
xOrder: A Model Agnostic Post-Processing Framework for Achieving Ranking Fairness While Maintaining Algorithm Utility

Sen Cui^{1*} Weishen Pan^{1*} Changshui Zhang¹ Fei Wang²

¹Institute for Artificial Intelligence, Tsinghua University (THUIAI),
State Key Lab of Intelligent Technologies and Systems,
Department of Automation, Tsinghua University, Beijing, China

²Department of Healthcare Policy and Research, Weill Cornell Medicine
Cornell University, New York, NY 10065
cuis19@mails.tsinghua.edu.cn, pws15@mails.tsinghua.edu.cn
zcs@mail.tsinghua.edu.cn, few2001@med.cornell.edu

Abstract

Algorithmic fairness has received lots of interests in machine learning recently. In this paper, we focus on the bipartite ranking scenario, where the instances come from either the positive or negative class and the goal is to learn a ranking function that ranks positive instances higher than negative ones. In an unfair setting, the probabilities of ranking the positives higher than negatives are different across different protected groups. We propose a general post-processing framework, xOrder, for achieving fairness in bipartite ranking while maintaining the algorithm classification performance. In particular, we optimize a weighted sum of the utility and fairness by directly adjusting the relative ordering across groups. We formulate this problem as identifying an optimal warping path across different protected groups and solve it through a dynamic programming process. xOrder is compatible with various classification models and applicable to a variety of ranking fairness metrics. We evaluate our proposed algorithm on four benchmark data sets and one real world patient electronic health record repository. The experimental results show that our approach can achieve great balance between the algorithm utility and ranking fairness. Our algorithm can also achieve robust performance when training and testing ranking score distributions are significantly different.

1 Introduction

Machine learning algorithms have been widely applied in a variety of real world applications including the high-stakes scenarios such as loan approvals, criminal justice, healthcare, etc. An increasing concern is whether these algorithms make fair decisions in these cases. For example, ProPublica reported that an algorithm used across US for predicting a defendant’s risk of future crime produced higher scores to African-Americans than Caucasians on average [1]. This stimulates lots of research on improving the fairness of the decisions made by machine learning algorithms.

Existing works on fairness in machine learning have mostly focused on the disparate impacts of binary decisions informed by algorithms with respect to different groups formed from the protected variables (e.g., gender or race). Demographic parity requires the classification results to be independent of the group memberships. Equalized odds [15] seeks for equal false positive and negative rates across different groups. Accuracy parity [35] needs equalized error rates across different groups.

*equal contributions

Another scenario that frequently involves computational algorithms is ranking. For example, Model for End-stage Liver Disease (MELD) score, which is derived from a simple linear model from several features, has been used for prioritizing candidates who need liver transplantation [33]. Studies have found that women were less likely than men to receive a liver transplant within 3 years with the MELD score [29]. To quantify ranking fairness, Kallus *et al.* [19] proposed $xAUC$, which measures the probability of positive examples of one group being ranked above negative examples of another group. Beutel *et al.* [2] proposed a similar definition *pairwise ranking fairness* (PRF), which requires equal probabilities for positive instances from each group ranked above all negative instances.

To address the potential disparity induced from risk scores, Kallus *et al.* [19] proposed a post-processing approach that adjusts the risk scores of the instances in the disadvantaged group with a parameterized monotonically increasing function. This method is model agnostic and aims to achieve equal $xAUC$, but it does not consider algorithm utility (i.e., classification performance) explicitly. Beutel *et al.* [2] studied the balance between algorithm utility and ranking fairness and proposed an optimization framework by minimizing an objective including the classification loss and a regularization term evaluating the absolute correlation between the group membership and pairwise residual predictions. This method, although considers both algorithm utility and fairness, is model-dependent and does not directly optimize PRF disparity but an approximated proxy.

In this paper, we develop a model agnostic post-processing framework, `xOrder`, to achieve ranking fairness while at the same time maximally maintain the algorithm utility. Specifically, we show that both algorithm utility and ranking fairness are essentially determined by the ordering of the instances involved, therefore, `xOrder` makes direct adjustments of the cross-group instance ordering (while existing post-processing algorithms mostly aimed at adjusting the ranking scores). The optimal adjustments can be obtained through a dynamic programming procedure of minimizing an objective comprising a weighted sum of algorithm utility loss and ranking disparity. The learned ordering adjustment can be easily transferred to the testing data through linear interpolation.

We evaluate `xOrder` empirically on four popular benchmark data sets for studying algorithm fairness and a real-world electronic health record data repository. The results show `xOrder` can achieve low ranking disparities on all data sets while at the same time maintaining good algorithm utilities. The source codes of `xOrder` are made publicly available at <https://github.com/xOrder-code/xOrder>.

2 Related Work

Algorithm fairness is defined as the disparities in the decisions made across groups formed by protected variables, such as gender and race. Many previous works on this topic focused on binary decision settings. Researchers have used different proxies as fairness measures which are required to be the same across different groups for achieving fairness. Examples of such proxies include the proportion of examples classified as positive [5, 6], as well as the prediction performance metrics such as true/false positive rates and error rates [9, 10, 15, 35, 18]. A related concept that is worthy of mentioning here is calibration [25]. A model with risk score S on input X to generate output Y is considered calibrated by group if for $\forall s \in [0, 1]$, we have $\Pr(Y = 1 | S = s, A = a) = \Pr(Y = 1 | S = s, A = b)$ where A is the group variable [8]. Recent studies have shown that it is impossible to satisfy both error rate fairness and calibration simultaneously when the prevalence of positive instances are different across groups [21, 8]. Plenty of approaches have been proposed to achieve fairness in binary classification settings. One type of methods is to train a classifier without any adjustments and then post-process the prediction scores by setting different thresholds for different groups [15]. Other methods have been developed for optimization of fairness metrics during the model training process through adversarial learning [38, 26, 4, 27, 39] or regularization [20, 36, 3].

Ranking fairness is an important issue in applications where the decisions are made by algorithm produced ranking scores, such as the example of liver transplantation candidate prioritization with MELD score [33]. This problem is related to but different from binary decision making [30, 28]. There are prior works formulating this problem in the setting of selecting the top- k items ranked based on the ranking scores for any k [7, 34, 37, 13]. For each sub-problem with a specific k , the top- k ranked examples can be treated as positive while the remaining examples can be treated as negative, so that these sub-problems can be viewed as binary classification problems. There are also works trying to assign a weight to each instance according to the orders and study the difference of such weights across different groups [31, 32]. Our focus is the fairness on bipartite ranking, which

seeks for a good ranking function that ranks positive instances above negative ones [28]. Kallus *et al.* [19] defined xAUC (Area under Cross-Receiver Operating Characteristic curve) as the probability of positive examples of one group being ranked above negative examples of another group. They require equal xAUC to achieve ranking fairness. Beutel *et al.* proposed a similar definition pairwise ranking fairness (PRF) as the probability that positive examples from one group are ranked above all negative examples [2] and use the difference of PRF across groups as ranking fairness metric. They further proved that some traditional fairness metrics (such as calibration and MSE) are insufficient for guaranteeing ranking fairness under PRF metric.

To address ranking fairness problem, Kallus *et al.* [19] proposed a post-processing technique. They transformed the prediction scores in the disadvantaged group with a logistic function and optimized the empirical xAUC disparity by exhaustive searching on the space of parameters without considering the trade-off between algorithm utility and fairness. Beutel *et al.* proposed a pairwise regularization for the objective function [2]. The regularization is computed as the absolute correlation between the residual prediction scores of the positive and negative example and the group membership of the positive example. However, PRF disparity is determined by judging whether a positive example is ranked above a negative one using an indicator function. The proposed pairwise regularization can be seen as an approximation of PRF disparity by replacing the indicator function with the residual prediction scores. This regularization does not guarantee ranking fairness under PRF metric. Moreover, it is difficult to apply this regularization to some learning framework such as Rankboost [11].

3 Notations and Problem Settings

Suppose we have data (X, A, Y) on features $X \in \mathcal{X}$, sensitive attribute $A \in \mathcal{A}$ and binary label $Y \in \{0, 1\}$. We are interested in the performance and the fairness issue of a predictive ranking score function $R: \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}$. Here we focus on the case where R returns an estimated conditional probability positive label corresponding to a given individual's ranking score, and we use $S = R(X, A) \in [0, 1]$ to denote the individual ranking score variable.

Given S , we can derive a binary classifier with a given threshold θ , such that $\hat{Y}_\theta = \mathbb{I}[S \geq \theta]$ and \mathbb{I} is the indicator function. To evaluate the performance of R , the receiver operator characteristic (ROC) curve is widely adopted with the false positive rate (FPR) on x-axis and the true positive rate (TPR) on y-axis as the threshold θ varies. The area under the ROC curve (AUC) quantitatively measures the quality of the binary classifier induced from R .

AUC can also be understood as the probability that a randomly drawn ranking score from the positive class is ranked above a randomly drawn score from the negative class [14]

$$\text{AUC} = \Pr[S_1 > S_0] = \frac{1}{n_1 \cdot n_0} \cdot \sum_{i: Y_i=1} \sum_{j: Y_j=0} \mathbb{I}[R(X_i) > R(X_j)], \quad (1)$$

where S_1 and S_0 represent a ranking score of a random positive and negative sample. n_1 and n_0 correspond to the number of positives and negatives, respectively. Note that we dropped the group variable in the R function because it is irrelevant to the measure of AUC (i.e., X_i, X_j can be from any groups). The two group-level ranking fairness metrics can be measured by the following Cross-Area Under the Curve (xAUC) metric [19].

Definition 1 (xAUC [19]). *The xAUC of group a over b is defined as*

$$\text{xAUC}(a, b) = \Pr[S_1^a > S_0^b] = \frac{1}{n_1^a n_0^b} \cdot \sum_{i: i \in a, Y_i=1} \sum_{j: j \in b, Y_j=0} \mathbb{I}[R(X_i, a) > R(X_j, b)], \quad (2)$$

where a and b are two groups formed by the sensitive variable A . S_1^a is the ranking score of a random positive sample in a . S_0^b is the ranking score of a random negative sample in b . i is the index of a particular positive sample from group a , whose corresponding ranking score is $R(X_i, a)$. j is the index of a particular negative sample from group b , whose corresponding ranking score is $R(X_j, b)$.

From Eq.(2) we can see that xAUC measures the probability of a random positive sample in a ranked higher than a random negative sample in b . Correspondingly, $\text{xAUC}(b, a)$ means $\Pr(S_1^b > S_0^a)$, and the ranking disparity can be measured by

$$\Delta \text{xAUC}(a, b) = |\text{xAUC}(a, b) - \text{xAUC}(b, a)| = \left| \Pr(S_1^a > S_0^b) - \Pr(S_1^b > S_0^a) \right|. \quad (3)$$

Definition 2 (Pairwise Ranking Fairness (PRF) [2]). *The PRF for group a is defined as*

$$\text{PRF}(a) = \Pr[S_1^a > S_0] = \frac{1}{n_1^a \cdot n_0} \sum_{i:i \in a, Y_i=1} \sum_{j:Y_j=0} \mathbb{I}[R(X_i, a) > R(X_j)], \quad (4)$$

where sample j can belong to either group a or group b .

From Eq.(4) we can see that the PRF for group a measures the probability of a random positive sample in a ranked higher than a random negative sample in either a or b . Then we can also define the following ΔPRF metric to measure the ranking disparity

$$\Delta\text{PRF}(a, b) = \left| \Pr[S_1^a > S_0] - \Pr[S_1^b > S_0] \right|. \quad (5)$$

From above definitions we can see the utility (measured by AUC as in Eq.(1)) and fairness (measured by ΔxAUC in Eq.(2) or ΔPRF in Eq.(4)) of ranking function R are essentially determined by the ordering of data samples induced by the predicted ranking scores. In the following, we use p^a and p^b to represent the data sample sequences in a and b with their ranking scores ranked in descending orders. That is, $p^a = [p^{a(1)}, p^{a(2)}, \dots, p^{a(n^a)}]$ with $R(X_{p^{a(i)}}, a) \geq R(X_{p^{a(j)}}, a)$ if $0 \leq i < j \leq n^a$, and p^b is defined in the same way, then we have the following definition.

Definition 3 (Cross-Group Ordering O). *Given ordered instance sequences p^a and p^b , the cross-group ordering $o(p^a, p^b)$ defines a ranked list combining the instances in groups a and b while keeps within group instance ranking orders preserved.*

One example of such cross-group ordering is $o(p^a, p^b) = [p^{a(1)}, p^{b(1)}, p^{a(2)}, \dots, p^{a(n^a)}, \dots, p^{b(n^b)}]$. With this definition, we have the following proposition.

Proposition 1. *Given ordered instance sequences p^a and p^b , there exists a crossing-group ordering $o(p^a, p^b)$ that can achieve $\Delta\text{xAUC} \leq \min(\max(1/n_1^b, 1/n_0^b), \max(1/n_1^a, 1/n_0^a))$ or $\Delta\text{PRF} \leq \min(\max(n_0^b/(n_1^a n_0), 1/n_0), \max(n_0^a/(n_1^b n_0), 1/n_0))$ with the two ranking fairness measures.*

The proof of Proposition 1 is provided in the supplemental material.

From the above definitions we can see that we only need cross-group ordering $o(p^a, p^b)$ to estimate both algorithm utility measured by AUC and ranking fairness measured by either ΔxAUC or ΔPRF , i.e., we do not need the actual ranking scores. Our proposal in this paper is to look for an optimal cross-group ordering $o^*(p^a, p^b)$, which can achieve ranking fairness and maximally maintain algorithm utility, through post-processing. We will use xAUC as the ranking fairness metric in our detailed derivations. The same procedure can be similarly developed for PRF based ranking fairness, and the details are provided in supplemental material.

One important issue to consider is that the cross-group ordering that achieves the same level of ranking fairness is not unique. In Figure 1, we demonstrate an illustrative example with 9 samples showing that different cross-group ordering can result in different AUCs with the same ΔxAUC . The middle row in Figure 1 shows the original predicted ranking scores and their induced sample ranking, which achieves a ranking disparity $\Delta\text{xAUC} = 0.75$. The top row shows one cross-group ordering with ranking disparity $\Delta\text{xAUC} = 0$ and algorithm utility $\text{AUC} = 0.56$. The bottom row shows another cross-group ranking with $\Delta\text{xAUC} = 0$ but $\text{AUC} = 0.83$.

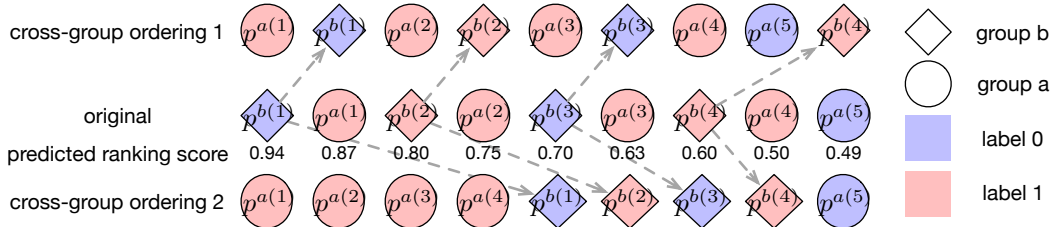


Figure 1: An example to illustrate the post-processing. Original scores are in the middle. The first row is the ordering after post-processing, while the optimal ordering is on the bottom.

Problem Setting. Our goal is to identify an optimal cross-group ordering $o^*(p^a, p^b)$ that leads to the minimum ranking disparity (measured by ΔxAUC) with algorithm utility (measured by AUC)

maximally maintained. We can maximize the following objective

$$J(o(p^a, p^b)) = \text{AUC}(o(p^a, p^b)) - \lambda \cdot \Delta \text{xAUC}(o(p^a, p^b)), \quad (6)$$

where we use $\text{AUC}(o(p^a, p^b))$ to denote the AUC induced by the ordering $o(p^a, p^b)$, which is calculated in the same way as in Eq.(1) if we think of $R(X_i)$ returning the rank of X_i instead of the actual ranking score. Please note that in the rest of this paper we will use similar notations for xAUC without causing further confusions. Similarly, $\Delta \text{xAUC}(o(p^a, p^b))$ is the ranking disparity induced by ordering $o(p^a, p^b)$ calculated as in Eq.(3). Then we have the following proposition.

Proposition 2. *The objective function in Eq.(6) is equivalent to:*

$$G(o(p^a, p^b)) = k_{a,b} \cdot \text{xAUC}(o(p^a, p^b)) + k_{b,a} \cdot \text{xAUC}(o(p^b, p^a)) - \lambda k \cdot \Delta \text{xAUC}(o(p^a, p^b)), \quad (7)$$

where $k_{a,b} = n_1^a n_0^b$, $k_{b,a} = n_0^a n_1^b$, $k = n_0 n_1$, $\text{xAUC}(o(p^a, p^b))$ indicates $\text{xAUC}(a, b)$ in Eq. 2 ($\text{xAUC}(o(p^b, p^a))$ indicates $\text{xAUC}(b, a)$ in Eq. 2), and λ is the hyperparameter trading off the algorithm utility and ranking disparity.

The proof of this proposition is provided in the supplemental material.

4 Algorithm

To intuitively understand the post-processing process, we treat the cross-group ordering as a path from the higher ranked instances to lower ranked instances. With the same example shown in Figure 1, we demonstrate the original ordering of those samples induced by their predicted ranking scores (middle row of Figure 1) on the top of Figure 2(a), where the direction on the path indicates the ordering. The ordering corresponding to the bottom row of Figure 1 is demonstrated at the bottom of Figure 2(a).

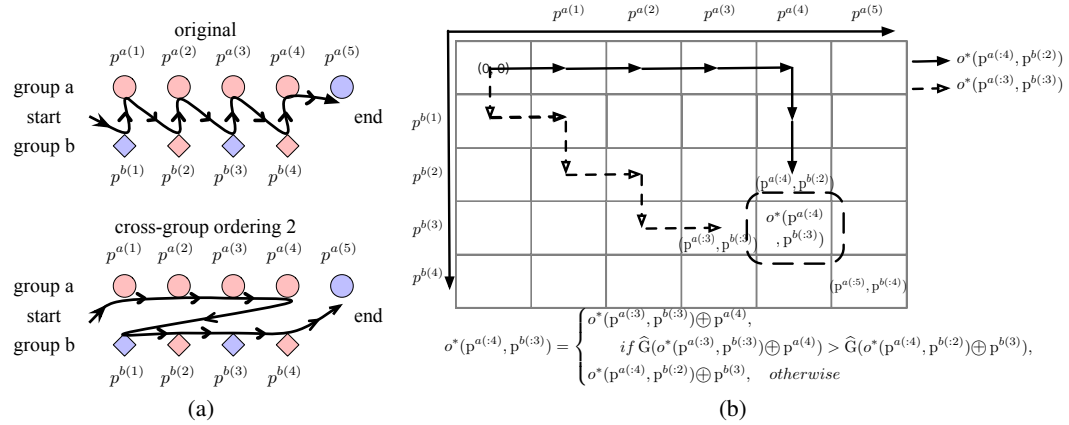


Figure 2: (a) Illustrations of the ordering procedure as path finding for the example in Figure 1. (b) Illustration of how our proposed method optimizes such relative ordering.

With this analogy, the optimal cross-group ordering $o^*(p^a, p^b)$ can be achieved by a greedy path finding process. The path must start from $p^{a(1)}$ or $p^{b(1)}$, and end with $p^{a(n^a)}$ or $p^{b(n^b)}$. Each instance in p^a and p^b can only appear once in the final path, and the orders of the instances in the final path must be the same as their orders in p^a and p^b . The path can be obtained through a dynamic programming process. In particular, we first partition the entire decision space into a $(n^b + 1) \times (n^a + 1)$ grid. Each location (i, j) $0 \leq i \leq n^a, 0 \leq j \leq n^b$ on the lattice corresponds to a decision step on determining whether to add $p^{a(i)}$ or $p^{b(j)}$ into the current path, which can be determined with the following rule:

$$\begin{aligned} & \text{Given } o^*(p^{a(i-1)}, p^{b(j)}), o^*(p^{a(i)}, p^{b(j-1)}) \\ & \text{if: } \hat{G}(o^*(p^{a(i-1)}, p^{b(j)}) \oplus p^{a(i)}) > \hat{G}(o^*(p^{a(i)}, p^{b(j-1)}) \oplus p^{b(j)}) \\ & \quad o^*(p^{a(i)}, p^{b(j)}) = o^*(p^{a(i-1)}, p^{b(j)}) \oplus p^{a(i)}; \\ & \text{otherwise: } o^*(p^{a(i)}, p^{b(j)}) = o^*(p^{a(i)}, p^{b(j-1)}) \oplus p^{b(j)}, \end{aligned} \quad (8)$$

where $p^{a(i)}$ represents the first i elements in p^a (and $p^{a(i-1)}, p^{b(j-1)}, p^{b(j)}$ are similarly defined). $o^*(p^{a(i-1)}, p^{b(j)}) \oplus p^{a(i)}$ means appending $p^{a(i)}$ to the end of $o^*(p^{a(i-1)}, p^{b(j)})$. The value of function $\widehat{G}(o(p^{a(i)}, p^{b(j)}))$ is defined as follows

$$\begin{aligned} \widehat{G}(o(p^{a(i)}, p^{b(j)})) &= k_{a,b} \cdot \text{xAUC}(o(p^{a(i)}, p^{b(j)}) \oplus p^{b(j+1:n^b)}) \\ &+ k_{b,a} \cdot \text{xAUC}(o(p^{b(j)}, p^{a(i)}) \oplus p^{a(i+1:n^a)}) \\ &+ (\lambda \cdot k) \cdot \left| \text{xAUC}(o(p^{a(i)}, p^{b(j)}) \oplus p^{b(j+1:n^b)}) - \text{xAUC}(o(p^{b(j)}, p^{a(i)}) \oplus p^{a(i+1:n^a)}) \right|, \end{aligned} \quad (9)$$

in which $\widehat{G}(o(p^{a(i)}, p^{b(j)})) = G(o(p^a, p^b))$ in Eq. 15 when $i = n^a, j = n^b$.

Figure 2(b) demonstrates the case of applying our rule to step $i = 4, j = 3$ in the example shown in Figure 1.

Algorithm 1 summarized the whole pipeline of identifying the optimal path. In particular, our algorithm calculates a cost for every point (i, j) in the decision lattice as the value of the \widehat{G} function evaluated on the path reaching (i, j) from $(0, 0)$. Please note that the first row ($j = 0$) only involves the instances in a , therefore the path reaching to the points in this row are uniquely defined considering the within group instance order should be preserved in the path. The decision points in the first column ($i = 0$) enjoy similar characteristics. After the cost values for the decision points in first row and column are calculated, the costs values on the rest of the decision points in the lattice can be calculated iteratively until $i = n^a$ and $j = n^b$.

Algorithm 1 can also be viewed as a process to maximize the objective function in Eq.(6). Different λ values trade-off the algorithm utility and ranking fairness differently and the solution Algorithm 1 converges to is a local optima. Moreover, we have the following proposition.

Proposition 3. *xOrder can achieve the global optimal solution of maximizing Eq.(6) with $\lambda = 0$.*

The proof of this proposition is provided in the supplemental material.

Algorithm 1: xOrder: optimize the cross-group ordering with post-processing

Input: λ , the ranking scores S^a, S^b from a predictive ranking function R

Sort the ranking scores S^a and S^b in descending order. Get the instances ranking

$p^a = [p^{a(1)}, p^{a(2)}, \dots, p^{a(n^a)}]$ and $p^b = [p^{b(1)}, p^{b(2)}, \dots, p^{b(n^b)}]$

Initialize cross-group ordering $o^*(p^{a(i)}, p^{b(j)})$ ($0 \leq i \leq n^a, o^*(p^{a(0)}, p^{b(j)})$ ($0 \leq j \leq n^b$),

forall $i = 1, 2, 3 \dots n^a$ **do**

forall $j = 1, 2, 3 \dots n^b$ **do**

 Calculate $\widehat{G}(o^*(p^{a(i-1)}, p^{b(j)}) \oplus p^{a(i)})$, $\widehat{G}(o^*(p^{a(i)}, p^{b(j-1)}) \oplus p^{b(j)})$ using the Eq.9

 Update $o^*(p^{a(i)}, p^{b(j)})$ according to the Eq.8

end

end

Output: the learnt cross-group ordering $o^*(p^a, p^b)$

The testing phase. Since the learned ordering in training phase cannot be directly used in testing stage, we propose to transfer the information of the learned ordering by rearranging the ranking scores of the disadvantaged group (which is assumed to be group b without the loss of generality). In particular, the process contains two steps:

1. Ranking score adjustment for the disadvantaged group in training data. Fixing the ranking scores for training instances in group a , the adjusted ranking scores for training instances in group b will be obtained by uniform linear interpolation according to their relative positions in the learned ordering. For example, if we have an ordered sequence $(p^{a(1)}, p^{b(1)}, p^{b(2)}, p^{a(2)})$ with the ranking scores for $p^{a(1)}$ and $p^{a(2)}$ being 0.8 and 0.5, then the adjusted ranking scores for $p^{b(1)}$ and $p^{b(2)}$ being 0.7 and 0.6.
2. Ranking score adjustment for the disadvantaged group in testing data. For testing instances, we follow the same practice of just adjusting the ranking scores of the instances from group b but keep

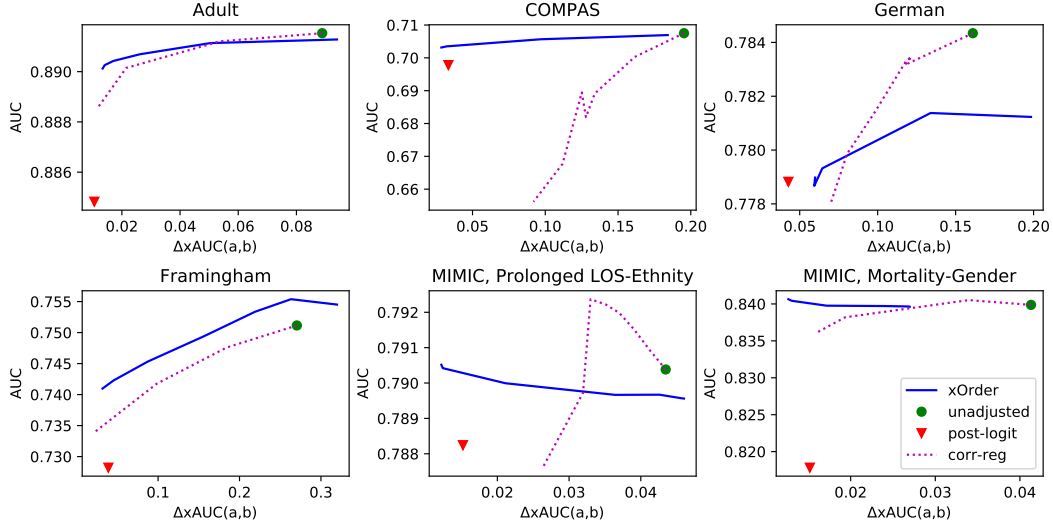
Table 1: Summary of benchmark data sets

Dataset	n	p	A	Y
COMPAS[1]	6,167	400	Race(white, non-white)	Non-recidivism within 2 years
Adult[22]	30,162	98	Race(white, non-white)	Income \geq 50K
German [24]	1,000	57	Age(\geq 25, $<$ 25)	Creditworthiness
Framingham [23]	4,658	7	Gender(male,female)	10-year CHD incidence

the ranking scores for instances from group a unchanged. For the adjustment process, we first rank both training and testing instances from group b according to their raw unadjusted ranking scores to get an ordered list. Then the adjusted ranking scores for testing instances in b can be obtained by linear interpolation from the adjusted ranking scores of the training instances in b .

5 Experiment

Data sets. We conduct experiments on 4 popular benchmark data sets for studying algorithm fairness using the same setting as in [19]. The basic information for these data sets are summarized in Table 1, where n and p are the number of instances and features, and CHD is the abbreviation of coronary heart disease. We also use MIMIC-III, a real world electronic health record repository for ICU patients [17], to study ranking fairness in real world clinical ranking prediction scenarios. The data set was preprocessed as in [16] with $n = 21,139$ and $p = 714$. Each instance is a specific ICU stay. For label Y , we consider in-hospital mortality and prolonged length of stay (whether the ICU stay is longer than 1 week). For protected variable A , we consider gender (male, female) and ethnicity (white, non-white).

Figure 3: AUC- Δ xAUC trade-off with logistic regression model.

Evaluation Settings. We adopt two classification models, logistic regression and bipartite rank-boost [11], in our empirical evaluations. Δ xAUC and Δ PRF are used as ranking fairness metrics and AUC is used to measure the algorithm utility. The algorithms that originally proposed xAUC [19] and PRF [2] are evaluated as baselines, which are denoted as **post-logit** and **corr-reg**, respectively. We also report the performance obtained by the two original classification models without fairness considerations (called **unadjusted**).

During evaluation, for post-processing algorithms including post-logit and xOrder, we first train the classifiers on training data without any specific considerations on ranking fairness to obtain the unadjusted prediction ranking scores, and then apply them to adjust these scores. For corr-reg, we directly train the classifiers with additional fairness regularization. The total AUC and ranking fairness metrics on test data are reported. In addition, we plot a curve showing the trade-off between ranking fairness and algorithm utility for xOrder and corr-reg with varying the trade-off parameter. We repeat each setting of the experiment ten times and report the average results.

To make the comparisons fair, we report the results of different methods under the same classification model separately. Since it is difficult to add the pairwise fairness regularization into the training framework of bipartite rankboost, we only implement corr-reg with logistic regression.

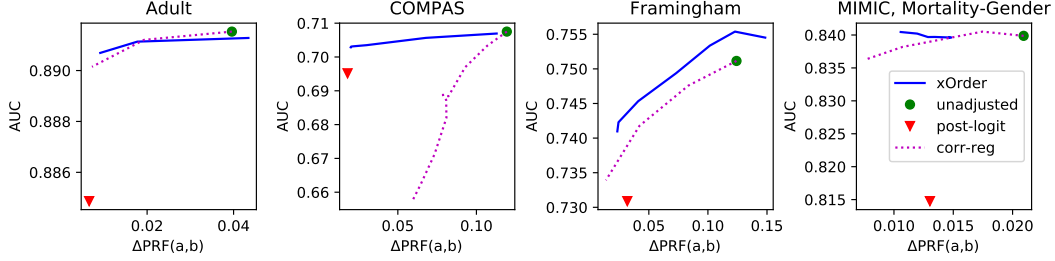


Figure 4: AUC- ΔPRF trade-off with logistic regression model.

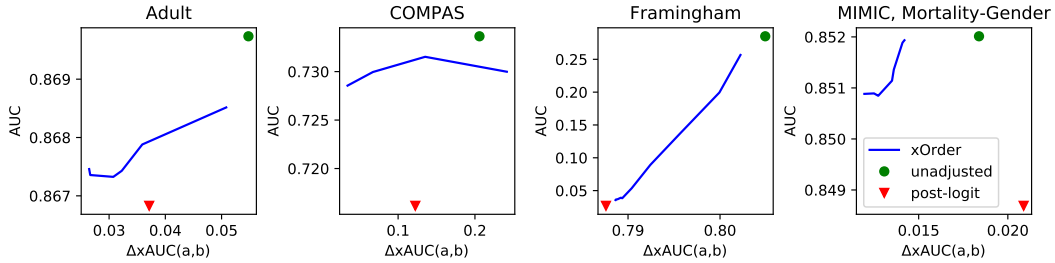


Figure 5: AUC- ΔxAUC trade-off with bipartite rankboost model.

Evaluation Results. The results obtained with logistic regression and ΔxAUC metric are shown in Figure 3. For MIMIC-III, we show the results as Y – A combinations (mortality-gender and prolonged length of stay (LOS)-ethnicity).

All three methods considering ranking fairness (xOrder, corr-reg, post-logit) are able to obtain lower ΔxAUC comparing to the unadjusted results. xOrder and post-logit can achieve ΔxAUC closed to zero on almost all the datasets. This supports proposition 1 empirically that post-processing by changing cross-group ordering have the potential to achieve ΔxAUC closed to zero. corr-reg fails to obtain results with low ranking disparities on COMPAS and German. One possible reason is that the correlation regularizer is only an approximation of ranking disparity under PRF metric.

Another observation is that xOrder can achieve better trade-off between utility and fairness. It can obtain competitive or obviously better AUC under the same level of ΔxAUC when compared with other methods, especially in the regiment with low ΔxAUC . corr-reg performs worse than other methods in COMPAS, while post-logit performs worse in Adult and Framingham. xOrder performs well consistently on all data sets.

Figure 4 illustrates the results with logistic regression and ΔPRF metric on four data sets. The findings are similar to those in Figure 3, since ΔPRF and ΔxAUC are highly correlated on these data sets. Although the regularization term in corr-reg is related to the definition of ΔPRF , xOrder maintains its superiority over corr-reg.

The results on bipartite rankboost model and the associated ΔxAUC are shown in Figure 5. It can still be observed that xOrder achieves higher AUC than post-logit under the same ΔxAUC . Moreover, post-logit cannot achieve ΔxAUC as low as xOrder on adult and COMPAS. This is because the distributions of the prediction scores from the bipartite rankboost model on training and testing data are significantly different. Therefore, the function in post-logit which achieves equal xAUC on training data may not generalize well on test data. Our method is robust against such difference. We empirically illustrate the relation between the distribution of the prediction ranking scores and the generalization ability of post-progressing algorithms in Section F.3 of supplemental material.

6 Conclusion

In this paper, we propose a general post-processing framework, `xOrder`, to achieve a good balance between ranking fairness and model utility by direct adjustment of the cross group ranking orders. We formulate `xOrder` as an optimization problem and propose a dynamic programming process to solve it. Empirical results on both benchmark and real world medical data sets demonstrated that `xOrder` can achieve a low ranking disparity while keep a maximum algorithm utility.

Broader Impact Statement

In this paper we investigate the problem of algorithmic fairness, which has attracted lots of attentions in a variety of application domains involving algorithmic decision making such as criminal justice, healthcare insurance plan enrollment, job recruitment, etc. Algorithms that can lead to potential decision disparity can be disastrous in these applications. Therefore it is crucial to develop algorithms that can make fair decisions. The approach we proposed in this paper improves both algorithm fairness and utility, and thus greatly enhances its generalization ability and trustworthiness.

References

- [1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. *And it's biased against blacks*. *ProPublica*, 23, 2016.
- [2] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H Chi, et al. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2212–2220, 2019.
- [3] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H Chi. Putting fairness principles into practice: Challenges, metrics, and improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 453–459, 2019.
- [4] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017.
- [5] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18. IEEE, 2009.
- [6] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.
- [7] L Elisa Celis, Damian Straszak, and Nisheeth K Vishnoi. Ranking with fairness constraints. *arXiv preprint arXiv:1704.06840*, 2017.
- [8] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.
- [9] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, 2018.
- [10] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.
- [11] Yoav Freund, Raj Iyer, Robert E Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *Journal of machine learning research*, 4(Nov):933–969, 2003.
- [12] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 329–338, 2019.
- [13] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2221–2231, 2019.
- [14] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.
- [15] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.
- [16] Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):1–18, 2019.
- [17] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.
- [18] Nathan Kallus and Angela Zhou. Residual unfairness in fair machine learning from prejudiced data. In *International Conference on Machine Learning*, pages 2439–2448, 2018.

- [19] Nathan Kallus and Angela Zhou. The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric. In *Advances in Neural Information Processing Systems*, pages 3433–3443, 2019.
- [20] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650. IEEE, 2011.
- [21] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.
- [22] Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pages 202–207, 1996.
- [23] Daniel Levy. *50 years of discovery: medical milestones from the National Heart, Lung, and Blood Institute’s Framingham Heart Study*. Center for Bio-Medical Communication, Inc., 1999.
- [24] Moshe Lichman et al. Uci machine learning repository, 2013.
- [25] Sarah Lichtenstein, Baruch Fischhoff, and Lawrence D Phillips. Calibration of probabilities: The state of the art to 1980. Technical report, DECISION RESEARCH EUGENE OR, 1981.
- [26] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.
- [27] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*, 2018.
- [28] Aditya Krishna Menon and Robert C Williamson. Bipartite ranking: a risk-theoretic perspective. *The Journal of Machine Learning Research*, 17(1):6766–6867, 2016.
- [29] Cynthia A Moylan, Carla W Brady, Jeffrey L Johnson, Alastair D Smith, Janet E Tuttle-Newhall, and Andrew J Muir. Disparities in liver transplantation before and after introduction of the meld score. *JAMA*, 300(20):2371–2378, 2008.
- [30] Harikrishna Narasimhan and Shivani Agarwal. On the relationship between binary classification, bipartite ranking, and binary class probability estimation. In *Advances in Neural Information Processing Systems*, pages 2913–2921, 2013.
- [31] Ashudeep Singh and Thorsten Joachims. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2219–2228, 2018.
- [32] Ashudeep Singh and Thorsten Joachims. Policy learning for fairness in ranking. In *Advances in Neural Information Processing Systems*, pages 5427–5437, 2019.
- [33] Russell Wiesner, Erick Edwards, Richard Freeman, Ann Harper, Ray Kim, Patrick Kamath, Walter Kremers, John Lake, Todd Howard, Robert M Merion, et al. Model for end-stage liver disease (meld) and allocation of donor livers. *Gastroenterology*, 124(1):91–96, 2003.
- [34] Ke Yang and Julia Stoyanovich. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, pages 1–6, 2017.
- [35] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017.
- [36] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*, 2015.
- [37] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1569–1578, 2017.
- [38] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.
- [39] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

A Analysis on AUC, xAUC and PRF

A.1 Decomposing AUC into xAUC and iAUC

In fact, the AUC of the risk function can be decomposed into xAUC and iAUC, while iAUC means the probability of positive instances rank above negative instances in the same group:

$$\begin{aligned}
\text{AUC} &= \frac{1}{k} \cdot (k_a \cdot \text{iAUC}(a) + k_{a,b} \cdot \text{xAUC}(a, b) + k_{b,a} \cdot \text{xAUC}(b, a) + k_b \cdot \text{iAUC}(b)) \\
\text{iAUC}(a) &= \Pr[S_1^a > S_0^a] = \frac{1}{k_a} \sum_{i:i \in a, Y_i=1} \sum_{j:j \in a, Y_j=0} \mathbb{I}[R(X_i, a) > R(X_j, a)] \\
\text{iAUC}(b) &= \Pr[S_1^b > S_0^b] = \frac{1}{k_b} \sum_{i:i \in b, Y_i=1} \sum_{j:j \in b, Y_j=0} \mathbb{I}[R(X_i, b) > R(X_j, b)] \\
\text{xAUC}(a, b) &= \Pr[S_1^a > S_0^b] = \frac{1}{k_{a,b}} \cdot \sum_{i:i \in a, Y_i=1} \sum_{j:j \in b, Y_j=0} \mathbb{I}[R(X_i, a) > R(X_j, b)] \\
\text{xAUC}(b, a) &= \Pr[S_1^b > S_0^a] = \frac{1}{k_{b,a}} \cdot \sum_{i:i \in b, Y_i=1} \sum_{j:j \in a, Y_j=0} \mathbb{I}[R(X_i, b) > R(X_j, a)]
\end{aligned} \tag{10}$$

in which $k_{a,b}, k_{b,a}$ are the same as in the main text while $k = n_0 n_1$, $k_a = n_0^a n_1^a$, $k_b = n_0^b n_1^b$.

A.2 Analysis on PRF

From the decomposition Eq. 10, the metric of PRF can be decomposed into xAUC and iAUC as follows:

$$\begin{aligned}
\text{PRF}(a) &= \Pr[S_1^a > S_0] = \frac{1}{n_1^a n_0} \cdot \sum_{i:i \in a, Y_i=1} \sum_{j:Y_j=0} \mathbb{I}[R(X_i, a) > R(X_j)] \\
\text{PRF}(b) &= \Pr[S_1^b > S_0] = \frac{1}{n_1^b n_0} \cdot \sum_{i:i \in b, Y_i=1} \sum_{j:Y_j=0} \mathbb{I}[R(X_i, a) > R(X_j)]
\end{aligned} \tag{11}$$

According to the Eq. 10, the probability $\Pr[S_1^a > S_0]$, $\Pr[S_1^b > S_0]$ can be rewritten as follows:

$$\begin{aligned}
\Pr[S_1^a > S_0] &= \frac{n_0^b}{n_0} \cdot \text{xAUC}(a, b) + \frac{n_0^a}{n_0} \cdot \text{iAUC}(a) \\
\Pr[S_1^b > S_0] &= \frac{n_0^a}{n_0} \cdot \text{xAUC}(b, a) + \frac{n_0^b}{n_0} \cdot \text{iAUC}(b),
\end{aligned} \tag{12}$$

B Proof of Proposition 1

B.1 (ΔxAUC)

Proof. Denote $T_1^a(i) = \frac{\sum_{k \leq i} \mathbb{I}[Y_{p^{a(k)}}=1]}{n_1^a}$, $T_1^a(i)$ is monotonically increasing on i from 0 to 1. Reversely, denote $\overline{T}_0^a(i) = \frac{\sum_{k > i} \mathbb{I}[Y_{p^{a(k)}}=0]}{n_0^a}$ and $\overline{T}_0^a(i)$ is monotonically decreasing on i from 1 to 0. Then $T_1^a(i) - \overline{T}_0^a(i)$ is monotonically increasing on i from -1 to 1. So there exists an i' that $T_1^a(i') - \overline{T}_0^a(i') < 0$ and $T_1^a(i'+1) - \overline{T}_0^a(i'+1) \geq 0$. Since the increment of $T_1^a(i) - \overline{T}_0^a(i)$ when i increases by 1 satisfies $\Delta(T_1^a(i) - \overline{T}_0^a(i)) \leq \max(\frac{1}{n_1^a}, \frac{1}{n_0^a})$, we can get $-\max(\frac{1}{n_1^a}, \frac{1}{n_0^a}) \leq T_1^a(i') - \overline{T}_0^a(i') < 0$.

Consider a cross-group ordering which is generated by inserting the whole sequence p^b between $p^{a(i')}$ and $p^{a(i'+1)}$, this operation will result in that positive examples $p^{a(k)}$ with $Y_{p^{a(k)}} = 1, k \leq i'$ will be ranked higher than all the negative examples in p^b . And positive examples $p^{a(k)}$ with $Y_{p^{a(k)}} = 1, k > i'$ will be ranked lower than all the negative examples in p^b . Then $\text{xAUC}(a, b)$ equals $T_1^a(i')$ with this cross-group ordering. Similarly, we can obtain $\text{xAUC}(b, a) = \overline{T}_0^a(i')$. Then $\Delta\text{xAUC} = |\overline{T}_0^a(i') - T_1^a(i')|$. According to the discussion above, $\Delta\text{xAUC} \leq \max(\frac{1}{n_1^a}, \frac{1}{n_0^a})$.

If we consider the cross-group ordering generated by inserting the whole sequence p^a between $p^{b(j)}$ and $p^{b(j+1)}$, symmetrically we will find that there exists a corresponding j' that $\Delta\text{xAUC} \leq \max(\frac{1}{n_1^b}, \frac{1}{n_0^b})$ with this cross-group ordering. We can choose one of these two cross-group ordering operations to achieve $\Delta\text{xAUC} \leq \min(\max(\frac{1}{n_1^b}, \frac{1}{n_0^b}), \max(\frac{1}{n_1^a}, \frac{1}{n_0^a}))$ \square

B.2 (Δ PRF)

Proof. Denote $C = \frac{n_0^b}{n_0} \text{iAUC}(b) - \frac{n_0^a}{n_0} \text{iAUC}(a)$, we can get $-\frac{n_0^a}{n_0} \leq C \leq \frac{n_0^b}{n_0}$. Since changing cross-group ordering does not affect inner-group ordering, C is constant for given p^b and p^a .

With the same definition of $T_1^a(i)$ and $\overline{T}_0^a(i)$ in the Section B.1, we will have $\frac{n_0^b}{n_0} T_1^a(i) - \frac{n_0^a}{n_0} \overline{T}_0^a(i)$ is monotonically increasing on i from $-\frac{n_0^a}{n_0}$ to $\frac{n_0^b}{n_0}$. Since $0 \in [-\frac{n_0^a}{n_0} - C, \frac{n_0^b}{n_0} - C]$, there exists i' satisfying $\frac{n_0^b}{n_0} T_1^a(i') - \frac{n_0^a}{n_0} \overline{T}_0^a(i') - C < 0$, $\frac{n_0^b}{n_0} T_1^a(i'+1) - \frac{n_0^a}{n_0} \overline{T}_0^a(i'+1) - C \geq 0$ or $\frac{n_0^b}{n_0} T_1^a(i') - \frac{n_0^a}{n_0} \overline{T}_0^a(i') - C \leq 0$, $\frac{n_0^b}{n_0} T_1^a(i'+1) - \frac{n_0^a}{n_0} \overline{T}_0^a(i'+1) - C > 0$. Since the increment of $\frac{n_0^b}{n_0} T_1^a(i) - \frac{n_0^a}{n_0} \overline{T}_0^a(i) - C$ when i increases by 1 satisfies $\Delta(\frac{n_0^b}{n_0} T_1^a(i) - \frac{n_0^a}{n_0} \overline{T}_0^a(i) - C) \leq \max(\frac{n_0^b}{n_1 n_0}, \frac{1}{n_0})$, $-\max(\frac{n_0^b}{n_1 n_0}, \frac{1}{n_0}) \leq \frac{n_0^a}{n_0} T_1^a(i') - \frac{n_0^b}{n_0} \overline{T}_0^a(i') - C \leq \max(\frac{n_0^b}{n_1 n_0}, \frac{1}{n_0})$.

Consider an cross-group ordering generated by inserting the whole sequence p^b between $p^{a(i')}$ and $p^{a(i'+1)}$. $\text{xAUC}(a, b) = T_1^a(i')$ and $\text{xAUC}(b, a) = \overline{T}_0^a(i')$. As in the calculation of $\Delta \text{PRF} = |\frac{n_0^b}{n_0} \text{xAUC}(a, b) - \frac{n_0^a}{n_0} \text{xAUC}(b, a) - (\frac{n_0^b}{n_0} \text{iAUC}(b) - \frac{n_0^a}{n_0} \text{iAUC}(a))| = |\frac{n_0^b}{n_0} T_1^a(i') - \frac{n_0^a}{n_0} \overline{T}_0^a(i') - C|$, we can get $\Delta \text{PRF} \leq \max(\frac{n_0^b}{n_1 n_0}, \frac{1}{n_0})$.

If we consider the cross-group ordering generated by inserting the whole sequence p^a between $p^{b(j)}$ and $p^{b(j+1)}$, symmetrically we will find there exists a corresponding j' that $\Delta \text{PRF} \leq \max(\frac{n_0^a}{n_1 n_0}, \frac{1}{n_0})$. We can choose one from these two cross-group ordering operations and achieve $\Delta \text{PRF} \leq \min(\max(\frac{n_0^b}{n_1 n_0}, \frac{1}{n_0}), \max(\frac{n_0^a}{n_1 n_0}, \frac{1}{n_0}))$. \square

C Analysis on Proposition 2

C.1 Proof of Proposition 2

Proof. As we keep the with-in group ordering invariant, $\text{iAUC}(a)$ and $\text{iAUC}(b)$ remain the same after post-processing procedure. For the objective function:

$$J(o(p^a, p^b)) = \text{AUC}(o(p^a, p^b)) - \lambda \cdot \Delta \text{xAUC}(o(p^a, p^b)) \quad (13)$$

the item $\text{AUC}(o(p^a, p^b))$ can be decomposed according to Eq. 10. We subtract the constant part $k_a \cdot \text{iAUC}(a)$, $k_b \cdot \text{iAUC}(b)$ and multiply the formula by a constant k :

$$J(o(p^a, p^b)) = \frac{1}{k} (k_{a,b} \text{xAUC}(o(p^a, p^b)) + k_{b,a} \text{xAUC}(o(p^b, p^a)) - \lambda k \Delta \text{xAUC}(o(p^a, p^b))) + C \quad (14)$$

where C is constant $C = \frac{k_a}{k} \cdot \text{iAUC}(p^a) + \frac{k_b}{k} \cdot \text{iAUC}(p^b)$

From Eq. 14, the target Eq. 13 is equivalent to:

$$G(o(p^a, p^b)) = k_{a,b} \cdot \text{xAUC}(o(p^a, p^b)) + k_{b,a} \cdot \text{xAUC}(o(p^b, p^a)) - \lambda \cdot k \cdot \Delta \text{xAUC}(o(p^a, p^b)). \quad (15)$$

\square

Consider the definition of $\widehat{G}(o(p^{a(i)}, p^{b(j)}))$ in main text Eq. (9) in which $\widehat{G}(o(p^{a(i)}, p^{b(j)}))$ is induced by $\text{xAUC}(o(p^{a(i)}, p^b))$ and $\text{xAUC}(o(p^{b(j)}, p^a))$. The cross-group ordering $o(p^{a(i)}, p^b)$ means appending the sequence $p^{b(j+1:n^b)}$ to the given $o(p^{a(i)}, p^{b(j)})$. The meanings of partial $\text{xAUC}(o(p^{a(i)}, p^b))$ is as follows:

$$\text{xAUC}(o(p^{a(i)}, p^b)) = \frac{1}{k_{a,b}} \cdot \sum_{k: k \leq i, Y_{p^{a(k)}} = 1} \sum_{h: h \leq n^b, Y_{p^{a(h)}} = 0} \mathbb{I}[p^{a(k)} \succ p^{b(h)}]. \quad (16)$$

C.2 Proposition 2 and the Objective \widehat{G}

Considering the definition of $\widehat{G}(o(p^{a(i)}, p^{b(j)}))$ in main text Eq. (9) in which $\widehat{G}(o(p^{a(i)}, p^{b(j)}))$ is induced by $\text{xAUC}(o(p^{a(i)}, p^{b(j)}) \oplus p^{b(j+1:n^b)})$ and $\text{xAUC}(o(p^{b(j)}, p^{a(i)}) \oplus p^{b(j+1:n^b)})$, the cross-

group ordering $o(p^{a(i)}, p^{b(j)}) \oplus p^{b(j+1:n^b)}$ means appending the sequence $p^{b(j+1:n^b)}$ to the given cross-group ordering $o(p^{a(i)}, p^{b(j)})$ ($o(p^{b(j)}, p^{a(i)}) \oplus p^{a(i+1:n^a)}$ is similarly defined). For expression convenience, we use $o(p^{a(i)}, p^b)$ to replace $o(p^{a(i)}, p^{b(j)}) \oplus p^{b(j+1:n^b)}$, and $o(p^{a(i+1)}, p^b)$ means $o(p^{a(i)}, p^{b(j)}) \oplus p^{a(i+1)} \oplus p^{b(j+1:n^b)}$. And the property of the partial xAUC is as follows:

$$\begin{aligned} \text{xAUC}(o(p^{a(i)}, p^b)) &= \frac{1}{k_{a,b}} \cdot \sum_{k:k \leq i, Y_{p^a(k)}=1} \sum_{h:h \leq n^b, Y_{p^a(h)}=0} \mathbb{I}[p^{a(k)} \succ p^{b(h)}] \\ \text{xAUC}(o(p^{a(i+1)}, p^b)) &= \text{xAUC}(o(p^{a(i)}, p^b)) + \mathbb{I}[Y_{p^a(i+1)}=1] \cdot \left(\sum_{h:h > j} \mathbb{I}(Y_{p^b(h)}=0) \right) \end{aligned} \quad (17)$$

Obviously, when $i = n^a, j = n^b$, \widehat{G} equals to G in Proposition 2 of the main text.

C.3 Proposition 2 on PRF metric

The objective function under pairwise ranking fairness metric (PRF) is as follows:

$$J(o(p^a, p^b)) = \text{AUC}(o(p^a, p^b)) - \lambda \cdot \Delta\text{PRF}(o(p^a, p^b)) \quad (18)$$

As post-processing procedure does not change iAUC(a) and iAUC(b), the optimization target in Eq. 18 is equivalent to:

$$\begin{aligned} \text{Maximizing:} \quad & G(o(p^a, p^b)) \\ G(o(p^a, p^b)) &= k_{a,b} \cdot \text{xAUC}(o(p^a, p^b)) + k_{b,a} \cdot \text{xAUC}(o(p^a, p^b)) - \lambda \cdot k \cdot \Delta\text{PRF}(o(p^a, p^b)). \end{aligned} \quad (19)$$

D Proof of Proposition 3

Proof. We will decompose the problem that maximizing AUC into $(n^a + 1) \cdot (n^b + 1)$ subproblems, and \widehat{G} in the main text in Eq. (9) when $\lambda = 0$ is equivalent to the objective in Eq. 20. We will use mathematical induction to prove the conclusion that *xOrder can achieve the global optimal solution to maximize $\text{AUC}(o(p^a, p^b))$* . For each subproblem given i, j with $0 \leq i \leq n^a, 0 \leq j \leq n^b$, the optimization target to maximize is:

$$\begin{aligned} \widehat{G}(o(p^{a(i)}, p^{b(j)})) &= \sum_{k:k \leq i, Y_{p^a(k)}=1} \sum_{h:h \leq n^b, Y_{p^b(h)}=0} \mathbb{I}[p^{a(k)} \succ p^{b(h)}] + \\ &\quad \sum_{k:k \leq n^a, Y_{p^a(k)}=0} \sum_{h:h \leq j, Y_{p^b(h)}=1} \mathbb{I}[p^{b(h)} \succ p^{a(k)}] \end{aligned} \quad (20)$$

For any $0 < i < n^a, 0 \leq j \leq n^b$, according to the property in Eq. 21, the update equation when $p^{a(i+1)}$ is appended to $o(p^{a(i)}, p^{b(j)})$:

$$\widehat{G}(o(p^{a(i)}, p^{b(j)}) \oplus p^{a(i+1)}) = \widehat{G}(o(p^{a(i)}, p^{b(j)})) + \mathbb{I}[Y_{p^a(i+1)}=1] \cdot \left(\sum_{h:h > j} \mathbb{I}(Y_{p^b(h)}=0) \right). \quad (21)$$

Consider two trivial cases of $o(p^{a(0)}, p^{b(j)})$ ($0 < j \leq n^b$) and $o(p^{a(i)}, p^{b(0)})$ ($0 < i \leq n^a$), there is only one possible path. The unique solution is obtained at the initialized stage of xOrder algorithm.

For any given i and j satisfying $i > 0$ and $j > 0$, suppose xOrder has got the optimal solutions $o^*(p^{a(i-1)}, p^{b(j)})$ and $o^*(p^{a(i)}, p^{b(j-1)})$ of the subproblems to maximize $\widehat{G}(o(p^{a(i-1)}, p^{b(j)}))$ and $\widehat{G}(o(p^{a(i)}, p^{b(j-1)}))$ respectively.

Now, we will prove the solution $o^*(p^{a(i)}, p^{b(j)})$ returned by xOrder is optimal by contradiction. Suppose there exists $\bar{o}(p^{a(i)}, p^{b(j)})$ satisfying $\widehat{G}(\bar{o}(p^{a(i)}, p^{b(j)})) > \widehat{G}(o^*(p^{a(i)}, p^{b(j)}))$. There are two possible situations: $\bar{o}(p^{a(i)}, p^{b(j)})$ ends with $1, p^{a(i)}, 2, p^{b(j)}$. Without the loss of generality, we assume $\bar{o}(p^{a(i)}, p^{b(j)})$ ends with

$p^{a(i)}$ and define $\bar{o}(p^{a(i-1)}, p^{b(j)})$ by $\bar{o}(p^{a(i)}, p^{b(j)}) = \bar{o}(p^{a(i-1)}, p^{b(j)}) \oplus p^{a(i)}$. We can get the following inequation:

$$\begin{aligned}
\widehat{G}(\bar{o}(p^{a(i-1)}, p^{b(j)})) &= \widehat{G}(\bar{o}(p^{a(i)}, p^{b(j)})) - \mathbb{I}[Y_{p^{a(i)}} = 1] \cdot \left(\sum_{h:h>j} \mathbb{I}(Y_{p^{b(h)}} = 0) \right) \\
&> \widehat{G}(o^*(p^{a(i)}, p^{b(j)})) - \mathbb{I}[Y_{p^{a(i)}} = 1] \cdot \left(\sum_{h:h>j} \mathbb{I}(Y_{p^{b(h)}} = 0) \right) \\
&\geq \widehat{G}(o^*(p^{a(i-1)}, p^{b(j)}) \oplus p^{a(i)}) - \mathbb{I}[Y_{p^{a(i)}} = 1] \cdot \left(\sum_{h:h>j} \mathbb{I}(Y_{p^{b(h)}} = 0) \right), \\
&= \widehat{G}(o^*(p^{a(i-1)}, p^{b(j)}))
\end{aligned} \tag{22}$$

where \geq in the third line holds due to the update process of `xOrder` in the main text Eq. (8). This violates the assumption that $o^*(p^{a(i-1)}, p^{b(j)})$ is optimal. For the situation that $\bar{o}(p^{a(i)}, p^{b(j)})$ ends with $p^{b(j)}$, similarly we can derive $\widehat{G}(\bar{o}(p^{a(i)}, p^{b(j-1)})) > \widehat{G}(o^*(p^{a(i)}, p^{b(j-1)}))$, which violates the assumption $o^*(p^{a(i)}, p^{b(j-1)})$ is optimal. Summarizing the deduction above, we can get $o^*(p^{a(i)}, p^{b(j)})$ must be optimal.

Due to the generality of i and j , $o^*(p^a, p^b)$ with $i = n^a$ and $j = n^b$ is the global optimal solution to maximize $AUC(o(p^a, p^b))$. □

E Implementation details

E.1 Data preprocess

For the four fairness benchmark data sets, we use the preprocessed data sets from the resource of [19], while three of them (COMPAS, Adult and German) are from the resource of [12]. All categorical variables will be encoded as one-hot features. In german dataset, the situations sometimes happen that $xAUC(a, b) > xAUC(b, a)$ in training data while $xAUC(a, b) < xAUC(b, a)$ in testing data. It is odd that one group is privileged in training data but disadvantaged in testing data. So we avoid this situation when we split the data in german dataset.

For MIMIC-III data set, we preprocess the original data as in [16] for each ICU admission. EHR data of 17 selected clinical variables from the first 48 hours are used to extract features. For all clinical variables, different statistics (mean, std, etc) of different time slices are extracted to form a vector with 714 features.

E.2 Training process of model

We first train a model without any fairness regularization to obtain the unadjusted result. The logistic regression model is optimized by gradient descent, with learning rate of 0.1. The model is trained for at most 100 epochs. If the training loss failed to reduce after 5 consecutive epochs, the training will be stopped. For bipartite rankboost model, the number of estimators is 50 and learning rate is 1.0, which are the same as in the experiments of [19].

Post-logit With transformation function $f(x) = \frac{1}{1+e^{-(\alpha x + \beta)}}$, We follow the same procedure in [19] to optimize empirical disparity ($\Delta xAUC$ or ΔPRF) with fixed $\beta = -2$. The value of α is chosen from $[0, 10]$. However, we find that this setting can not obtain equal $xAUC$ or PRF in MIMIC data set which has not been used in [19]. So we change β from $[-1, -3]$ and α from $[0, 10]$ when optimizing the empirical disparity on MIMIC data set.

Corr-reg For corr-reg, we train models of the same structure with various weights of fairness regularization. Since the correlation regularization is only an approximation of the pairwise ranking disparity, we use the corresponding ranking fairness metrics ($\Delta xAUC$ or ΔPRF) as the criterion to determine the range of weights. We initialize the weight to be 0 (equivalent to unadjusted) and increase it until the average ranking disparity on training data is lower than 0.01 or does not decrease in 2 consecutive steps. We also apply this strategy to `xOrder`.

E.3 Computing infrastructure and consumption

We run `xOrder` on computer with Intel i7-9750H CPU ($@2.6\text{GHz} \times 6$) and 16 gb RAM. We report the average running time of 10 evaluation runs in Table 2.

Table 2: Time consumption of xOrder

Data set	COMPAS	Adult	German	Framingham	MIMIC-III
n	6,167	30,162	1,000	4,658	21,139
time/sec	3.27 ± 0.44	38.67 ± 1.28	0.45 ± 0.43	2.11 ± 0.41	16.71 ± 0.61

Table 3: Summary of ranking fairness metrics on unadjusted result over 10 repeat experiments

Data set	Logistic Regression		Bipartite Rankboost	
	$\Delta xAUC$	ΔPRF	$\Delta xAUC$	ΔPRF
COMPAS	0.195 ± 0.024	0.119 ± 0.015	0.206 ± 0.027	0.134 ± 0.019
Adult	0.089 ± 0.011	0.040 ± 0.009	0.055 ± 0.016	0.020 ± 0.008
German	0.161 ± 0.046	$0.060 \pm 0.031^*$	$0.128 \pm 0.091^*$	$0.056 \pm 0.030^*$
Framingham	0.270 ± 0.022	0.124 ± 0.018	0.285 ± 0.075	0.136 ± 0.038
MIMIC, mortality-gender	0.050 ± 0.012	0.020 ± 0.009	$0.018 \pm 0.011^*$	$0.010 \pm 0.010^*$
MIMIC, mortality-ethnicity	$0.025 \pm 0.012^*$	$0.024 \pm 0.012^*$	$0.019 \pm 0.013^*$	$0.016 \pm 0.010^*$
MIMIC, prolonged LOS-gender	$0.025 \pm 0.020^*$	$0.022 \pm 0.014^*$	$0.019 \pm 0.010^*$	$0.012 \pm 0.011^*$
MIMIC, prolonged LOS-ethnicity	0.043 ± 0.008	0.021 ± 0.008	0.025 ± 0.010	0.012 ± 0.007

F Additional Experiment Results

F.1 Ranking fairness analysis on unadjusted result

We report the ranking fairness metrics($\Delta xAUC$, ΔPRF) in Table 3. Large $\Delta xAUC$ and ΔPRF are observed on the four benchmark data sets. We use t-test with p-value as 0.0001 to evaluate whether the average $\Delta xAUC$ and ΔPRF do not equal to 0. We mark the results which do not pass the test with * on the table. For MIMIC-III data set, disparities are significant with certain $Y - A$.

F.2 Complete experiment results

The complete results with model-metric combinations of logistic regression- $\Delta xAUC$, logistic regression- ΔPRF , bipartite rankboost- $\Delta xAUC$ and bipartite rankboost- ΔPRF are shown in Figure 8, 9, 10 and 11 respectively. The source codes to reproduce these results on public data sets are at <https://github.com/xOrder-code/xOrder>. Most subfigures in Figure 8, 9 and 10 have been discussed in the main text. In addition, we can find that all the methods do not work well on German data set, especially with bipartite rankboost model. It may be due to its relatively small sample size. With different training-test split, the ranking disparities on training and test data are obviously different.

F.3 Further comparison between xOrder and post-logit

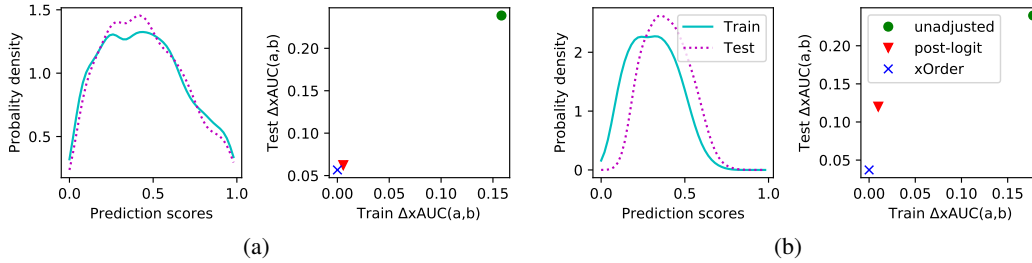


Figure 6: Result analysis on COMPAS data set with $\Delta xAUC$ metric. (a) illustrates the result with logistic regression model. (b) illustrates the result with bipartite rankboost model. The left part of each sub-figure is the distributions of prediction scores on training and test data, the right part of each sub-figure plots $\Delta xAUC$ on training data and testing data.

According to Figure 10, post-logit fails to achieve $\Delta xAUC$ as low as xOrder with bipartite rankboost model, while both methods can achieve low $\Delta xAUC$ with logistic regression. To analyze this phenomenon, we use COMPAS and adult as examples in Figure 6 and 7. For different models, we illustrate the distributions of prediction scores S on training and testing data. We further plot $\Delta xAUC$ on training data versus $\Delta xAUC$ on testing data. With logistic regression model, the distributions of S on training and testing data are closed to each other. In this situation, the transform relations learnt from post-logit and xOrder can both obtain results

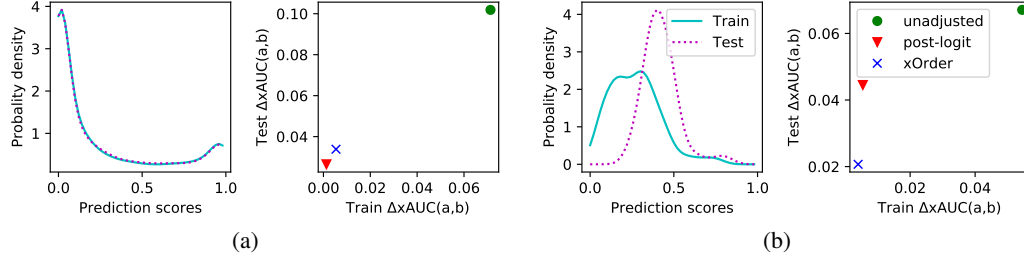


Figure 7: Result analysis on Adult data set with $\Delta xAUC$ metric. (a) illustrates the result with logistic regression. (b) illustrates the result with bipartite rankboost model. The left part of each sub-figure is the distributions of prediction scores on training and test data, the right part of each sub-figure plots $\Delta xAUC$ on training data and test data.

with low $\Delta xAUC$ on testing data. While the distributions of the scores S on training and testing data become obviously different, the function learned from post-logit may not be generalized well on testing data to achieve low $\Delta xAUC$. Similar results can be observed on adult data set in Figure. 7. These phenomena occur in repeat experiments on both data sets. This suggests that $xOrder$ is more robust to such distribution difference.

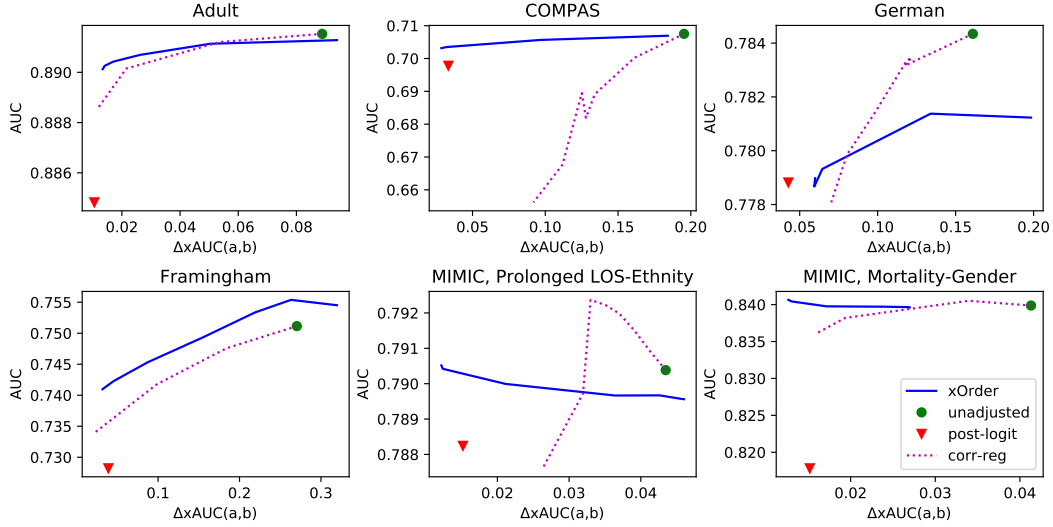


Figure 8: AUC- $\Delta xAUC$ with logistic regression model.

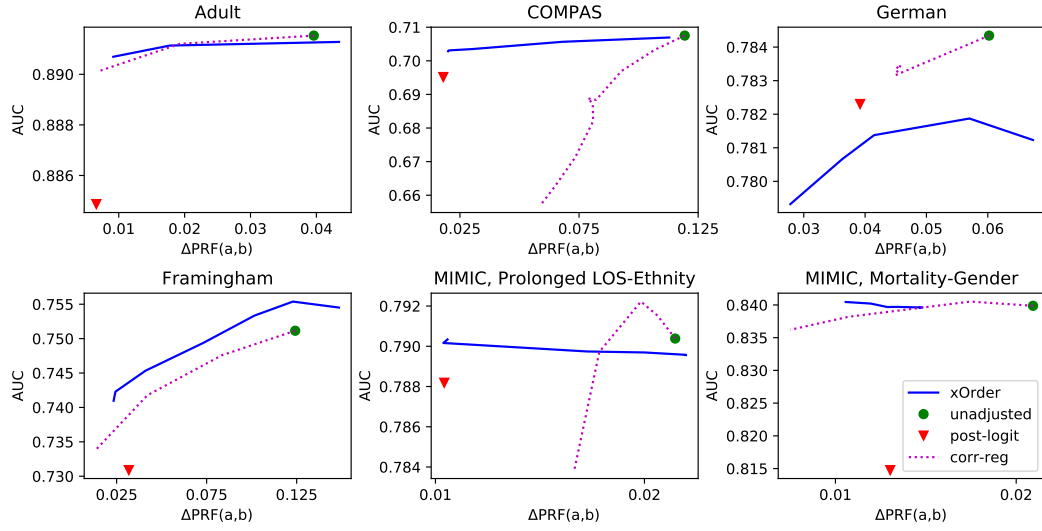


Figure 9: AUC- ΔPRF with logistic regression model.

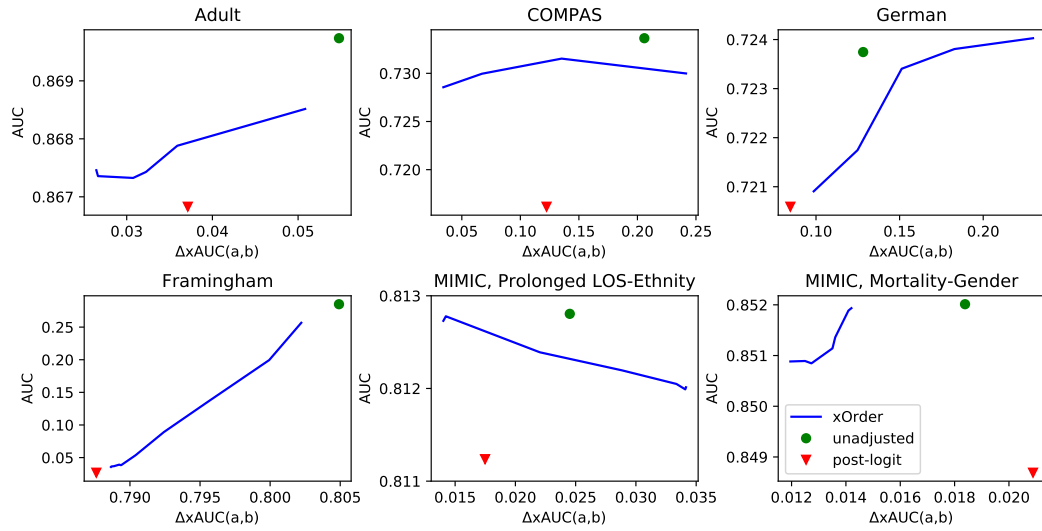


Figure 10: AUC- $\Delta x\text{AUC}$ with bipartite rankboost model.

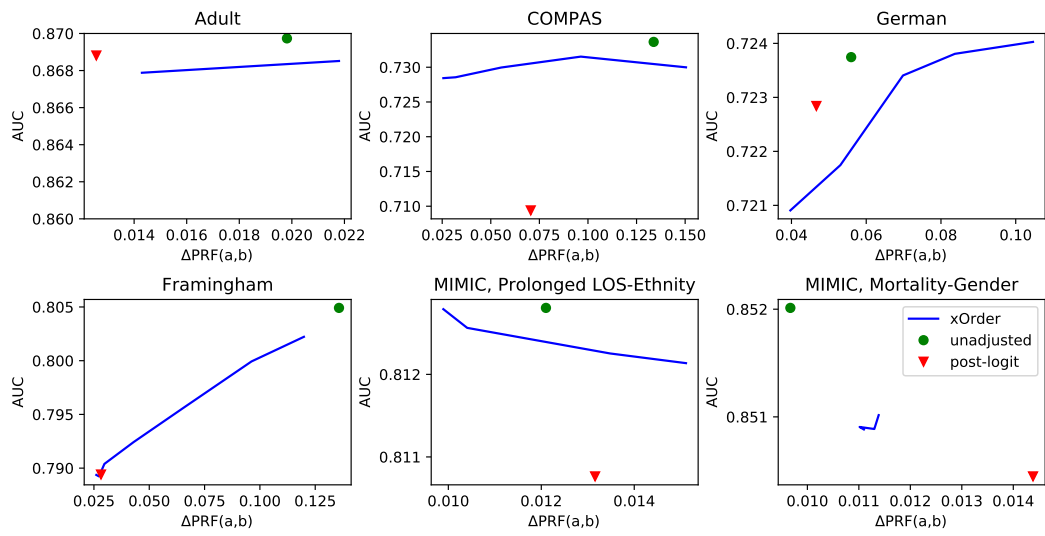


Figure 11: AUC- ΔPRF with bipartite rankboost model.