) scores between the pre- and post-fairness corrected thresholds are represented. The three horizontal maroon lines correspond to the pre-fairness corrected thresholds and the four facets correspond to the assignment of scores (i.e.* $S = 1, S = 2, S = 3, S = 4$*) under the post-fairness corrected threshold combinations. For each score (i.e. within each facet), the predicted probability of every individual in the test who received that score under the post-fairness corrected thresholds is provided, broken down by level of the protected attribute. This illustration indicates that relative to the pre-fairness corrected thresholds, the post-fairness corrected threshold combinations require more evidence from the APW individuals, and less evidence from the ANO and BLK individuals, before increasing the corresponding score.*

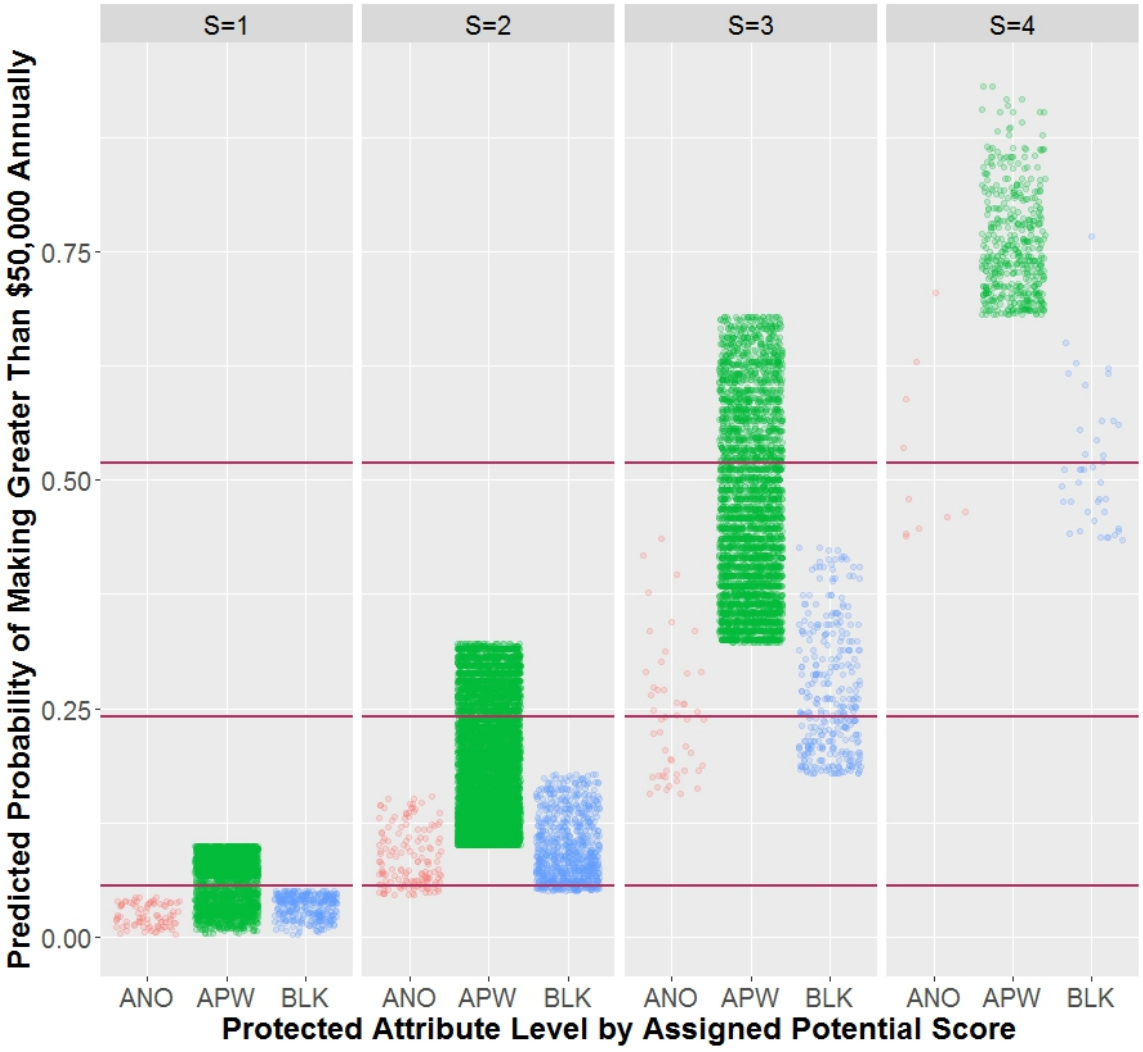

Figure 2: *The pre- and post-fairness corrected pairwise error rate ratios for the test set across the three pairings of levels of the protected attribute (ANO-APW, ANO-BLK, and APW-BLK) are displayed for the high potential threshold (top graphic), prevalence threshold (middle graphic), and low potential threshold (bottom graphic). The solid lines correspond to false positive error rates, while the dashed lines correspond to false negative error rates. The solid black horizontal line corresponds to a desired error rate ratio of 0.80.*

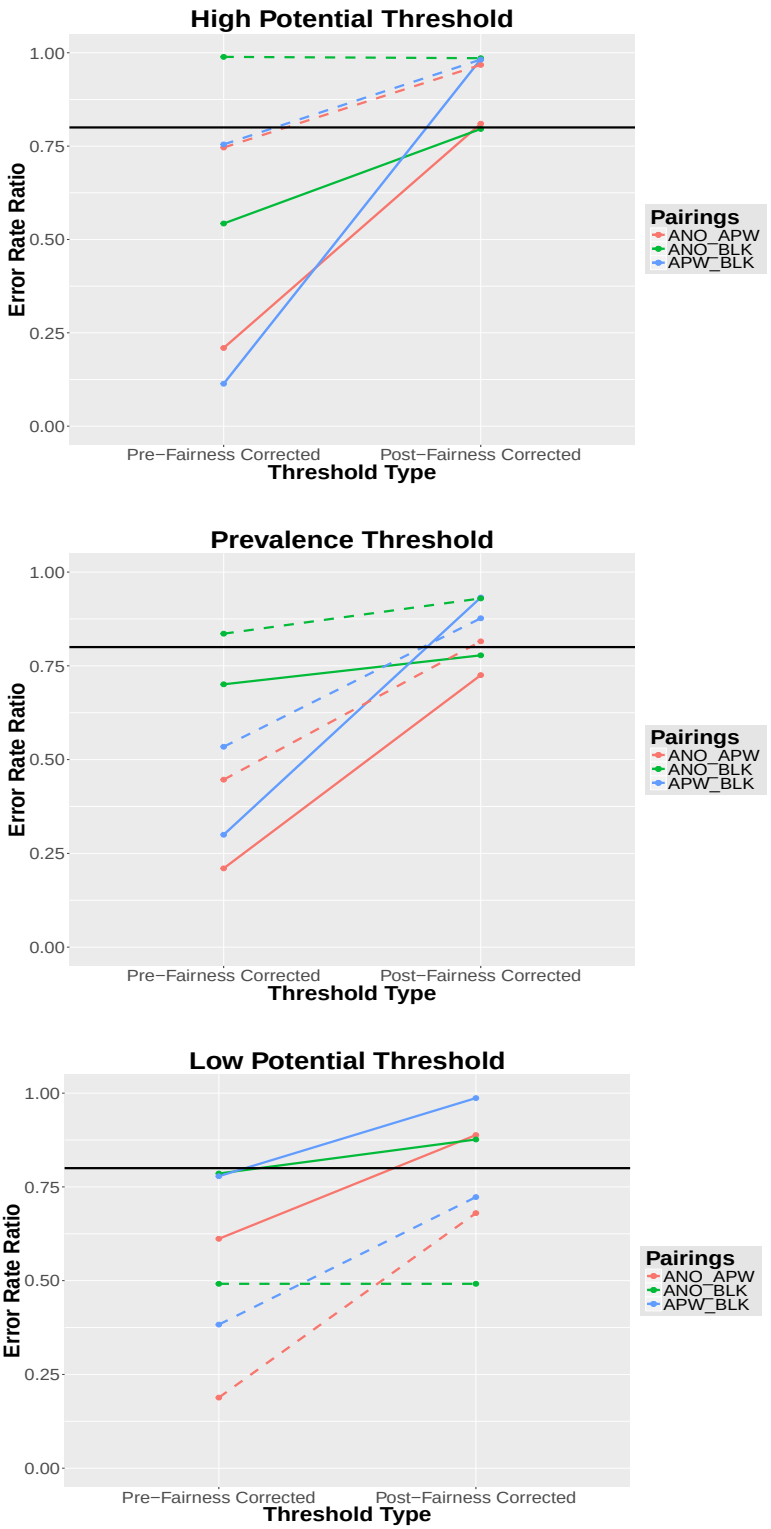

Figure 3: *The bootstrap distribution of the 1000 fairness corrected prevalence threshold combinations for each of the three levels of the protected attribute (ANO, APW, and BLK) are displayed in the top graphic. These 1000 prevalence threshold combinations were then applied to the test set; the resulting bootstrap distribution of 1000 overall fairness measures at the prevalence threshold are displayed in the bottom graphic.*