Research Project - Tabular DS

Danielle Shrem and Jonathan Mandl

February 2025

1 Abstract

Machine learning models are increasingly used in high-stakes decision-making systems—such as loan approvals, hiring, and healthcare—but even when models appear unbiased, their performance may differ across demographic groups, potentially reinforcing hidden biases and undermining trust in AI-driven systems. Existing fairness mitigation techniques typically operate at a specific stage of the modeling pipeline and may not generalize well across diverse data types. In this work, we develop an automated framework that, given a dataset with sensitive attributes, applies multiple fairness-enhancing methods at the preprocessing, model training, and postprocessing stages. Our system systematically tests various combinations of techniques and presents the user with Pareto-optimal configurations based on both accuracy and fairness metrics, enabling tailored decision-making. We evaluate our framework on four widely used fairness benchmark datasets and demonstrate that our approach outperforms baseline fairness methods. Our system integrates multiple methods to achieve a balanced trade-off between predictive performance and fairness, thereby promoting more ethical and trustworthy AI-driven decision-making.

2 Problem Description

In high-stakes decision-making systems—such as loan approvals, hiring, and healthcare—ensuring fairness is critical. Even when overall model performance appears high, disparities in error rates and predictions across demographic groups can lead to systematic discrimination. For instance, a model may have high accuracy overall but produce significantly different positive prediction rates for different groups. This discrepancy can reinforce hidden biases, erode trust in AI systems, and result in unjust outcomes

Two common fairness metrics used to evaluate these disparities are Demographic Parity and Equalized Odds. Demographic Parity requires that different demographic groups receive positive outcomes at similar rates. In contrast, Equalized Odds ensures that error rates—specifically, the true positive and false positive rates—are consistent across groups. Although many bias mitigation techniques aim to improve these fairness metrics, existing approaches typically address only one stage of the data science pipeline. Most bias mitigation algorithms are applied at one of three stages in the model pipeline: preprocessing, training, or postprocessing. In addition, many methods perform well in specific cases but may not generalize to other scenarios or data types. Often, the only way to determine whether a method will work effectively for a particular dataset is through empirical evaluation. We propose an automated approach to test various bias mitigation methods across all stages of the modeling pipeline. Our strategy aims not only to enhance fairness but also to maintain overall model performance across diverse real-world scenarios.

3 Solution Overview

Our solution is an automated framework that evaluates and selects bias mitigation techniques across all stages of the model pipeline—preprocessing, training, and postprocessing—to simultaneously optimize for both model performance and fairness. The key components of our approach are described below.

3.1 Baseline Model

Our framework is built on a logistic regression model, a model we studied which is widely-used as a baseline for classification tasks where fairness is a concern. For every combination of preprocessing, in-training, and postprocessing methods, a new logistic regression model is initialized and evaluated using both performance and fairness metrics. We chose logistic regression for its simplicity and fast runtime in testing various bias mitigation strategies. However, the modular design of our framework allows the base model to be easily replaced with more complex models that we learned about (e.g., XGBoost) with minimal modifications to code.

3.2 Pipeline Stages and Methods

1. Preprocessing: At the preprocessing stage, we test several methods that modify the input data to reduce bias before model training:

- No Intervention: The raw dataset is used without modification.
- Correlation Removal: We apply a technique that decorrelates non-sensitive features from the sensitive attribute to reduce the influence of the sensitive attribute
- Sensitive Resampling: This method balances the dataset by oversampling or undersampling based on the distribution of the sensitive attribute, ensuring more equal representation of all groups.

2. In-processing: During model training, we test several bias mitigation strategies:

- Baseline Training: A standard model (Logistic Regression) is trained without any bias constraints.
- Reweighting: We set sample weights based on sensitive group frequencies so that the model gives balanced consideration to under or over-represented groups in the loss function.

- Fairness-Constrained Training: We use an iterative approach, known as the Exponentiated Gradient algorithm, that trains a base classifier in each iteration and updates weights on fairness constraints, increasing the emphasis on constraints that are most violated.

3. Postprocessing: After training, we further refine the model's predictions:

- Threshold Optimization: The decision threshold of the logistic regression model is adjusted to meet fairness criteria, either by enforcing similar positive prediction rates across groups (demographic parity) or by balancing error rates (equalized odds).
- No Postprocessing: The model's raw predictions are used as a baseline for comparison.

3.3 Automated Evaluation and Selection

For every combination of preprocessing, in-processing, and postprocessing methods, our system trains a separate model and evaluates it based on its accuracy, F1 score and fairness metrics (Demographic Parity and Equalized Odds). Only the Pareto optimal configurations—those for which no other configuration achieves better performance on all metrics simultaneously—are presented to the user. This means that the selected configurations represent the best trade-offs between predictive performance and fairness. In addition, the user can set thresholds on predictive performance (accuracy and F1 score) and fairness metrics (Demographic Parity and Equalized Odds) to filter and display only those combinations that meet their specific criteria.

4 Experimental Evaluation

To validate our framework, we applied it to four widely used fairness benchmark datasets:

- **Adult Dataset:** A dataset containing census income data, where gender serves as sensitive attributes.
- **Bank Marketing:** A dataset from a Portuguese banking institution, predicting whether a client subscribes to a term deposit, where middle age (25-60) is a protected attibute
- **COMPAS:** A dataset assessing the likelihood of criminal recidivism, where race is a sensitive attribute.
- **German Credit:** A dataset classifying credit risk, where the "personal status and sex" feature serves as the sensitive attribute.

For each dataset, we compared our framework's best model—optimized for the specific fairness metrics — with both a standard baseline model and a naive approach that simply removes the sensitive attribute. The graphs below illustrate the comparison of fairness metrics (Demographic Parity and Equalized Odds) across the baseline, the naive solution, and the best model combination identified by our framework for each of the 4 datasets.

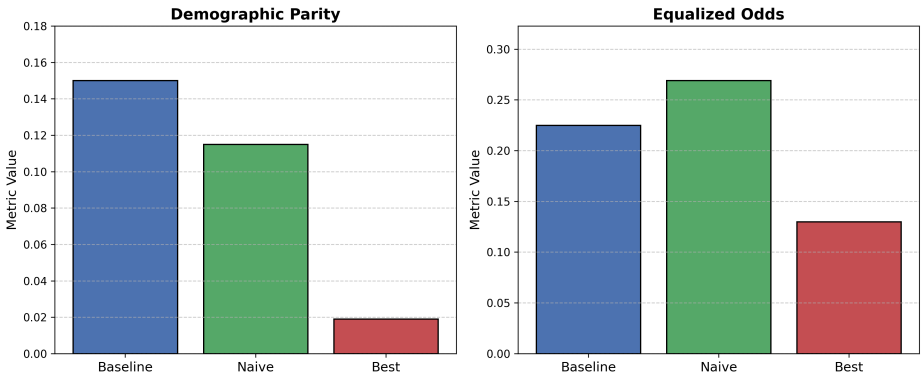

Figure 1: Comparison on the German Credit dataset

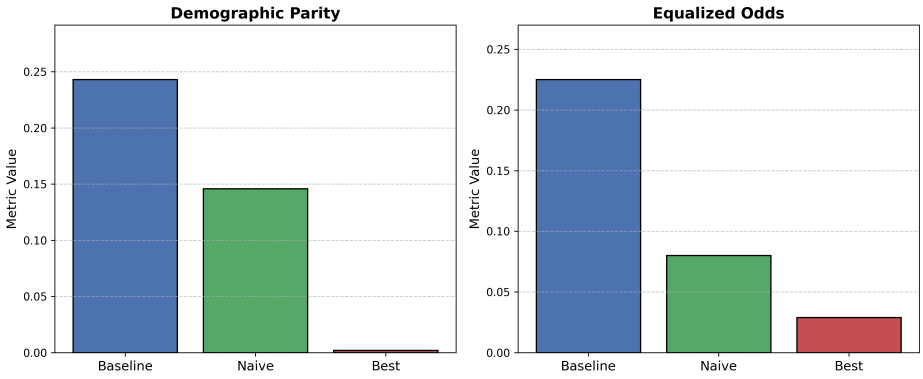

Figure 2: Comparison on the Bank Marketing dataset

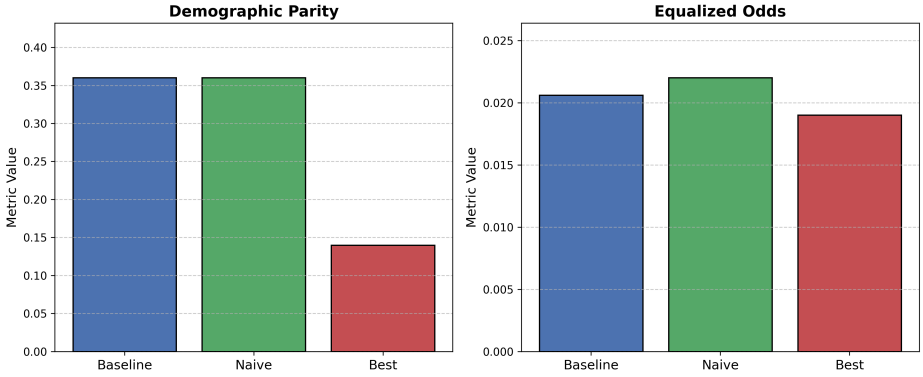

Figure 3: Comparison on the COMPAS dataset

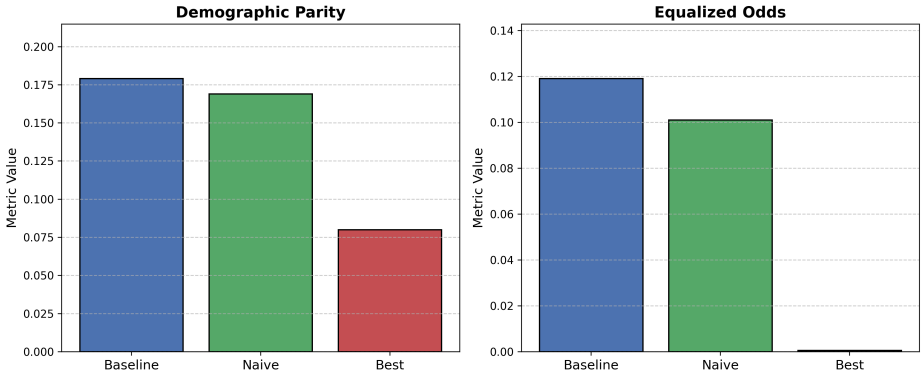

Figure 4: Comparison on the Adult dataset

As illustrated in the graphs, our framework effectively reduced bias across all datasets as measured by both Demographic Parity and Equalized Odds. We enforced minimum performance thresholds by ensuring that accuracy and F1 scores remain at least 90% of their original values. We observed that the best method for reducing Demographic Parity was different from the best method for reducing Equalized Odds, which highlights the challenge of optimizing both fairness metrics at once.

5 Related work

There are numerous existing methods for mitigating bias and improving fairness in machine learning applications. These methods are typically applied at one of three stages of the model pipeline: preprocessing, training, or postprocessing [1]. For instance, preprocessing methods include applying linear transformations to remove correlations with sensitive features [2] and using oversampling techniques to balance classes [3]. During training, fairness metrics can be improved by incorporating fairness regularization terms into the loss function [4] or by employing adaptive sample weighting to better represent under-represented groups [5]. Finally, postprocessing techniques adjust the classifier's decision threshold to achieve more equal outcomes for protected groups [6].

Our project builds on these existing fairness methods by automating the selection of the best combination of methods across the three stages of the model pipeline. The best combinations of methods can be filtered to retain predictive performance.

6 Conclusion

In this work, we presented an automated fairness optimization framework for tabular datasets. Our experiments on several benchmark datasets demonstrated that our approach significantly improves fairness metrics while maintaining competitive predictive performance. Throughout this project, we gained valuable experience in applying machine learning models and ad-

dressing fairness issues, learned how to design and implement experimental evaluations, and improved our ability to effectively report and interpret our findings.

References

[1] Tiago P Pagano, Rafael B Loureiro, Fernanda VN Lisboa, Rodrigo M Peixoto, Guilherme AS Guimarães, Gustavo OR Cruz, Maira M Araujo, Lucas L Santos, Marco AS Cruz, Ewerton LS Oliveira, et al. Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big data and cognitive computing*, 7(1):15, 2023.

[2] Hilde Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. Fairlearn: Assessing and improving fairness of ai systems. *Journal of Machine Learning Research*, 24(257):1–8, 2023.

[3] Md Alamgir Kabir, Mobyen Uddin Ahmed, Shahina Begum, Shaibal Barua, and Md Rakibul Islam. Balancing fairness: unveiling the potential of smote-driven oversampling in ai model enhancement. In *Proceedings of the 2024 9th International Conference on Machine Learning Technologies*, pages 21–29, 2024.

[4] Bhanu Jain, Manfred Huber, and Ramez Elmasri. Increasing fairness in predictions using bias parity score based loss function regularization. *arXiv preprint arXiv:2111.03638*, 2021.

[5] Junyi Chai and Xiaoqian Wang. Fairness with adaptive weights. In *International Conference on Machine Learning*, pages 2853–2866. PMLR, 2022.

[6] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.