Technical Report: Mitigating Bias

Rohan Tikotekar
rtiko001@ucr.edu
Univeristy of California Riverside
Riverside, California, USA

ACM Reference Format:
Rohan Tikotekar. 2018. Technical Report: Mitigating Bias. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Project Goals

Bias in machine learning models is a significant challenge, particularly in high-stakes applications like criminal justice. This project aims to analyze and mitigate bias in the COMPAS dataset, which is widely used to predict recidivism risk. Our goal is to evaluate various fairness metrics and apply bias mitigation techniques to improve the equity of model predictions while maintaining performance. The COMPAS dataset contains information about individuals assessed for recidivism risk, including demographic details, prior offenses, and recidivism outcomes. Given the high impact of predictive modeling in legal and correctional systems, ensuring fairness in such models is crucial to prevent discrimination and bias in decision-making. Machine learning models, when left unmitigated, often inherit biases present in historical data, exacerbating systemic disparities rather than alleviating them. By implementing fairness-aware strategies, we seek to create predictive models that are more equitable and just in their outcomes. Bias in criminal justice predictions can have profound societal implications, as algorithmic misclassifications can lead to harsher sentencing or unnecessary supervision for specific demographic groups. This study explores the impact of bias in predictive modeling and evaluates approaches to minimize its effects while maintaining the overall efficacy of risk assessment models.

2 Dataset

The dataset used in this study includes various attributes that are pertinent to the assessment of recidivism risk, with key features such as race, age, number of prior offenses, and two-year recidivism outcomes. These attributes provide the foundation for analyzing bias in predictive models, as disparities in recidivism predictions can disproportionately affect different demographic groups. If a model disproportionately classifies individuals from a particular race as high-risk when they are not, it can contribute to systemic discrimination. The COMPAS dataset has been historically criticized for racial bias, with previous analyses showing that Black defendants were more likely to be misclassified as high-risk compared to White defendants. Given the implications of this bias in criminal justice, it is essential to conduct a thorough analysis of the dataset and identify any disparities that may arise due to underlying biases in historical data. Without intervention, models trained on such datasets can inadvertently reinforce existing social inequalities, making fairness-aware modeling a crucial component of responsible AI development. By ensuring that data preprocessing steps are designed to detect and address imbalances, we can work toward developing predictive models that produce more equitable outcomes. Fairness-aware data collection and preprocessing strategies must be implemented to reduce potential biases that could skew predictions, allowing machine learning models to make informed decisions that do not disproportionately harm specific groups. This step is critical in ensuring that machine learning algorithms are not perpetuating historical patterns of discrimination but instead contributing to fairer and more just decision-making processes in criminal justice applications.

3 Evaluation Metrics

To evaluate model performance and fairness, we used several key metrics that allow for a comprehensive assessment of predictive strength and equity. Accuracy measures the overall correctness of the model in classifying outcomes, while the F1-score balances precision and recall, ensuring that the model does not over-prioritize one over the other. Recall is a key measure that assesses the model's ability to identify true positives, whereas precision evaluates the proportion of correctly identified positive cases. In the context of fairness, demographic parity difference measures whether positive predictions are proportionally distributed across groups, ensuring equal representation. Equalized odds is another critical metric that compares true positive and false positive rates across demographic groups to determine whether the model treats them fairly. These metrics allow us to quantify the extent of bias in model predictions and ensure that mitigation strategies are implemented effectively. Without fairness constraints, models trained on biased datasets tend to favor majority groups, leading to inequitable decision-making. Evaluating models across these multiple dimensions is essential for identifying disparities in prediction outcomes and ensuring that any fairness interventions lead to genuine improvements rather than surface-level adjustments. A model that fails to maintain balance in these metrics can perpetuate existing inequities rather than correcting them, making fairness analysis an essential step in developing responsible AI solutions. By integrating fairness metrics into model evaluation, we can systematically measure the impact of different bias mitigation techniques and ensure that our models achieve both high performance and equitable decision-making outcomes.

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

4 Model Performace Pre-Mitigation

Before applying bias mitigation techniques, multiple machine learning models were used to evaluate fairness and overall performance. The Naive Bayes model served as a baseline, revealing disparities in recall and fairness metrics, with non-white individuals experiencing a higher false positive rate, indicating racial bias. Logistic Regression, commonly used for binary classification, was employed to assess disparities in predictions between demographic groups. The Decision Tree model, known for its interpretability, was included in the analysis, though it exhibited fairness concerns similar to other models. The Random Forest model, a more robust approach, demonstrated disparities in precision and recall across demographic groups. The pre-mitigation results highlighted the existing biases in the dataset, showing that non-white individuals were more likely to be misclassified as high-risk, which can have severe consequences in real-world applications. The disparities in fairness metrics underscored the need for bias mitigation techniques to improve the model's equitable treatment of different demographic groups while maintaining its predictive power. The presence of such disparities suggests that the dataset itself contains embedded biases that machine learning models learn and reinforce. If left unaddressed, these biases may lead to outcomes that disproportionately impact marginalized communities, reinforcing existing inequalities in the criminal justice system. Evaluating pre-mitigation performance enables us to quantify the extent of bias and provides a benchmark for measuring the effectiveness of mitigation strategies. Without intervention, machine learning algorithms may unintentionally perpetuate societal biases, making it imperative to develop models that account for fairness considerations in high-risk decision-making environments.

5 Model Performance Post Mitigation

To mitigate bias, various fairness techniques were applied, each demonstrating different trade-offs between predictive performance and fairness. The post-mitigation Naive Bayes model showed an improvement in accuracy from 73.53% to 74.36%, with the F1-score increasing from 56.19% to 58.14%, indicating a better balance between precision and recall. The recall improved from 39.52% to 41.45%, reflecting a higher capability of identifying true positives, while precision remained high at 97.2%, ensuring minimal false positives. This improvement suggested that the post-processing strategy enhanced the model's ability to fairly classify positive cases while maintaining its reliability. The SMOTE resampling technique, applied to the Logistic Regression model, resulted in similar accuracy (69.23% to 69.09%), suggesting minimal impact on overall correctness. However, the demographic parity difference decreased from 0.1672 to 0.1571, improving fairness by ensuring a more balanced distribution of positive predictions between demographic groups. The False Positive Rate (FPR) for non-white individuals decreased, reducing racial bias and making the model less likely to wrongfully classify them as high-risk. However, a small trade-off was observed as the accuracy and F1-score for white individuals slightly decreased, indicating that the model adjustments redistributed predictive power to achieve a more equitable outcome. Threshold adjustments applied to the Decision Tree model led to a slight decrease in accuracy from 68.68% to 67.36%, along with a precision drop from 62.03% to 59.97%. However, recall increased from 69.84% to 72.26%, indicating that the model became more sensitive in detecting true positives across groups. The most notable improvement was in the demographic parity difference, which significantly decreased from 0.6911 to 0.0653, suggesting that the model's predictions became more balanced across demographic groups. The Exponentiated Gradient Reduction method, applied to the Random Forest model, demonstrated similar trends, with a slight reduction in accuracy from 65.07% to 64.59%, precision dropping from 59.48% to 58.89%, and recall decreasing from 58.71% to 58.23%. However, the demographic parity difference improved significantly from 0.1315 to 0.0473, indicating that the model's fairness was enhanced. Additionally, the True Positive Rate and False Positive Rate differences significantly decreased, ensuring that predictions were more equitable between demographic groups. These results showed that while there was a minor loss in predictive power, the model became considerably fairer in its classification outcomes, reducing systemic bias in recidivism predictions.

6 Results and Discussion

The results indicate that bias mitigation strategies effectively reduced racial bias in predictions while maintaining overall predictive performance. Different techniques demonstrated varying levels of effectiveness, but all approaches contributed to making the model fairer. SMOTE resampling reduced false positives for non-white individuals, leading to more equitable outcomes by ensuring that individuals were not disproportionately classified as high-risk based on race. While this improved demographic parity, it slightly decreased accuracy for the majority group, highlighting the common trade-off in fairness-aware machine learning—enhancing fairness for disadvantaged groups often comes at the cost of minor reductions in predictive performance for the majority. Threshold adjustments further demonstrated this trade-off, as accuracy decreased slightly while recall improved, ensuring that fewer high-risk individuals were incorrectly classified as low-risk, which is particularly critical in criminal justice applications. The Exponentiated Gradient Reduction method showed the most significant improvement in fairness metrics, particularly by reducing disparities in equalized odds, minimizing differences in True Positive Rate (TPR) and False Positive Rate (FPR) across demographic groups. Although this approach led to a minor decrease in overall accuracy, it ensured that predictions were made more equitably, reducing racial disparities in risk assessments. These results highlight the complexity of bias mitigation in machine learning, where improving fairness often requires carefully managing trade-offs with performance. While SMOTE and threshold adjustments provided moderate fairness improvements with minimal performance impact, more aggressive in-processing techniques, like Exponentiated Gradient Reduction, significantly enhanced fairness at a slight cost to accuracy and precision. Additionally, fairness-aware techniques had varying effects on demographic groups—while SMOTE improved fairness for non-white individuals, it slightly reduced precision for the white group, emphasizing the challenge of balancing fairness across populations. Similarly, threshold adjustments improved recall across both groups but had a more pronounced impact on reducing disparities in high-risk classifications for non-white individuals, while

Exponentiated Gradient Reduction provided the most consistent fairness improvements but introduced minor accuracy reductions across all groups. These findings underscore the importance of a careful, context-driven approach to bias mitigation, where models must be rigorously tested across multiple metrics to ensure that fairness improvements do not introduce unintended negative consequences. The study highlights that while no single method can completely eliminate bias, a combination of mitigation strategies can lead to substantial fairness improvements while maintaining predictive accuracy. Considering both fairness and performance is essential when developing machine learning models for high-stakes applications like criminal justice. Bias mitigation techniques can help reduce systemic disparities in predictive outcomes, but they must be implemented thoughtfully to avoid introducing new imbalances. The study also underscores the necessity for continuous monitoring and evaluation of model fairness, as biases can shift over time with changes in data distributions. Future work should explore hybrid approaches that integrate multiple bias mitigation techniques to maximize fairness while maintaining accuracy. Moreover, policymakers and practitioners should be actively involved in developing and deploying fairness-aware models to ensure alignment with ethical and legal standards. Machine learning models hold the potential to support fairer decision-making processes, but achieving this requires a committed effort to identify and mitigate bias at every stage of model development. Through rigorous evaluation and responsible implementation of fairness-aware techniques, we can work toward more equitable machine learning models that contribute to reducing disparities in criminal justice outcomes and beyond.

7 Conclusion

The findings of this study demonstrate that bias mitigation strategies can significantly improve fairness in machine learning models while maintaining reasonable predictive performance. While trade-offs between accuracy and fairness are inevitable, techniques such as SMOTE resampling, threshold adjustments, and Exponentiated Gradient Reduction effectively reduced disparities in demographic parity and equalized odds, making predictions more equitable across racial groups. The study highlights the importance of carefully selecting bias mitigation techniques based on the specific context and ethical considerations of the application. Given the high stakes of criminal justice predictions, ongoing monitoring and refinement of these models are crucial to prevent unintended biases from resurfacing. Future work should explore hybrid approaches that integrate multiple fairness-aware techniques to maximize both equity and predictive power. By implementing responsible AI practices and involving policymakers in model development, we can work toward more ethical and just decision-making systems that mitigate systemic biases rather than reinforcing them.