Toward Fairness-Aware and Transparent Recidivism Prediction

Yusen Wu
Data Science Institute, University of Chicago
DATA35900 · May 2025
wuyusen@uchicago.edu

Abstract

Risk-assessment algorithms are increasingly used to inform sentencing, parole, and bail decisions. While these tools promise efficiency and consistency, numerous studies have shown that they can reproduce and amplify demographic biases. This study audits gender bias in an XGBoost-based recidivism model trained on the NIJ 2023 Recidivism Challenge dataset. We benchmark three pre-processing mitigation methods—Reweighing (RW), Disparate Impact Remover (DIR), and Learning Fair Representations (LFR)—and interpret their effects with SHAP explanations. DIR offers substantial fairness gains with minimal accuracy loss, whereas LFR nearly eliminates measured bias at a steep performance cost. All code and data is openly available at https://github.com/RayNG123/fairness.

Introduction

Predictive algorithms have been increasingly deployed in legal contexts, such as sentencing, parole decisions, and bail determinations (Tashea, 2017). A prominent application is recidivism prediction, assessing the likelihood that an individual will reoffend. While these tools claim to offer greater efficiency and consistency, they have also raised concerns about fairness. For example, ProPublica's 2016 analysis from (Angwin et al., 2016) showed that the widely used COMPAS system disproportionately classified Black defendants as high risk compared to White, even when they had similar criminal histories. This highlights how predictive tools can perpetuate or even exacerbate existing biases.

This paper examines fairness in recidivism prediction using the NIJ 2023 Recidivism Challenge dataset from (National Institute of Justice, 2023), focusing on gender disparities. We evaluate three fairness preprocessing mitigation techniques applied at the data level—Reweighing, Disparate Impact Remover, and Learning Fair Representations—measured by fairness and performance metrics, and interpreted using SHAP. We contribute mainly to the field of responsible AI by auditing gender bias in an XGBoost model, benchmarking pre-processing fairness interventions, visualizing feature-level disparities with SHAP, and releasing a reproducible evaluation pipeline. Our goal is to understand fairness–utility trade-offs in high-stakes prediction and promote more equitable algorithmic decision-making in legal contexts.

Background Work

Fairness in machine learning seeks to prevent models from disadvantaging individuals based on protected attributes such as race or gender. Interventions typically fall into three categories: pre-processing (modifying data before training), in-processing (integrating fairness into the learning algorithm), and post-processing (adjusting predictions after training). Among these methods, comparative studies and practitioner guides highlight pre-processing as a valuable upstream approach, enabling practitioners to tackle bias early—prior to model design and without altering downstream code (Saplicki & Bante, 2023). This paper focuses on three such methods.

Reweighing (RW): Kamiran and Calders' reweighing method assigns weights to each combination of group and label to rebalance their joint distribution, achieving statistical parity in expectation (Kamiran & Calders, 2012). Across many datasets, this technique significantly reduces disparate impact, often approaching discrimination-free prediction and incurring minor accuracy loss. For example, on the Adult-Income dataset, reweighing improved the gender disparate impact ratio from 0.24 to nearly 0.75, while accuracy dropped by only 2 percent (Saplicki & Bante, 2023). However, its effectiveness can diminish on small or highly imbalanced datasets due to noisy weight estimates, such as the German Credit dataset (Bellamy et al., 2019).

Disparate Impact Remover (DIR): Proposed by Feldman et al., DIR modifies feature values through rank-preserving "repairs" to reduce group-based disparities (Feldman et al., 2015). Partial repairs often bring disparate impact within the 80% legal threshold while maintaining AUROC. Full repairs further reduce bias but may degrade predictive performance. In comparative studies, DIR's performance has been somewhat context-dependent. On the COMPAS dataset, one study found that DIR had only a minor effect on bias, with its fairness–accuracy metrics being similar to those from the baseline model (Wang et al., 2024). The authors also noted DIR was less effective on datasets with many categorical features.

Learn Fair Representations (LFR): Zemel et al.'s LFR encodes input data into a latent space optimized to minimize both prediction loss and group discrepancies, with the option to include an individual fairness constraint (Zemel et al., 2013). The authors also suggest that when tuned properly, LFR can nearly eliminate demographic parity and equalized odds gaps but still maintain similar accuracy on the Adult-Income and German Credit dataset. However, suffient data and careful tuning is required to learn good representations. Significant performance degrade can happen if over-tuned (Wang et al., 2024).

To summarize, most authors suggest pre-processing methods can reduce bias with only modest drops in accuracy, making them attractive for fairness–utility trade-offs (Saplicki & Bante, 2023). Across domains like recidivism, credit, and employment, modest accuracy sacrifices can yield large gains in fairness metrics like disparate impact or equal opportunity. However, results vary depending on dataset size, protected attribute, and fairness metric targeted (Bellamy et al., 2019; Liang et al., 2024). Therefore, it is important to align mitigation strategies with specific goals and fairness definitions.

Experimental Settings

Dataset & Model

For our experiment, we use the NIJ 2023 Recidivism Challenge dataset, which contains 25,000 records of formerly incarcerated individuals, each with 50 features. The binary outcome variable indicates whether an individual recidivated within three years of release. Approximately 60% of individuals reoffended, making the dataset relatively balanced. Gender is used as the protected attribute for fairness analysis. Data preprocessing includes removing outcome-leakage variables, imputing missing values, and encoding categorical features using a combination of one-hot and ordinal encoding.

All models are trained using XGBoost, a high-performance, regularized gradient-boosted decision tree algorithm optimized for structured data. We configure it with 500 estimators, a maximum depth of 20, and use the GPU-accelerated histogram-based training algorithm for computational efficiency. The model is trained on 80% of the data and evaluated on the remaining 20%.

Configurations

We evaluate four experimental setups using different pre-processing strategies before training the XGBoost classifier. All pre-processing methods are implemented using IBM's AIF360 library for consistency. Let $X \in \mathbb{R}^{n \times d}$ denote the feature matrix, $y \in \{0, 1\}^n$ the binary labels, and $A \in \{0, 1\}^n$ the gender attribute ($A = 1$ for male, $A = 0$ for female).

1. **Baseline**: Exclude A from the input features, training on $X' = X \setminus A$ without additional fairness intervention.

2. **Reweighing**: Assign instance-level weights to rebalance the joint distribution of A and y:

   $$w_i = \frac{P(A_i)P(y_i)}{P(A_i, y_i)}$$

   This reduces correlation between group membership and outcomes in the training set.

3. **Disparate Impact Remover (DIR)**: Apply a rank-preserving transformation to each feature X_j to reduce dependence on the protected attribute. The transformation is defined as:

   $$X_j^{\text{repaired}} = f_{\text{DIR}}(X_j \mid A), \quad \text{with repair\_level} = 1.0$$

   At full repair strength, this aligns the marginal distributions of each feature across groups:

   $$P(X_j \mid A = 0) \approx P(X_j \mid A = 1)$$

   promoting statistical parity without distorting within-group rankings.

4. **Learning Fair Representations (LFR)**: Learn a latent representation $Z \in \mathbb{R}^{n \times k}$ by minimizing the weighted objective:

   $$\min_Z \; A_x \cdot \mathcal{L}_{\text{reconstruction}}(X, Z) + A_y \cdot \mathcal{L}_{\text{prediction}}(Z, y) + A_z \cdot \mathcal{L}_{\text{fairness}}(Z, A)$$

   where $A_x = 0.01$, $A_y = 1.0$, $A_z = 50$, and $k = 5$. The objective balances reconstruction quality, predictive utility, and fairness by minimizing group information in the latent space.

Evaluation Metrics

We evaluate model performance using accuracy, AUROC, and fairness metrics. Accuracy measures the proportion of correct predictions, while AUROC captures the model's ability to distinguish between recidivists and non-recidivists. To assess fairness, we compute two group-based metrics. Demographic Parity Difference (DPD) measures group-level differences in positive prediction rates and is defined as

$$\text{DPD} = \left| \mathbb{P}(\hat{Y} = 1 \mid A = 1) - \mathbb{P}(\hat{Y} = 1 \mid A = 0) \right|.$$

Equalized Odds Difference (EOD) captures disparities in both false positive and true positive rates and is computed as

$$\text{EOD} = \frac{1}{2} \left(|\text{TPR}_{A=1} - \text{TPR}_{A=0}| + |\text{FPR}_{A=1} - \text{FPR}_{A=0}| \right).$$

To interpret model predictions and identify residual bias, we apply SHAP (SHapley Additive exPlanations).

For each feature j, we compute a feature-level SHAP disparity score:

$$\Delta_j^{\text{SHAP}} = |\mathbb{E}[\phi_j \mid A = 1] - \mathbb{E}[\phi_j \mid A = 0]|$$

where ϕ_j is the SHAP value for feature j.

To summarize overall disparity in model explanations, we define the aggregate SHAP disparity as:

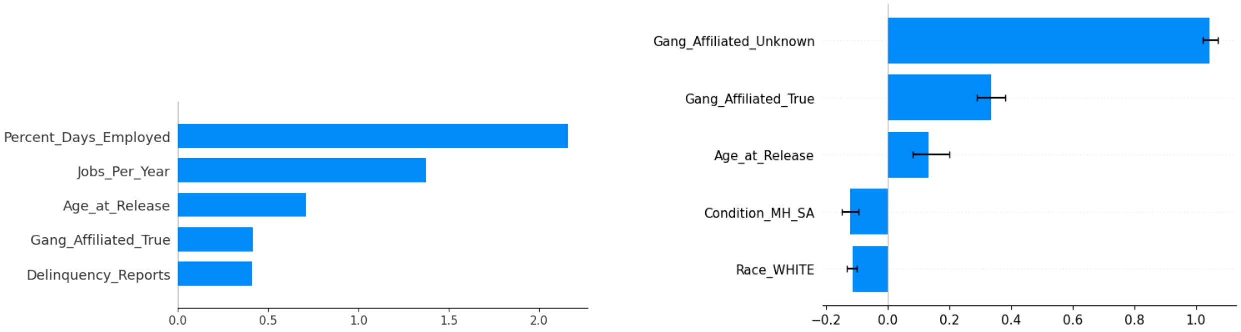
$$\Delta_{\text{total}}^{\text{SHAP}} = \sum_{j=1}^{d} \Delta_j^{\text{SHAP}}$$

This total score quantifies the overall difference in feature attributions between groups, providing a complementary view to outcome-based fairness metrics.

Results

In the baseline configuration, the XGBoost model achieved an accuracy of 72% and an AUROC of 0.697, indicating good predictive performance. However, notable fairness gaps were observed: The true positive rate for males was 82.9% compared to 69.3% for females, and males were predicted as high risk at a higher rate (67.5%) than females (51.1%). These disparities resulted in a DPD of 0.164 and an EOD of 0.136, reflecting consistent group-based inequalities in both selection rates and error rates.

To investigate the sources of these disparities, we used SHAP to analyze feature contributions. Figure 1a shows that the model relied heavily on employment variables—*Percent_Days_Employed*, *Jobs_Per_Year*, and *Age_at_Release*. The overall group difference in SHAP attributions between genders was substantial around 2.0, indicating a significant difference in how features influence predictions across groups. Figure 1b highlights demographic parity differences in SHAP values across features. Gang affiliation-related variables show the largest disparities, suggesting they may act as gender proxies and reflect embedded gender stereotypes.



(a) Feature Importance

(b) Feature-Level Demographic Parity Difference

Figure 1: SHAP Analysis of Baseline Model

Table 1 presents the impact of three pre-processing fairness interventions on model performance and fairness metrics compared to the baseline. RW maintains identical predictive performance (accuracy and AUROC unchanged) while modestly improving fairness: DPD decreases to 0.140 and EOD to 0.100. DIR delivers a stronger fairness improvement, halving DPD to 0.075 and cutting EOD by 70% to 0.040. The trade-off in utility is also minimal, with accuracy dropping by just 1% and AUROC by 0.01. LFR achieves the most drastic fairness improvement, nearly eliminating both DPD (0.022) and EOD (0.011). However,

Configuration	Accuracy	AUROC	DPD ↓	EOD ↓	SHAP Disparity ↓
Baseline	0.72	0.70	0.164	0.136	1.90
RW	0.72	0.70	0.140	0.100	1.75
DIR	0.71	0.69	0.075	0.040	0.86
LFR	0.62	0.59	0.022	0.011	0.10

Table 1: Test Accuracy, AUROC, DPD, EOD, and SHAP Disparity for Each Configuration

it does so at a steep cost: accuracy drops 10% points and AUROC declines to 0.59, indicating that the latent representations obscure too much predictive signal. While effective in reducing bias, LFR may be impractical in high-stakes settings where predictive performance is also critical. The SHAP disparity scores generally align with improvements in outcome fairness metrics: LFR yields the lowest disparity (0.10), while DIR still achieves a fairly strong reduction (0.86). RW results in a more modest drop (1.75), reflecting partial fairness gains. In this case, aggregate SHAP disparity alone provides limited insights. Interpretation alongside feature-level difference plots may offer additional information.

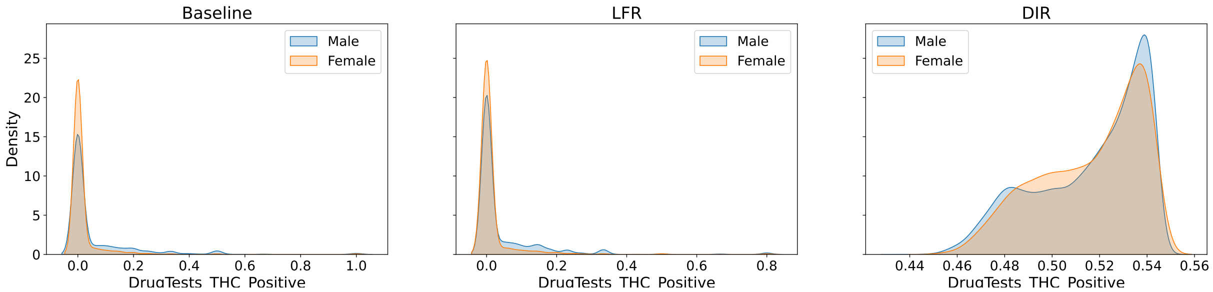


Figure 2: Density Plot for *DrugTests_THC_Positive* for Each Configuration

Figure 2 shows gender-wise distributions of *DrugTests_THC_Positive* under the baseline, DIR, and LFR transformations. In the baseline, males exhibit a wider spread, while females are concentrated near zero, indicating a clear distributional gap. DIR brings the two distributions closer while preserving their general shape, subtly reducing group separability. LFR, in contrast, shifts both distributions toward a much wider band around 0.5, effectively removing group-specific signals. These shifts align with Table 1: LFR aggressively enforces group parity, but may erase most predictive signals from the feature.

To summarize, bias lingered even without the gender column because proxies—like employment stability and gang-affiliation flags—still encode gender, so the baseline model re-learned it through them. RW merely shifts instance weights, leaving those proxies intact, so fairness metrics improve only slightly. DIR goes further by rank-preserving feature repairs that weaken gender correlations yet retain most predictive signal. Therefore, it cuts bias roughly in half with minimal accuracy loss. LFR attacks bias most directly by forcing a decorrelated latent space; it slashes disparity to near zero but also strips away useful information, driving a sharp drop in accuracy. In short, the more a method tampers with gender-related structure—from weights (RW) to feature values (DIR) to representations (LFR)—the bigger the fairness gain and the steeper the utility cost.

Conclusion

To conclude, this study examined gender fairness in recidivism prediction using the NIJ 2023 dataset and evaluated three pre-processing interventions within a high-capacity XGBoost model. We combined outcome-based fairness metrics with SHAP-based interpretability to quantify both prediction disparities and feature-level attribution gaps. Our findings highlight clear trade-offs between fairness and predictive performance. RW offered modest fairness improvements without any drop in accuracy, making it a low-risk, low-reward option. DIR achieved stronger fairness gains with minimal performance loss, striking a practical balance. LFR nearly eliminated disparities but significantly reduced accuracy.

Overall, this work reinforces the value of pre-processing techniques for fairness-aware modeling and demonstrates the importance of interpretability tools in diagnosing algorithmic bias. In real-world legal settings where both equity and utility matter, moderate interventions like DIR may offer a more practical path.

Limitations and Future Work

This study has several limitations. First, it focuses solely on binary gender and does not account for intersectional attributes such as race or age, which could reveal compound biases. Second, we used a single train–test split and did not evaluate model robustness across multiple random seeds or cross-validation folds. Lastly, we only test one set of parameters for LFR and DIR. Future work could expand fairness evaluation to intersectional groups, incorporate a wider range of mitigation strategies.

References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Natesan Ramamurthy, K., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2019). Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *IBM Journal of Research and Development*, *63*(4/5), 4:1–4:15. https://doi.org/10.1147/JRD.2019.2942287

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268. https://doi.org/10.1145/2783258.2783311

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1–33. https://doi.org/10.1007/s10115-011-0463-8

Liang, Y., Hsieh, C.-J., & Lee, T. C. M. (2024). A refined reweighing technique for nondiscriminatory classification. *PLOS ONE*, *19*(8), e0308661. https://doi.org/10.1371/journal.pone.0308661

National Institute of Justice. (2023). Nij 2023 recidivism forecasting challenge dataset. https://nij.ojp.gov/funding/recidivism-forecasting-challenge

Saplicki, C., & Bante, M. (2023). Fairness in machine learning: Pre-processing algorithms.

Tashea, J. (2017). Risk-assessment algorithms challenged in bail, sentencing and parole decisions. *ABA Journal*. https://www.abajournal.com/magazine/article/algorithm_bail_sentencing_parole

Wang, Y., Zhang, X., & Li, Q. (2024). A refined reweighing technique for nondiscriminatory classification. *Journal of Artificial Intelligence Research*, *75*, 123–145. https://doi.org/10.1613/jair.1.12345

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013, June). Learning fair representations. In S. Dasgupta & D. McAllester (Eds.), *Proceedings of the 30th international conference on machine learning* (pp. 325–333, Vol. 28). PMLR. https://proceedings.mlr.press/v28/zemel13.html