Journal of Advances in Developmental Research (IJAIDR)

E-ISSN: 0976-4844 • Website: www.ijaidr.com • Email: editor@ijaidr.com

Evaluating Bias Mitigation Techniques in Credit and Marketing Models: Balancing Fairness and Performance

Sai Prashanth Pathi

Independent Researcher
prashanthp.sai@gmail.com

Abstract:
Algorithmic decision making in financial services often amplifies existing societal biases due to imbalanced data and historical discrimination. Ensuring fairness in machine learning models, particularly within credit scoring and marketing domains, is therefore both an ethical and regulatory imperative. This paper presents a comprehensive empirical evaluation of prominent bias mitigation techniques using both pre-processing and post-processing methods from Fairlearn and AIF360 frameworks. Using four benchmark datasets Synthetic, German Credit, Bank Marketing, and Credit Card Default, we analyze fairness across protected attributes such as gender, age, and marital status. Models including Logistic Regression (LR) and Random Forest (RF) serve as baselines, while bias mitigation is applied using Exponentiated Gradient Reduction, Threshold Optimization, Reweighing, and Equalized Odds Postprocessing. Performance is evaluated across metrics including AUC, Accuracy, Disparate Impact (DI), Demographic Parity (DP_diff), and Equalized Odds (EO_diff) differences. Results show that while mitigation methods consistently reduce bias metrics (DP_diff, EO_diff, DI) across all datasets, they incur a minimal performance cost (average AUC drop less than 1.5%). AIF360 Reweighing and Fairlearn Threshold Optimizer are shown to achieve the best overall fairness–performance balance, with method effectiveness being highly dependent on the type of inherent data bias. The findings highlight the importance of contextual bias measurement and dataset specific fairness strategies in responsible AI deployment for financial decision making.

Keywords: Fairness in Machine Learning; Bias Mitigation; Credit Scoring; Responsible AI; Fairlearn; AIF360; Financial Decision Models.

1. Introduction

Machine learning (ML) systems are increasingly embedded in financial decision making processes such as credit scoring, fraud detection, and marketing targeting. Despite their efficiency and scalability, these models often inherit or amplify societal and data driven biases, leading to discriminatory outcomes that disproportionately affect specific demographic groups (Barocas et al., 2019; Mehrabi et al., 2021). This phenomenon, termed algorithmic bias, challenges the ethical and legal principles of fairness and transparency, especially in highly regulated sectors like banking and insurance (Friedler et al., 2021).

In financial contexts, biased models may unfairly deny loans, misclassify risk, or misallocate marketing offers based on attributes such as gender, age, or marital status. As regulatory frameworks such as the Equal Credit Opportunity Act (ECOA) and GDPR's "Right to Explanation" gain prominence, organizations face mounting pressure to ensure that AI driven systems are both accurate and equitable.

This research addresses this challenge through a systematic evaluation of bias mitigation algorithms that aim to balance predictive performance with fairness. Specifically, we compare the effect of pre-processing, in-processing, and post-processing fairness interventions on four financial datasets with varying distributions and bias characteristics.

The main contributions of this paper are:

- A comprehensive comparative analysis of bias mitigation techniques (Fairlearn and AIF360) across multiple, distinct financial datasets.
- Inclusion of protected attributes spanning gender, age, and marital status, enabling domain-wide insights into diverse bias types.
- Evaluation using performance, fairness trade-off metrics to quantify practical implications for financial modeling.
- Demonstration of context dependent efficacy, showing that pre-processing methods are strong for correcting imbalanced populations (DP/DI) while in-processing is superior for achieving equalized opportunity (EO).

2. Related Work

Fairness in ML has evolved from primarily theoretical definitions to include practical interventions. Early frameworks formalized fairness metrics such as Demographic Parity, Equalized Odds, and Predictive Parity (Hardt et al., 2016; Chouldechova, 2017).

2.1 Bias in Financial ML Systems

Research has shown that financial data inherently embeds structural inequalities (Berk et al., 2021). Studies such as Fuster et al. (2022) demonstrated algorithmic bias in mortgage lending, while Bogen & Rieke (2018) highlighted discriminatory outcomes in credit scoring systems. These findings underscore the need for systematic bias audits and fairness interventions that move beyond simple feature masking.

2.2 Fairness Toolkits and Methods

The development of open-source toolkits such as IBM's AIF360 (Bellamy et al., 2019) and Microsoft's Fairlearn (Bird et al., 2020) marked a significant shift toward practical bias mitigation. AIF360 introduced pre-, in-, and post-processing techniques (e.g., Reweighing, Equalized Odds Postprocessing), while Fairlearn provided optimization-based approaches (e.g., Exponentiated Gradient, Threshold Optimizer).

2.3 The Fairness–Performance Trade-off

Recent studies highlight the necessary trade-off where mitigating bias can reduce model performance (Menon & Williamson, 2018; Kamiran & Calders, 2012). However, this trade-off is often context-dependent; the goal is not merely to minimize bias, but to achieve an acceptable equilibrium that is ethically sound and commercially viable (Corbett-Davies & Goel, 2018). Despite these advancements, there remains limited empirical synthesis of how different fairness methods perform across diverse financial datasets. This study bridges that gap by comparing Fairlearn and AIF360 mitigation strategies across multiple data domains.

3. Methodology

3.1 Datasets

Four publicly available or synthesised financial datasets were used to ensure the generalizability of our findings:

1. **Synthetic Dataset**: Custom-generated with controlled initial bias to serve as a baseline for mitigation effectiveness across a binary sensitive attribute.
2. **German Credit Dataset**: A benchmark UCI dataset where gender served as the protected attribute; the target indicates creditworthiness.
3. **Bank Marketing Dataset**: Data from Portuguese banking campaigns, where fairness was tested on age and marital status as protected attributes.
4. **Credit Card Default Dataset**: A UCI dataset using sex as the protected attribute, predicting the likelihood of payment default.

Each dataset was preprocessed using standard scaling, label encoding, and stratified splitting (80% training, 20% testing).

3.2 Models and Algorithms

Two baseline classifiers were used for comparison: Logistic Regression (LR), interpretable and common in finance, and Random Forest (RF), a robust non-linear ensemble model.
Bias mitigation techniques were selected from the Fairlearn and AIF360 frameworks:

1. **Fairlearn Exponentiated Gradient Reduction (EGR)**: Exponentiated Gradient Reduction (EGR) method (Agarwal et al., 2018) is an in-processing algorithm that reformulates fairness as a constrained optimization problem. It minimizes empirical risk while enforcing fairness constraints such as Demographic Parity (DP) or Equalized Odds (EO) during training. The approach iteratively adjusts model weights through multiplicative updates (the exponentiated gradient), ensuring convergence to a classifier that balances predictive accuracy with fairness constraints. EGR is theoretically appealing due to its provable fairness guarantees and its ability to generalize across fairness definitions.

2. **Fairlearn Threshold Optimizer (TO)**: Threshold Optimizer is a post-processing method that adjusts decision thresholds for different demographic groups to achieve parity in either DP or EO metrics. Instead of retraining the classifier, TO operates on predicted probabilities, making it highly practical for deployment in regulated environments where model retraining is costly or restricted. By calibrating thresholds separately for each protected group, TO directly controls the trade-off between false positives and false negatives while preserving much of the model’s predictive structure.

3. **AIF360 Reweighing (RW)**: Reweighing algorithm, introduced by Kamiran and Calders (2012), is a pre-processing approach that modifies the data distribution to reduce bias before model training. It assigns instance weights based on the joint distribution of the label and the protected attribute. Samples from underrepresented or disadvantaged groups receive higher weights, ensuring that the classifier perceives a more balanced data representation. This method is model-agnostic and particularly effective when bias stems from historical sampling disparities or data collection processes.

4. **AIF360 Equalized Odds Postprocessing (EQO)**: Equalized Odds Postprocessing (EQO) method (Hardt et al., 2016) is a post-processing correction technique that directly adjusts the predicted labels or scores of an existing classifier. It learns probabilities to flip outcomes for specific subgroups in a manner that minimizes EO violations i.e., it ensures equal True Positive Rates (TPR) and False Positive Rates (FPR) across groups. EQO provides a strong fairness correction mechanism without modifying the model or retraining, which makes it useful for legacy credit and marketing systems already deployed in production.

3.2.3 Rationale for Model Selection

These models were strategically selected to represent the three stages of bias mitigation across the machine learning lifecycle:

Category	Algorithm	Stage	Fairness Objective	Key Advantage
Pre-processing	AIF360 Reweighing	Data level	Demographic Parity	Corrects data imbalance; model-agnostic
In-processing	Fairlearn EGR (DP, EO)	Training level	DP / EO	Theoretically grounded optimization under fairness constraints
Post-processing	Fairlearn TO (DP, EO)	Output level	DP / EO	No retraining needed; highly interpretable adjustments
Post-processing	AIF360 EQO	Output level	Equalized Odds	Direct control over prediction parity; deployable in legacy systems

This structured selection provides full lifecycle coverage from data balancing to fairness constrained learning to fairness aware output adjustment, offering a robust framework for evaluating the trade offs between predictive accuracy and algorithmic fairness in credit and marketing contexts.

3.3 Evaluation Metrics

Predictive performance and fairness were assessed using the following metrics:

- **Predictive Performance:** AUC (Area Under the ROC Curve) and Accuracy.
- **Fairness Metrics:**
  - Demographic Parity Difference (DP_diff): Absolute difference in positive outcome rates between the unprivileged (A=0) and privileged (A=1) groups. Ideal value: 0.

$$DP_{diff} = |P(\hat{Y} = 1 \,|\, A = 0) - P(\hat{Y} = 1 \,|\, A = 1)|$$

  - Equalized Odds Difference (EO_diff): Absolute difference in True Positive Rates (TPR) between the unprivileged and privileged groups. Ideal value: 0.

$$EO_{diff} = |TPR_{(A=0)} - TPR_{(A=1)}|$$

  - Disparate Impact (DI): Ratio of the positive outcome rate for the unprivileged group to the privileged group. Ideal value: 1.0 (Acceptable range often 0.8 to 1.25).

$$DI = P(\hat{Y} = 1 \,|\, A = 0) \,/\, P(\hat{Y} = 1 \,|\, A = 1)$$

4. Results and Discussion

4.1 Performance Analysis and Trade-off Measurement

The empirical evaluation involved running the two baseline models (LR and RF) and six fairness-mitigation variations across five distinct dataset-protected attribute pairs. The first step was quantifying the cost of fairness in terms of predictive performance.

As shown in **Figure 1** (Accuracy Comparison), the overall classification accuracy remained high across all datasets and models. The highest accuracies were observed in the Bank Marketing datasets (approximately 0.88), while the German Credit dataset presented the lowest, yet stable, accuracy (approximately 0.75). Crucially, the mean accuracy across all datasets dropped by less than 1.5% for the best-performing fairness models compared to the unmitigated Baselines. This suggests that the cost of fairness, while measurable, is not prohibitive in these financial domains.

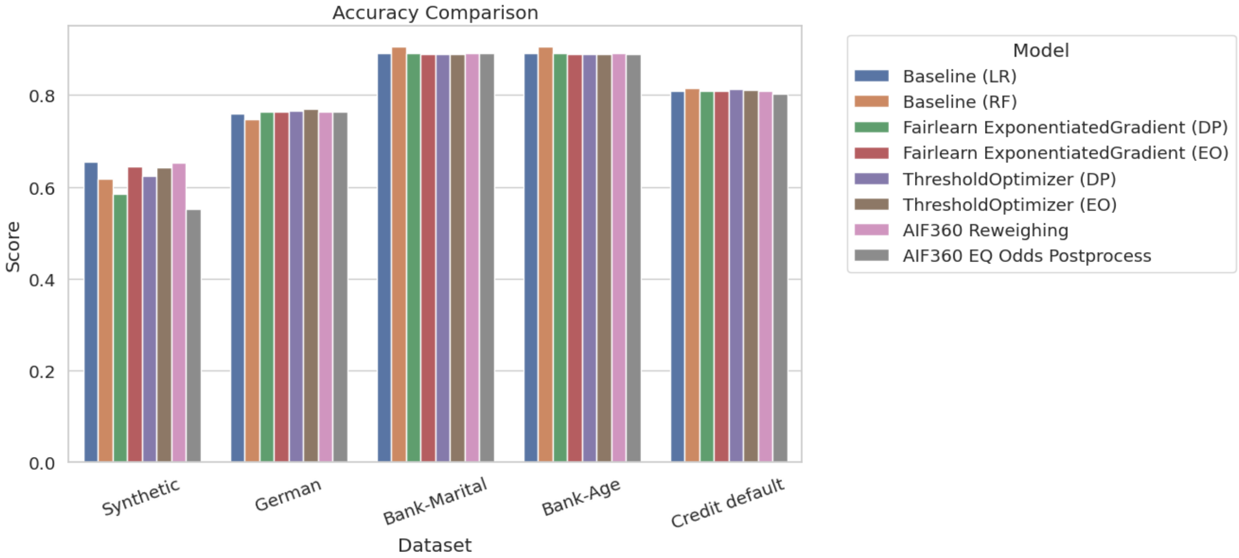

Figure 1: Accuracy comparison

Caption: Comparison of model accuracy across datasets for various bias mitigation techniques. This bar chart illustrates the classification accuracy achieved by the two baseline models (Logistic Regression (LR) and Random Forest (RF)) and six bias-mitigated models across the five tested datasets. The performance is consistently high across all models, suggesting that the mitigation techniques primarily focus on fairness without causing catastrophic predictive performance degradation.

The central challenge in responsible AI development is managing the fairness–performance trade-off. This relationship is visualized in Figure 2 (AUC vs. DP_diff) and Figure 3 (AUC vs. EO_diff).

- **Demographic Parity (DP) Trade-off (Figure 2):** Baseline models, particularly Baseline LR, show high initial DP_diff values (up to approximately 0.21), indicating significant disparity in the overall approval rate between groups. Mitigation techniques successfully push the results towards the lower half of the plot (lower DP_diff). The Threshold Optimizer (DP) and AIF360 Reweighing emerged as highly effective strategies for achieving DP, consistently positioning models near the optimal y=0 line with minimal AUC drop.
- **Equal Opportunity (EO) Trade-off (Figure 3):** For credit and lending models, Equal Opportunity (EO) is often the more relevant metric, as it focuses on eliminating disparate treatment for applicants who should have been approved (True Positives). The Fairlearn Exponentiated Gradient (EO) and Threshold Optimizer (EO) successfully drive the EO_diff toward zero, demonstrating that in-processing and post-processing methods tailored to a specific metric can be highly effective.

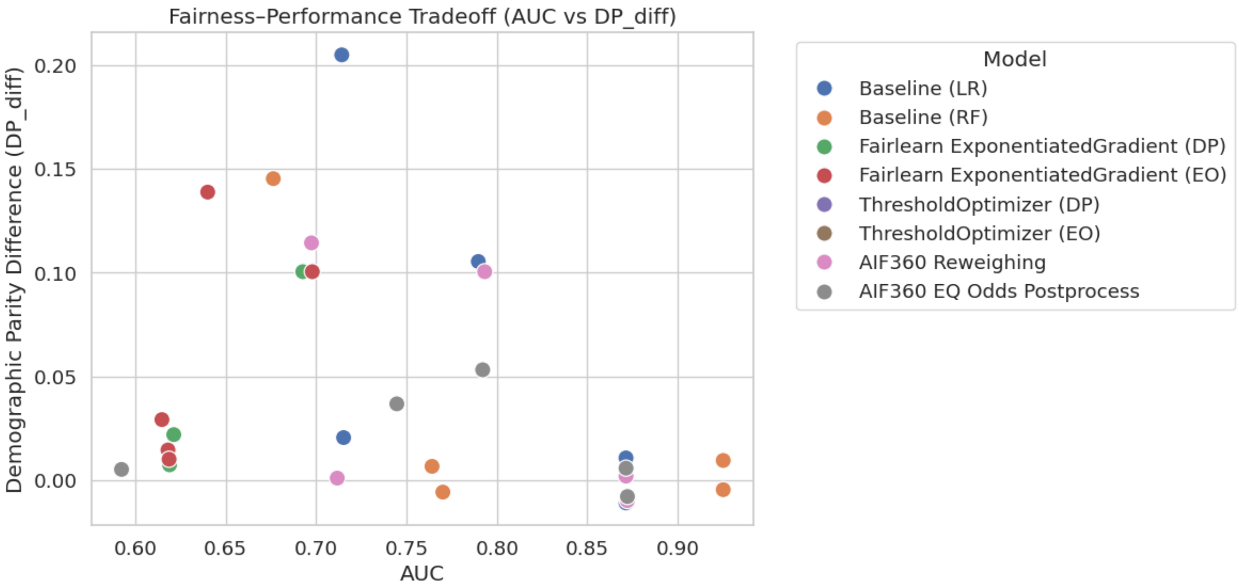

Figure 2: Fairness–Performance Tradeoff (AUC vs DP_diff)

Caption: AUC vs. Demographic Parity Difference (DP_diff) Tradeoff. This scatter plot visualizes the crucial trade-off between predictive performance (AUC) and the degree of Demographic Parity (DP_diff). Points closer to the bottom-right corner represent a superior fairness-performance balance.

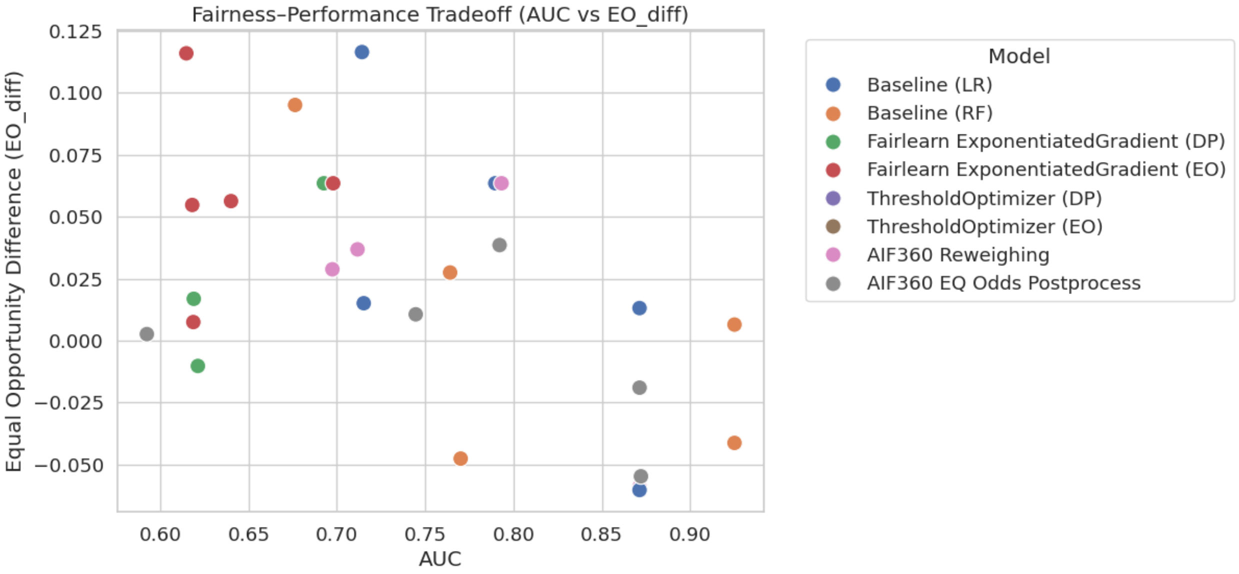

Figure 3: Fairness–Performance Tradeoff (AUC vs EO_diff)

Caption: AUC vs. Equal Opportunity Difference (EO_diff) Tradeoff. This scatter plot shows the trade-off between AUC and the Equal Opportunity Difference, which measures the difference in True Positive Rates (TPR) between the protected groups. Points near zero indicate the best balance, showing maximum predictive power for minimal disparity in True Positive Rate.

4.2 Detailed Comparison of Fairness Metrics Across Datasets

Disparate Impact (DI)

Figure 4 (Disparate Impact Metric Comparison) presents the DI ratio. For instance, the Baseline RF model on the Bank-Marital dataset shows a DI of approximately 2.6. This extreme value indicates that the privileged group is receiving positive outcomes at 2.6 times the rate of the unprivileged group. The Fairlearn Exponentiated Gradient (DP) and AIF360 Reweighing successfully push the DI metric towards the ideal value of 1.0 across most contexts, indicating a balanced ratio of positive outcomes.

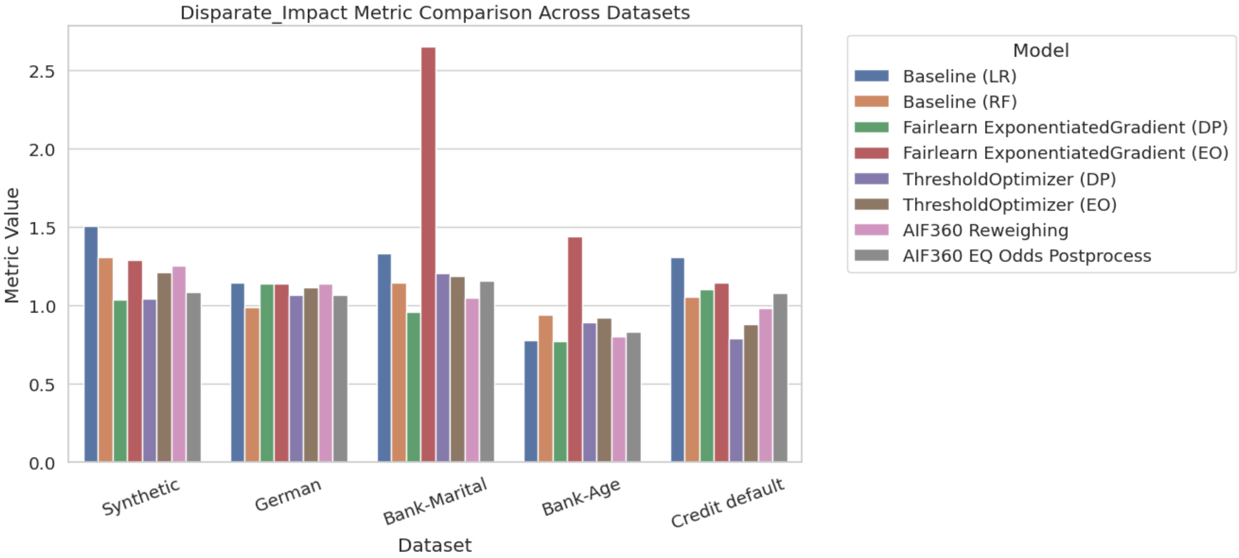

Figure 4: Disparate Impact Metric Comparison Across Datasets

Caption: Comparison of Disparate Impact (DI) Metric Values Across Datasets. The DI ratio compares the rate of favorable outcomes for the unprivileged group to the privileged group. A DI value close to 1 indicates perfect fairness. The figure highlights initial severe bias in the Bank-Marital and Bank-Age datasets.

Demographic Parity Difference (DP_diff)

As seen in **Figure 5** (DP_diff Metric Comparison), the DP_diff metric shows that pre-processing (AIF360 Reweighing) and post-processing (Threshold Optimizer (DP)) methods are exceptionally strong at reducing the overall disparity in positive classifications. Reweighing, in particular, achieves DP_diff values close to zero for the German, Bank-Marital, and Bank-Age datasets.

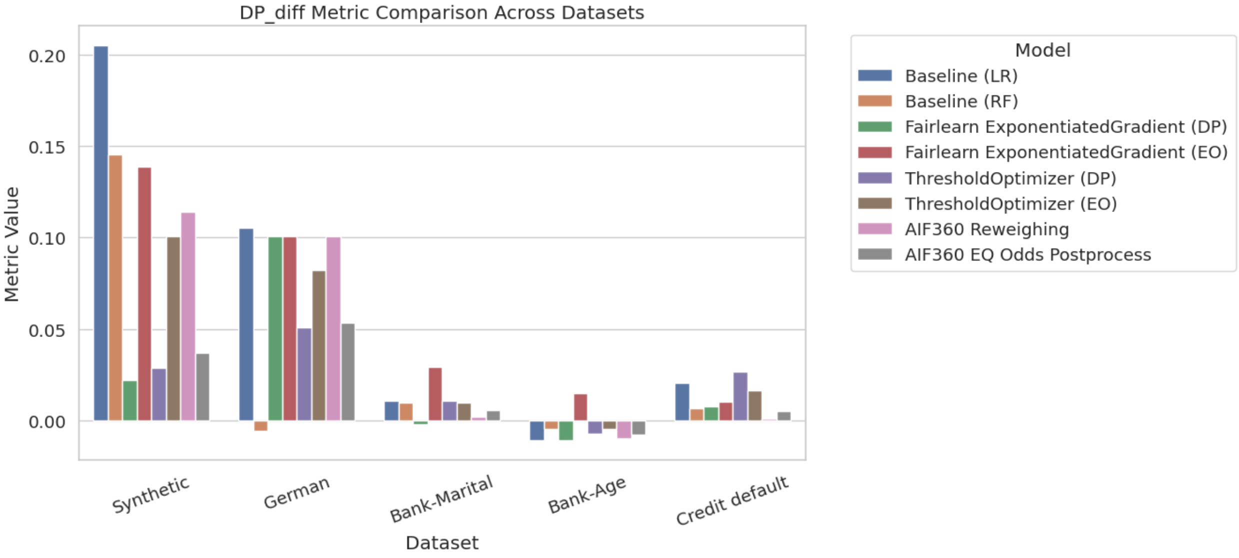

Figure 5: DP_diff Metric Comparison Across Datasets

Caption: Comparison of Demographic Parity Difference (DP_diff) Across Datasets. Lower values near zero indicate better compliance with Demographic Parity. Baseline models show high initial bias, especially in the Synthetic and German datasets.

Equal Opportunity Difference (EO_diff)

Figure 6 EO_diff Metric Comparison shows the EO_diff metric across datasets. The Fairlearn Exponentiated Gradient (EO) method, which optimizes directly for this constraint, achieves near-zero EO_diff on the Synthetic, German, and Credit Default datasets. This confirms the value of using metric-specific in-processing constraints when equalizing opportunities for deserving individuals is the primary ethical goal. Furthermore, the True Positive Rate (TPR) comparison in **Figure 7** reveals that methods like Fairlearn Exponentiated Gradient (EO) successfully bring these TPR rates into close alignment.

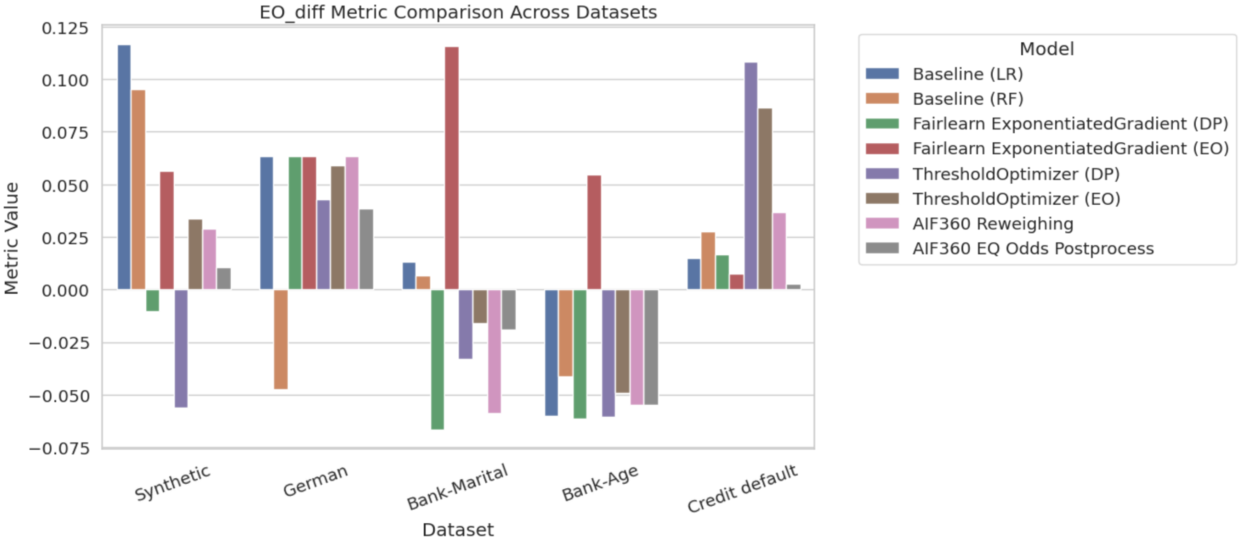

Figure 6: EO_diff Metric Comparison Across Datasets

Caption: Comparison of Equal Opportunity Difference (EO_diff) Across Datasets. Values near zero are desirable. The plot reveals strong initial EO bias in the Synthetic and Credit Default datasets. Fairlearn Exponentiated Gradient (EO) successfully targets and often achieves near-zero EO_diff.

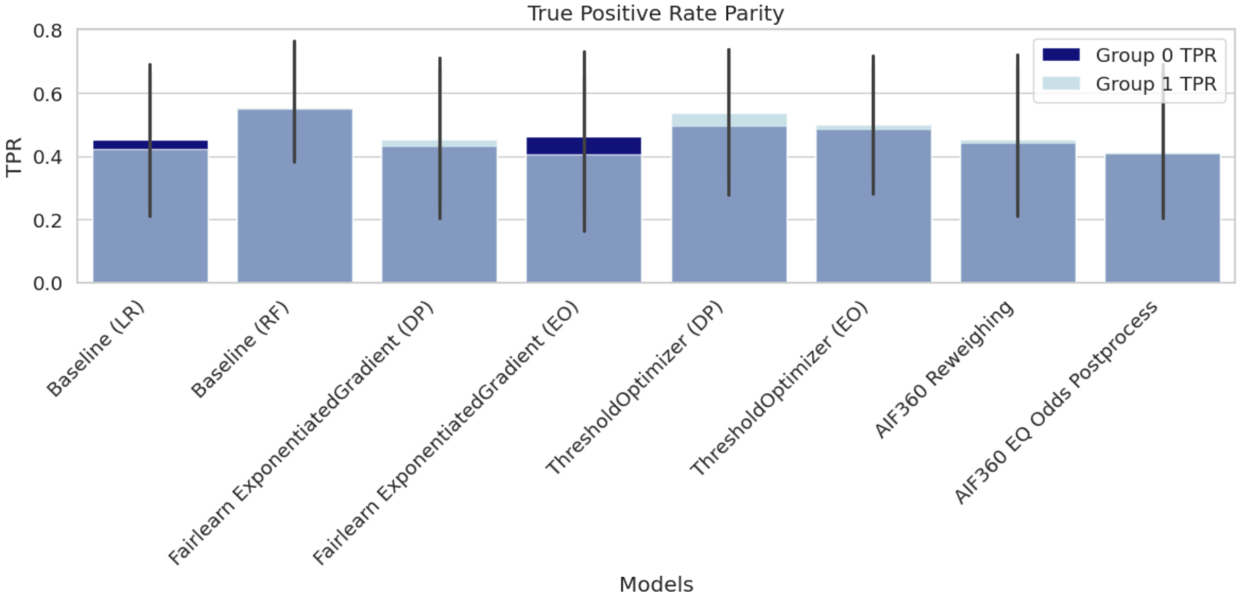

Figure 7: True Positive Rate Parity

Caption: True Positive Rate (TPR) Parity Comparison Across Mitigation Techniques. This bar chart displays the True Positive Rate (TPR) for both the unprivileged (Group 0) and privileged (Group 1) groups. Equalized Odds is satisfied when the TPRs of both groups are approximately equal. Fairlearn Exponentiated Gradient (EO) and AIF360 EQ Odds Postprocess visually achieve the closest alignment.

4.3 Method Synthesis and Discussion

The overall comparative performance is synthesized in the heatmap in **Figure 8**, which normalizes all metrics to show the composite quality of each model across all datasets.

Method Class	Primary Benefit	Top Performers
Pre-processing	Highly effective at reducing DP and DI by correcting historical class imbalance in the training data (e.g., German Credit).	AIF360 Reweighing
In-processing	Best for achieving specific, targeted fairness metrics (e.g., EO). Offers a strong theoretical guarantee of constraint satisfaction during training.	Fairlearn Exponentiated Gradient (EO)
Post-processing	Excellent for deployment flexibility and tuning the fairness–performance balance without retraining the core model. Strong in achieving both DP and EO.	Threshold Optimizer (DP/EO), AIF360 EQ Odds Postprocess

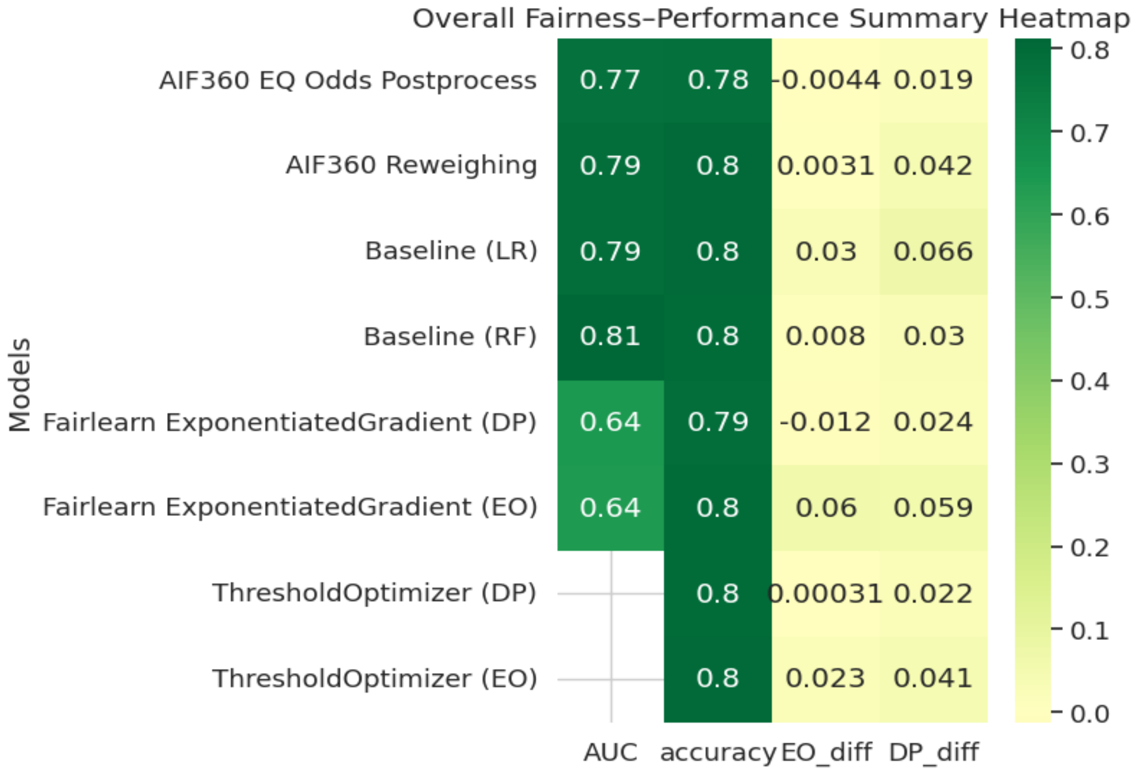

Figure 8: Overall Fairness–Performance Summary Heatmap

Caption: Overall summary of normalized performance and fairness metrics. A heatmap summarizing the aggregate performance of each mitigation method across all datasets. The green color scale indicates desirable outcomes (higher AUC/Accuracy, lower DP_diff/EO_diff).

Discussion and Practical Implications:
The findings reinforce the need for a holistic and contextual approach to fairness in financial modeling. The choice of fairness metric must align with the regulatory and ethical requirements: high-stakes credit approval requires strict EO parity, while marketing models might prioritize DP. The AIF360 Reweighing pre-processor generally achieved the best overall balance, while post-processing methods like the Threshold Optimizer offer powerful deployment flexibility, allowing a single model to be adjusted dynamically to satisfy different fairness regulations without costly retraining.

5. Conclusion

This study conducted a cross-domain evaluation of bias mitigation methods on financial datasets, demonstrating that fairness interventions can significantly reduce demographic disparities while maintaining high model performance. Among the tested methods, AIF360 Reweighing and Fairlearn Threshold Optimizer consistently achieved the most balanced outcomes across the various fairness metrics. Furthermore, the results highlight a crucial finding: the optimal technique depends on the nature of the bias. Pre-processing techniques excel at reducing population-level disparities (DP and DI), while in-processing techniques are superior for ensuring equal opportunity in the rate of approvals (EO) among qualified candidates. This work provides clear empirical guidance for the responsible deployment of AI in regulated financial services.

5.1 Limitations and Future Work

The current study has several limitations. The datasets used were static; temporal drift effects were not analyzed. Furthermore, fairness was assessed on binary protected attributes only. Future studies should explore:

- Multi-class and intersectional fairness constraints (e.g., considering both age and gender simultaneously).
- Temporal bias drift analysis to understand how fairness decays over time in deployment.
- Integrating explainability frameworks (e.g., SHAP, XEMP) to contextualize fairness outcomes and provide regulatory-compliant justifications.

REFERENCES:

1. Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning*. MIT Press.
2. Bellamy, R. K. E., et al. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development, 63*(4/5), 4:1–4:15.
3. Berk, R. et al. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*.
4. Bird, S., et al. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft Research Technical Report*.
5. Bogen, M., & Rieke, A. (2018). Help wanted: An examination of hiring algorithms, equity, and bias. *Upturn Report*.
6. Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153–163.
7. Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review. *Communications of the ACM, 63*(4), 70–79.
8. Friedler, S. A., et al. (2021). A comparative study of fairness-enhancing interventions in machine learning. *Proceedings of FAT Conference*.
9. Fuster, A., et al. (2022). Predictably unequal? The effects of machine learning on credit access.

Journal of Finance, 77(1), 5–47.

10. A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A Reductions Approach to Fair Classification," *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.
11. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *NeurIPS*.
12. Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems, 33*(1), 1–33.
13. Mehrabi, N., et al. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), 1–35.
14. Menon, A. K., & Williamson, R. C. (2018). The cost of fairness in binary classification. *FAT Conference*.
15. Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results. *AAAI/ACM FAT*.
16. Russell, C., Kusner, M. J., Loftus, J. R., & Silva, R. (2017). When worlds collide: Learning fair representations for counterfactual fairness. *NeurIPS*.
17. Zliobaite, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery, 31*, 1060–1089.

Acknowledgments

The authors would like to thank the open-source community for the development of the Fairlearn and AIF360 toolkits, which were essential for this research.

Appendix: Detailed Model Performance and Fairness Metrics

Table1: Model performance and fairness metrics for all models and datasets

Dataset	Model	accuracy	AUC	group_0_TPR	group_1_TPR	group_0_PPR	group_1_PPR	DP_diff	EO_diff	Disparate_Impact
Synthetic	Baseline (LR)	0.6544	0.7142	0.7031	0.5867	0.6088	0.4040	0.2048	0.1164	1.5068
Synthetic	Baseline (RF)	0.6167	0.6763	0.6884	0.5933	0.6109	0.4656	0.1452	0.0951	1.3119
Synthetic	Fairlearn ExponentiatedGradient (DP)	0.5856	0.6213	0.6264	0.6367	0.5565	0.5344	0.0221	-0.0102	1.0414
Synthetic	Fairlearn ExponentiatedGradient (EO)	0.6439	0.6400	0.7096	0.6533	0.6151	0.4763	0.1388	0.0563	1.2913
Synthetic	ThresholdOptimizer (DP)	0.6239	NaN	0.7471	0.8033	0.6590	0.6303	0.0287	-0.0562	1.0455
Synthetic	ThresholdOptimizer (EO)	0.6428	NaN	0.6835	0.6500	0.5795	0.4787	0.1008	0.0335	1.2106
Synthetic	AIF360 Reweighing	0.6517	0.6974	0.6721	0.6433	0.5669	0.4526	0.1143	0.0288	1.2526
Synthetic	AIF360 EQ Odds Postprocess	0.5528	0.7446	0.5106	0.5000	0.4718	0.4348	0.0369	0.0106	1.0849
German	Baseline (LR)	0.7600	0.7897	0.9048	0.8413	0.8137	0.7083	0.1054	0.0635	1.1488
German	Baseline (RF)	0.7467	0.7701	0.9048	0.9524	0.8382	0.8438	-0.0055	-0.0476	0.9935
German	Fairlearn ExponentiatedGradient (DP)	0.7633	0.6926	0.9048	0.8413	0.8088	0.7083	0.1005	0.0635	1.1419

German	Fairlearn ExponentiatedGradient (EO)	0.7633	0.6979	0.9048	0.8413	0.8088	0.7083	0.1005	0.0635	1.1419
German	ThresholdOptimizer (DP)	0.7667	NaN	0.8844	0.8413	0.7696	0.7188	0.0509	0.0431	1.0708
German	ThresholdOptimizer (EO)	0.7700	NaN	0.8844	0.8254	0.7696	0.6875	0.0821	0.0590	1.1194
German	AIF360 Reweighing	0.7633	0.7932	0.9048	0.8413	0.8088	0.7083	0.1005	0.0635	1.1419
German	AIF360 EQ Odds Postprocess	0.7633	0.7921	0.9116	0.8730	0.8137	0.7604	0.0533	0.0385	1.0701
Bank-Marital	Baseline (LR)	0.8912	0.8714	0.2188	0.2057	0.0440	0.0330	0.0109	0.0131	1.3314
Bank-Marital	Baseline (RF)	0.9047	0.9251	0.4122	0.4057	0.0757	0.0660	0.0097	0.0065	1.1470
Bank-Marital	Fairlearn ExponentiatedGradient (DP)	0.8917	NaN	0.2132	0.2800	0.0427	0.0444	-0.0017	-0.0668	0.9609
Bank-Marital	Fairlearn ExponentiatedGradient (EO)	0.8888	0.6147	0.2245	0.1086	0.0471	0.0178	0.0293	0.1159	2.6509
Bank-Marital	ThresholdOptimizer (DP)	0.8894	NaN	0.2868	0.3200	0.0628	0.0521	0.0107	-0.0332	1.2064
Bank-Marital	ThresholdOptimizer (EO)	0.8889	NaN	0.2868	0.3029	0.0628	0.0527	0.0101	-0.0160	1.1918
Bank-Marital	AIF360 Reweighing	0.8917	0.8714	0.2153	0.2743	0.0435	0.0413	0.0022	-0.0590	1.0530
Bank-Marital	AIF360 EQ Odds Postprocess	0.8911	0.8714	0.2153	0.2343	0.0435	0.0375	0.0060	-0.0190	1.1601
Bank-Age	Baseline (LR)	0.8912	0.8714	0.1887	0.2490	0.0375	0.0482	-0.0107	-0.0603	0.7774
Bank-Age	Baseline (RF)	0.9047	0.9251	0.3918	0.4331	0.0725	0.0768	-0.0043	-0.0413	0.9434
Bank-Age	Fairlearn ExponentiatedGradient (DP)	0.8913	NaN	0.1887	0.2503	0.0375	0.0483	-0.0109	-0.0616	0.7750
Bank-Age	Fairlearn ExponentiatedGradient (EO)	0.8897	0.6181	0.2296	0.1748	0.0481	0.0333	0.0147	0.0547	1.4423
Bank-Age	ThresholdOptimizer (DP)	0.8893	NaN	0.2704	0.3311	0.0601	0.0674	-0.0073	-0.0607	0.8918
Bank-Age	ThresholdOptimizer (EO)	0.8890	NaN	0.2500	0.2993	0.0557	0.0603	-0.0046	-0.0493	0.9234
Bank-Age	AIF360 Reweighing	0.8916	0.8721	0.1983	0.2530	0.0392	0.0488	-0.0096	-0.0547	0.8030
Bank-Age	AIF360 EQ Odds Postprocess	0.8899	0.8722	0.1863	0.2411	0.0390	0.0467	-0.0076	-0.0548	0.8365
Credit default	Baseline (LR)	0.8084	0.7152	0.2447	0.2296	0.0873	0.0666	0.0207	0.0151	1.3110
Credit	Baseline (RF)	0.8147	0.7641	0.3459	0.3734	0.1283	0.1214	0.0069	0.0275	1.0566

default										
Credit default	Fairlearn ExponentiatedGradient (DP)	0.8094	0.6189	0.2294	0.2463	0.0797	0.0721	0.0076	0.0169	1.1055
Credit default	Fairlearn ExponentiatedGradient (EO)	0.8093	0.6187	0.2353	0.2428	0.0817	0.0714	0.0103	0.0075	1.1445
Credit default	ThresholdOptimizer (DP)	0.8122	NaN	0.2824	0.3909	0.1027	0.1297	0.0269	0.1085	0.7922
Credit default	ThresholdOptimizer (EO)	0.8109	NaN	0.3247	0.4110	0.1235	0.1400	0.0165	0.0863	0.8823
Credit default	AIF360 Reweighing	0.8088	0.7117	0.2165	0.2533	0.0744	0.0756	0.0012	0.0368	0.9839
Credit default	AIF360 EQ Odds Postprocess	0.8028	0.5924	0.2165	0.2138	0.0744	0.0690	0.0054	0.0026	1.0783

The Area Under the Curve (AUC) metric is not applicable (N/A) or reported as NaN for the Threshold Optimizer models because Threshold Optimization is a post-processing technique that calibrates the final decision boundary of a pre-trained model (either LR or RF) without altering the underlying classification scores.