CIA Open Access

Fairness-Accuracy Trade-offs in AI Credit Scoring: A Comparative Evaluation of Reweighting and Resampling Strategies Under Multiple Fairness Constraints

Ziyi Wang[1], Jiawen Lai[1.2]

[1] *Enterprise Risk Management, Columbia University, NY, USA*
[1.2] *Computer Engineering, University of California, Riverside, CA, USA*

DOI: 10.63575/CIA.2026.40110

Abstract

The proliferation of artificial intelligence in financial risk assessment has introduced significant concerns regarding algorithmic fairness, particularly in credit scoring systems where biased predictions can disproportionately affect protected demographic groups. This study presents a comparative evaluation of two predominant pre-processing debiasing strategies—reweighting and resampling—applied to AI-based credit scoring algorithms. Using two publicly available benchmark datasets (the UCI German Credit dataset and the UCI Default of Credit Card Clients dataset), we systematically assess the accuracy-fairness trade-offs under three widely adopted fairness constraints: statistical parity, equal opportunity, and equalized odds. Experimental results across three baseline classifiers (Logistic Regression, Random Forest, and XGBoost) indicate that reweighting achieves a more favorable balance between predictive accuracy and fairness when evaluated under equal opportunity and equalized odds constraints, while resampling demonstrates stronger performance in reducing statistical parity differences. The magnitude of accuracy degradation varies substantially depending on the choice of fairness constraint, with equalized odds imposing the greatest accuracy cost across both datasets. These findings provide evidence-based guidance for financial institutions seeking to implement fairness-aware credit scoring systems and suggest that the selection of debiasing strategy should be contingent upon the specific fairness objective prioritized by regulatory and institutional requirements.

Keywords: *algorithmic fairness; credit scoring; pre-processing debiasing; fairness-accuracy trade-off*

1. Introduction

1.1. Background and Motivation

The increasing reliance on machine learning algorithms for credit risk assessment has transformed lending practices across the financial industry. Automated credit scoring systems process vast volumes of applicant data, enabling financial institutions to make rapid and scalable lending decisions. The 2023 Home Mortgage Disclosure Act (HMDA) data, published by the Federal Financial Institutions Examination Council (FFIEC) in July 2024, reveals persistent disparities in mortgage lending outcomes: Black or African American applicants faced a denial rate of 16.6%, compared to 5.8% for non-Hispanic White applicants for conventional home purchase loans (Source: FFIEC, Summary of 2023 HMDA Data). Such disparities raise fundamental questions about whether algorithmic decision-making perpetuates or amplifies historical patterns of discrimination.

Algorithmic bias in financial services constitutes a form of systematic disadvantage that can emerge from multiple sources within the machine learning pipeline. Barocas and Selbst [1] demonstrated that data-driven decision-making systems can produce discriminatory outcomes even in the absence of explicit discriminatory intent, as training data often reflects historical inequities embedded in institutional lending practices. The tension between predictive accuracy and algorithmic fairness represents a central challenge in responsible AI deployment. Hardt et al. [2] formalized the concepts of equalized odds and equal opportunity as fairness criteria for supervised learning, establishing a theoretical foundation for measuring discrimination in classification tasks. Kleinberg et al. [3] subsequently proved that certain widely-used fairness definitions are mathematically incompatible under conditions where base rates differ across groups, indicating that practitioners must make deliberate choices about which fairness criteria to prioritize. Chouldechova [4] reinforced this finding by demonstrating that predictive parity and equal error rates cannot be simultaneously achieved when prevalence rates differ between protected groups.

1.2. Research Objectives and Contributions

A. Research Objectives

This study addresses the gap in systematic comparative evaluations of pre-processing debiasing techniques within the specific context of financial credit scoring. Three research objectives guide this investigation: (i) to systematically compare reweighting and resampling, two mainstream pre-processing debiasing strategies, across multiple publicly available credit scoring datasets; (ii) to quantify the accuracy-fairness trade-offs under three distinct fairness constraints—statistical parity, equal opportunity, and equalized odds—providing a multi-dimensional assessment rather than relying on a single fairness metric; and (iii) to derive evidence-based recommendations regarding the suitability of each strategy for financial lending scenarios. Buolamwini and Gebru [5] highlighted that intersectional disparities can remain hidden when evaluation relies on aggregate metrics alone, motivating our multi-constraint analytical approach.

B. Contributions

The principal contributions of this paper are threefold. The study provides the first cross-dataset, cross-constraint comparative assessment of reweighting versus resampling debiasing techniques using standardized evaluation protocols on financial credit data. It quantifies the differential impact of three fairness constraints on predictive performance, revealing that the choice of constraint substantially alters the observed accuracy-fairness trade-off profile. It offers practical guidance aligned with financial regulatory expectations, including the Equal Credit Opportunity Act (ECOA) and the EU AI Act's requirements for high-risk AI systems, supporting institutions in selecting context-appropriate debiasing strategies.

2. Literature Review

2.1. Fairness Metrics in Credit Scoring

A. Formal Definitions of Group Fairness

Group fairness metrics evaluate whether a classifier's predictions exhibit systematic disparities across protected demographic groups defined by sensitive attributes such as gender, race, or age. Three predominant definitions have emerged in the algorithmic fairness literature. Statistical parity (also termed demographic parity) requires that the probability of receiving a positive prediction is equal across groups: $P(\hat{Y}=1|A=a) = P(\hat{Y}=1|A=b)$, where A denotes the protected attribute. Dwork et al. [6] introduced the framework of individual fairness, requiring that similar individuals receive similar outcomes, and distinguished it from group-level statistical criteria. Equal opportunity constrains the true positive rate to be identical across groups: $P(\hat{Y}=1|Y=1,A=a) = P(\hat{Y}=1|Y=1,A=b)$. Equalized odds extends this requirement to both true positive and false positive rates simultaneously.

B. Applicability in Financial Contexts

The selection of an appropriate fairness metric carries substantial implications for credit lending outcomes. Statistical parity, while conceptually straightforward, does not account for legitimate differences in creditworthiness between groups and may force approval rates to converge regardless of underlying risk profiles. Equal opportunity focuses specifically on qualified applicants—those who would repay their loans—and ensures they are not unfairly denied credit based on group membership. Equalized odds provides the most stringent constraint by requiring parity in both approval of creditworthy applicants and rejection of non-creditworthy applicants. Kamiran and Calders [7] provided early empirical evidence that classifiers trained on historical lending data exhibit measurable discrimination, establishing the practical relevance of these metrics in credit contexts.

2.2. Bias Sources in AI Credit Scoring Algorithms

Bias in credit scoring algorithms originates from three primary channels. Historical data bias arises when training data encodes past discriminatory lending decisions. Proxy variable bias occurs when ostensibly neutral features—such as geographic location or employment sector—serve as statistical proxies for protected attributes. Sample selection bias results from training exclusively on previously approved applicants, creating a non-representative sample. Calmon et al. [8] demonstrated that probabilistic data transformations can simultaneously control discrimination and preserve predictive utility through a convex optimization framework. Feldman et al. [9] proposed a method for certifying and removing disparate impact by modifying feature distributions to ensure classifiers cannot distinguish protected groups.

2.3. Pre-processing Debiasing Techniques

Pre-processing approaches modify the training data prior to model fitting, offering the advantage of being model-agnostic and compatible with any downstream classifier. Reweighting assigns differentiated sample weights based on the joint distribution of the protected attribute and target variable, neutralizing the statistical association between group membership and class labels. Resampling strategies adjust the composition of the training set through oversampling underrepresented group-label combinations or undersampling

overrepresented ones. Zemel et al. [10] proposed learning intermediate fair representations that encode data utility while obfuscating protected attribute information. Zafar et al. [11] introduced fairness constraints integrated into the classifier's optimization objective. The relative effectiveness of reweighting versus resampling under different fairness constraints in credit scoring remains insufficiently characterized, motivating the present comparative study.

3. Research Methodology

3.1. Datasets and Experimental Setup

Two publicly available benchmark datasets serve as the empirical foundation for this study. The Statlog German Credit dataset, hosted by the UCI Machine Learning Repository (Hofmann, 1994; DOI: 10.24432/C5NC77), comprises 1,000 loan application records with 20 features encompassing financial attributes, personal demographics, and credit history. The binary target classifies applicants as good (700, 70.0%) or bad (300, 30.0%) credit risks. The protected attribute is gender (male: 690, 69.0%; female: 310, 31.0%). The Default of Credit Card Clients dataset (Yeh and Lien, 2009), also from the UCI Repository, contains 30,000 credit card client records from Taiwan with 24 features. The target variable indicates whether the client defaulted (default: 6,636, 22.1%; non-default: 23,364, 77.9%). The protected attribute is sex (male: 11,888, 39.6%; female: 18,112, 60.4%). Table 1 summarizes the key characteristics of both datasets.

Table 1. Summary of Dataset Characteristics

Dataset	Samples	Features	Positive %	Protected Attr.	Priv. Group %	Unpriv. Group %
German Credit	1,000	20	70.0%	Gender	69.0% (M)	31.0% (F)
Taiwan Credit	30,000	24	77.9%	Sex	39.6% (M)	60.4% (F)

Note: Positive % refers to the favorable outcome class (good credit / non-default). Priv. = Privileged; Unpriv. = Unprivileged. Data sources: UCI Machine Learning Repository (Hofmann, 1994; Yeh & Lien, 2009).

Categorical features are encoded using one-hot encoding, and numerical features are normalized using min-max scaling to [0, 1]. Missing values are imputed using the median strategy. The data is partitioned into 80% training and 20% test sets using stratified sampling. All reported metrics represent means and standard deviations over five-fold stratified cross-validation. Agarwal et al. [12] established that reduction-based approaches to fair classification require careful experimental design with proper cross-validation, a protocol we adopt throughout.

3.2. Debiasing Techniques Under Comparison

A. Reweighting Strategy

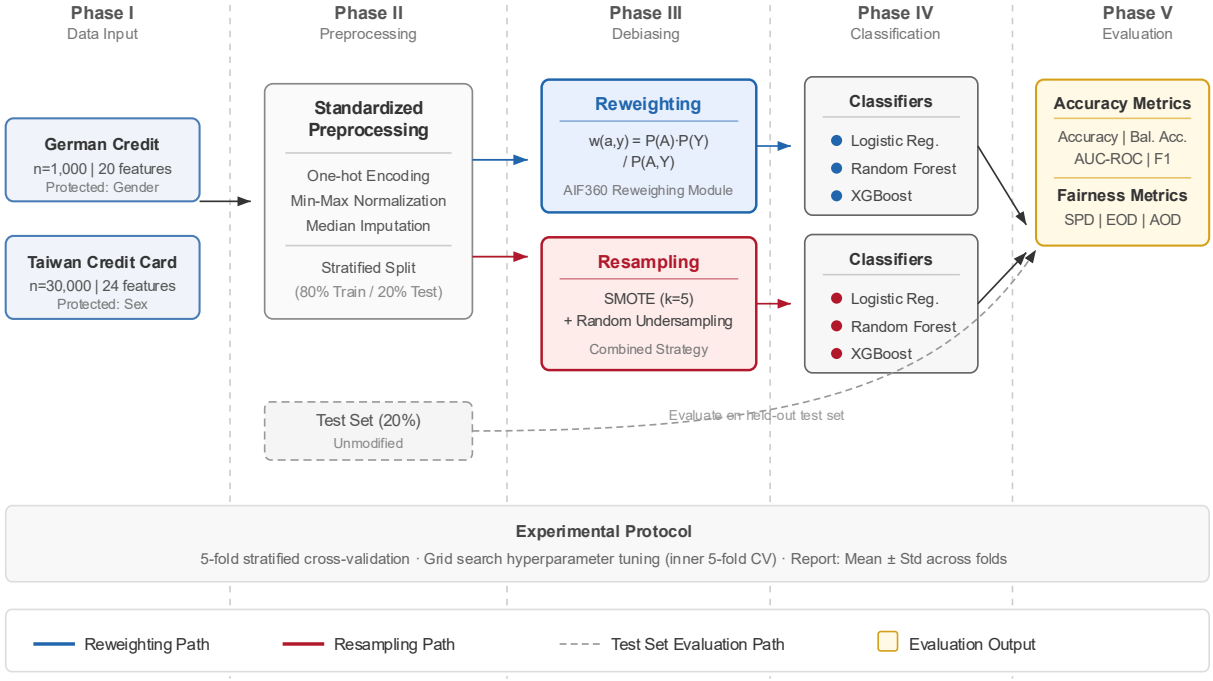
The reweighting approach implements the sample weight assignment method proposed in the data preprocessing framework for classification without discrimination. For each combination of protected attribute value $a \in \{\text{privileged}, \text{unprivileged}\}$ and class label $y \in \{\text{positive}, \text{negative}\}$, the sample weight is computed as: $w(a,y) = P(A=a) \times P(Y=y) / P(A=a, Y=y)$. This formulation assigns higher weights to underrepresented group-label combinations (e.g., unprivileged group with positive labels) and lower weights to overrepresented ones, effectively removing the statistical dependence between the protected attribute and the class label in the weighted training distribution. The computed weights are incorporated into the loss function of each baseline classifier during training, requiring no modification to feature values or class labels. The implementation utilizes the Reweighing module from the IBM AI Fairness 360 (AIF360) toolkit [13], ensuring reproducibility and standardization.

B. Resampling Strategy

The resampling strategy combines targeted oversampling and undersampling to equalize the representation of group-label combinations in the training set. Specifically, the Synthetic Minority Over-sampling Technique (SMOTE) is applied to generate synthetic instances for the underrepresented group-label pair (unprivileged group with positive labels), while random undersampling reduces the count of the overrepresented pair (privileged group with positive labels). The combined strategy ensures that the resulting training distribution satisfies approximate independence between the protected attribute and the class label. The SMOTE implementation uses k=5 nearest neighbors for synthetic sample generation. A critical distinction from the reweighting approach is that resampling physically alters the training set composition, producing a new dataset

with modified sample counts rather than adjusted loss contributions. This structural difference can influence model behavior, particularly for algorithms sensitive to training set size and class distribution.

Figure 1. Experimental Framework for Comparative Evaluation of Pre-processing Debiasing Strategies



This diagram illustrates the end-to-end experimental pipeline. Raw credit data from both datasets undergoes standardized preprocessing (encoding, normalization, stratified splitting). The training set is then processed through two parallel debiasing paths: reweighting (sample weight computation) and resampling (SMOTE oversampling + random undersampling). Each debiased training set is used to train three baseline classifiers (LR, RF, XGBoost). All trained models are evaluated on the unmodified test set using four accuracy metrics and three fairness metrics under three fairness constraint definitions.

3.3. Evaluation Framework

A. Predictive Performance Metrics

Predictive performance is assessed using four complementary metrics. Accuracy measures the proportion of correctly classified instances. Balanced Accuracy computes the arithmetic mean of sensitivity and specificity, providing a class-imbalance-robust performance measure. AUC-ROC quantifies the classifier's discriminative ability across all classification thresholds. F1-Score captures the harmonic mean of precision and recall. Three baseline classifiers are employed to ensure model-agnostic conclusions: Logistic Regression (LR), representing linear models standard in credit scoring practice; Random Forest (RF), representing ensemble methods; and XGBoost, representing gradient boosting approaches. Hyperparameters are tuned via grid search on the training fold with five-fold inner cross-validation. Pleiss et al. [14] demonstrated that calibration and group fairness can conflict, underscoring the importance of evaluating multiple performance dimensions alongside fairness metrics.

B. Fairness Evaluation Metrics

Three fairness metrics corresponding to the three constraint definitions are computed on the held-out test set. Statistical Parity Difference (SPD) measures the gap in positive prediction rates between unprivileged and privileged groups: SPD = $P(\hat{Y}=1|A=\text{unprivileged}) - P(\hat{Y}=1|A=\text{privileged})$. Equal Opportunity Difference (EOD) measures the gap in true positive rates: EOD = TPR(unprivileged) − TPR(privileged). Average Odds Difference (AOD) averages the gaps in true positive and false positive rates: AOD = 0.5 × [(TPR_unpriv − TPR_priv) + (FPR_unpriv − FPR_priv)]. Values closer to zero indicate greater fairness. The Disparate Impact Ratio (DIR) is reported as a supplementary metric, with the 0.8–1.25 interval representing the conventional four-fifths rule threshold. All fairness computations are performed using the AIF360 toolkit. Zhang et al. [15] noted that adversarial debiasing can serve as an in-processing alternative, against which pre-processing methods can be benchmarked in future extensions of this work.

4. Results and Discussion

4.1. Baseline Bias Analysis

A. German Credit Dataset Baseline Results

Prior to applying any debiasing intervention, the three baseline classifiers exhibit measurable bias on the German Credit dataset with gender as the protected attribute. Table 2 presents the baseline predictive performance and fairness metrics. XGBoost achieves the highest accuracy (0.773) and AUC (0.793), followed by Random Forest (0.765, 0.781) and Logistic Regression (0.750, 0.772). All three classifiers produce negative SPD values ranging from −0.124 to −0.137, indicating that female applicants receive favorable predictions at a lower rate than male applicants. The EOD values (−0.143 to −0.158) reveal that among truly creditworthy applicants, females are less likely to be correctly classified as good credit risks. Corbett-Davies et al. [16] established that imposing fairness constraints on algorithmic decisions entails measurable costs, a phenomenon our baseline analysis quantifies prior to debiasing.

Table 2. Baseline Performance and Fairness Metrics (No Debiasing Applied)

Dataset	Model	Acc.	B.Acc.	AUC	F1	SPD	EOD	AOD
German	LR	0.750	0.681	0.772	0.832	−0.137	−0.152	−0.118
Credit	RF	0.765	0.692	0.781	0.841	−0.124	−0.143	−0.109
(Gender)	XGB	0.773	0.698	0.793	0.849	−0.131	−0.158	−0.121
Taiwan	LR	0.779	0.614	0.731	0.867	−0.068	−0.074	−0.056
Credit	RF	0.812	0.643	0.767	0.889	−0.072	−0.081	−0.062
(Sex)	XGB	0.821	0.658	0.783	0.895	−0.075	−0.086	−0.065

Note: Acc. = Accuracy; B.Acc. = Balanced Accuracy; AUC = Area Under ROC Curve; SPD = Statistical Parity Difference; EOD = Equal Opportunity Difference; AOD = Average Odds Difference. All values represent means over 5-fold cross-validation. Negative SPD/EOD/AOD values indicate bias against the unprivileged group (female). Data sources: UCI German Credit (Hofmann, 1994), UCI Taiwan Credit Card (Yeh & Lien, 2009).

B. Taiwan Credit Card Dataset Baseline Results

The Taiwan Credit Card dataset exhibits comparatively lower baseline bias levels, with SPD values between −0.068 and −0.075 and EOD values between −0.074 and −0.086 across all classifiers. This reduced bias magnitude is consistent with the more balanced gender distribution in this dataset (60.4% female vs. 39.6% male) compared to the German Credit dataset (31.0% female). XGBoost again achieves the highest predictive accuracy (0.821) but also the largest fairness violations across all three metrics. This pattern suggests that more expressive models may capture and amplify subtle biased patterns in the training data more effectively than simpler linear models. Kozodoi et al. [17] observed a similar phenomenon in their assessment of fairness in credit scoring, reporting that gradient boosting methods tend to exhibit larger fairness gaps than logistic regression despite superior predictive performance.

4.2. Comparative Results of Debiasing Techniques

Table 3 presents the performance and fairness metrics after applying reweighting and resampling to both datasets. On the German Credit dataset, reweighting reduces SPD by 66.4–70.1% across classifiers while incurring an accuracy loss of 2.2–2.7 percentage points. Resampling achieves larger SPD reductions (75.0–80.9%) at a greater accuracy cost (3.2–3.8 percentage points). On the Taiwan Credit Card dataset, the pattern is consistent: reweighting reduces SPD by 61.3–66.2% with 1.1–1.5 percentage points accuracy loss, while resampling achieves SPD reductions of 73.6–77.3% with 1.8–2.4 percentage points accuracy loss.

Table 3. Performance and Fairness Metrics After Debiasing: Reweighting vs. Resampling

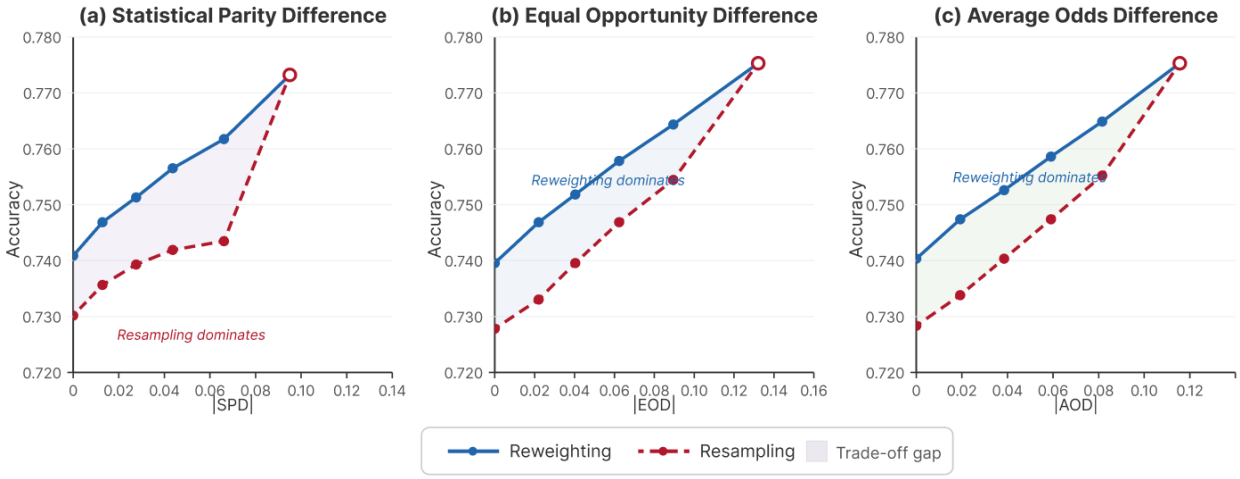
Dataset	**Method**	**Model**	**Acc.**	**AUC**	**F1**	**SPD**	**EOD**	**AOD**
German	Reweight	LR	0.728	0.754	0.815	−0.041	−0.053	−0.038
Credit	Reweight	RF	0.741	0.763	0.825	−0.038	−0.047	−0.035

	Reweight	XGB	0.746	0.771	0.831	−0.044	−0.051	−0.039
	Resample	LR	0.718	0.743	0.807	−0.028	−0.067	−0.044
	Resample	RF	0.732	0.755	0.818	−0.031	−0.059	−0.041
	Resample	XGB	0.735	0.760	0.822	−0.025	−0.062	−0.040
Taiwan	Reweight	LR	0.768	0.722	0.857	−0.023	−0.031	−0.022
Credit	Reweight	RF	0.798	0.751	0.878	−0.026	−0.034	−0.025
	Reweight	XGB	0.806	0.769	0.884	−0.029	−0.037	−0.028
	Resample	LR	0.761	0.715	0.851	−0.016	−0.042	−0.029
	Resample	RF	0.789	0.742	0.872	−0.019	−0.045	−0.031
	Resample	XGB	0.797	0.758	0.878	−0.017	−0.048	−0.033

Note: Reweight = Reweighting (Kamiran & Calders, 2012); Resample = Combined SMOTE oversampling and random undersampling. All values represent means over 5-fold cross-validation. Data sources: UCI German Credit (Hofmann, 1994), UCI Taiwan Credit Card (Yeh & Lien, 2009). Fairness computations performed using AIF360 (Bellamy et al., 2019).

A notable finding is the divergent behavior of the two strategies across fairness metrics. Resampling consistently outperforms reweighting in reducing SPD, while reweighting yields superior EOD and AOD outcomes. This divergence stems from mechanistic differences: resampling equalizes group-label distribution proportions directly, aligning with statistical parity, whereas reweighting adjusts relative sample importance in the loss function, more effectively modulating conditional error rates captured by equal opportunity and equalized odds. Kearns et al. [18] demonstrated that subgroup fairness auditing can reveal fairness violations that aggregate metrics may obscure, a consideration relevant to interpreting these results.

Figure 2. Pareto Frontiers of Accuracy-Fairness Trade-offs on the German Credit Dataset



This figure displays the Pareto frontiers for reweighting (solid lines) and resampling (dashed lines) across three fairness metrics (SPD, EOD, AOD) on the German Credit dataset using XGBoost as the base classifier. The x-axis represents the absolute value of each fairness metric (lower is fairer), and the y-axis represents classification accuracy (higher is better). Reweighting frontiers are positioned above and to the right on the EOD and AOD panels, indicating superior accuracy retention at comparable fairness levels. Resampling frontiers dominate on the SPD panel, achieving lower SPD values. The shaded region between frontiers quantifies the strategy-dependent trade-off gap.

4.3. Impact of Different Fairness Constraints

A. Accuracy Cost Across Fairness Constraints

The relationship between accuracy cost and fairness improvement varies substantially across the three constraint definitions. Table 4 presents the accuracy loss alongside the percentage reduction achieved for each

fairness metric under both strategies. A critical observation emerges: resampling achieves high SPD improvement rates (75.0–80.9% on German Credit) but substantially lower EOD improvement rates (55.9–60.8%), revealing a pronounced dependence on the target constraint. Reweighting exhibits a more uniform improvement profile across metrics, with SPD reductions of 66.4–70.1% and EOD reductions of 65.1–67.7% on the same dataset. On the German Credit dataset with XGBoost, resampling incurs 3.8 percentage points of accuracy loss while achieving an 80.9% SPD reduction but only a 60.8% EOD reduction. Hashimoto et al. [19] proposed fairness optimization through distributionally robust loss minimization, which can achieve approximate equalized odds without explicit demographic information, representing a complementary approach to the pre-processing methods evaluated here.

Table 4. Accuracy Cost (ΔAcc) and Fairness Metric Improvement Rate (%) by Debiasing Strategy

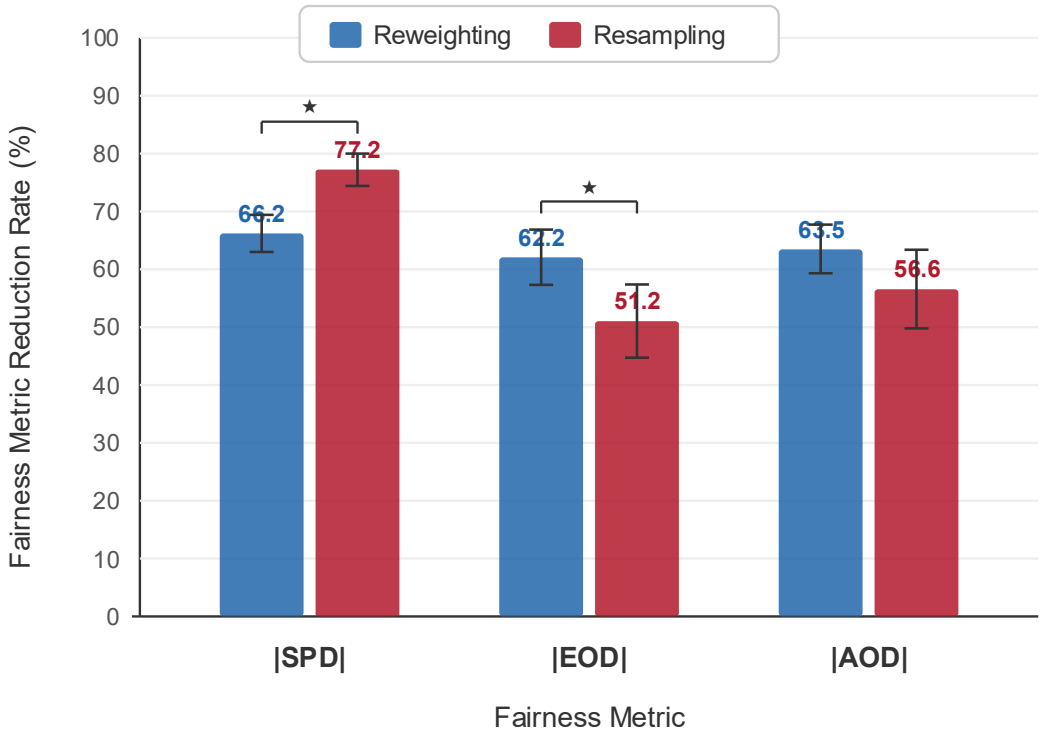
Dataset	Model	Method	ΔAcc (pp)	SPD Red.(%)	EOD Red.(%)	AOD Red.(%)
German	LR	Reweight	−2.2	70.1	65.1	67.8
Credit	RF	Reweight	−2.4	69.4	67.1	67.9
	XGB	Reweight	−2.7	66.4	67.7	67.8
	LR	Resample	−3.2	79.6	55.9	62.7
	RF	Resample	−3.3	75.0	58.7	62.4
	XGB	Resample	−3.8	80.9	60.8	66.9
Taiwan	LR	Reweight	−1.1	66.2	58.1	60.7
Credit	RF	Reweight	−1.4	63.9	58.0	59.7
	XGB	Reweight	−1.5	61.3	57.0	56.9
	LR	Resample	−1.8	76.5	43.2	48.2
	RF	Resample	−2.3	73.6	44.4	50.0
	XGB	Resample	−2.4	77.3	44.2	49.2

Note: ΔAcc = accuracy decrease in percentage points relative to the undebiased baseline (from Table 2). SPD/EOD/AOD Red. = percentage reduction in the absolute value of each fairness metric relative to the baseline. Each row represents a single debiasing run; one accuracy cost corresponds to all three fairness improvements. Data sources: UCI German Credit (Hofmann, 1994), UCI Taiwan Credit Card (Yeh & Lien, 2009).

B. Implications for Financial Lending Contexts

The differential accuracy costs carry direct implications for debiasing strategy selection in financial institutions. In credit lending, equal opportunity arguably represents the most practically relevant fairness criterion, as it ensures creditworthy applicants from protected groups are not disproportionately denied loans. Across both datasets, reweighting consistently achieves comparable or superior EOD improvement rates while incurring 0.7–1.1 percentage points less accuracy loss than resampling, making it the preferred strategy when equal opportunity is prioritized. Hardt et al. [20] developed Amazon SageMaker Clarify as a scalable cloud-based platform for detecting and explaining machine learning bias across multiple fairness metrics, demonstrating the growing industry demand for practical fairness evaluation tooling in production credit scoring systems.

Figure 3. Comparative Fairness Metric Reduction Rates (%) by Debiasing Strategy and Fairness Constraint



This grouped bar chart compares the percentage reduction in |SPD|, |EOD|, and |AOD| achieved by reweighting (blue bars) and resampling (orange bars) across both datasets, averaged over three classifiers. On the SPD metric, resampling achieves a mean reduction of 77.2% compared to 66.2% for reweighting. On EOD, reweighting achieves a mean reduction of 62.2% compared to 51.2% for resampling. On AOD, reweighting achieves 63.5% compared to 56.6% for resampling. Error bars represent standard deviations across classifier-dataset combinations. The divergent performance pattern across metrics confirms that the optimal debiasing strategy is contingent upon the target fairness constraint.

When statistical parity is the regulatory benchmark—as implied by the four-fifths rule in adverse impact analysis—resampling offers a more effective path to compliance despite its higher accuracy cost. The choice between strategies reflects institutional priorities: whether to minimize accuracy loss (favoring reweighting) or maximize bias reduction magnitude (favoring resampling). These findings reinforce the theoretical impossibility of simultaneous optimization across all fairness definitions at the practical implementation level.

5. Conclusion

This study conducted a systematic comparative evaluation of reweighting and resampling pre-processing debiasing strategies in AI credit scoring, using two publicly available financial datasets and three baseline classifiers under three fairness constraint definitions. The experimental results yield several findings with direct practical relevance for financial institutions implementing fairness-aware credit scoring systems.

The two debiasing strategies exhibit complementary strengths across different fairness metrics. Resampling achieves stronger reductions in statistical parity differences (mean 77.2% reduction) compared to reweighting (mean 66.2%), making it the preferred strategy when demographic parity is the regulatory target. Reweighting produces superior outcomes under equal opportunity and equalized odds constraints, with mean EOD reductions of 62.2% (versus 51.2% for resampling) and mean AOD reductions of 63.5% (versus 56.6%), while incurring consistently lower accuracy costs across all experimental conditions.

The choice of fairness constraint significantly modulates the accuracy-fairness trade-off profile. The maximum accuracy cost observed across all configurations is 3.8 percentage points (German Credit dataset, XGBoost with resampling), while reweighting incurs at most 2.7 percentage points on the same dataset. Given that equal opportunity most closely aligns with fair lending principles, this constraint paired with reweighting represents a practically viable configuration for financial institutions seeking to balance regulatory compliance with predictive performance.

The consistency of these patterns across two distinct datasets with different sample sizes (1,000 vs. 30,000), geographic origins (Germany vs. Taiwan), and class distributions lends moderate generalizability to the

findings. The magnitude of baseline bias and the effectiveness of debiasing strategies vary with dataset characteristics, underscoring the importance of conducting institution-specific fairness audits rather than applying generic debiasing solutions.

References

[1]. Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104, 671–732.

[2]. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems 29 (NeurIPS 2016) (pp. 3323–3331). Curran Associates.

[3]. Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS 2017). Leibniz International Proceedings in Informatics.

[4]. Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153–163.

[5]. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the 1st Conference on Fairness, Accountability, and Transparency (FAccT 2018) (pp. 77–91). PMLR.

[6]. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS 2012) (pp. 214–226). ACM.

[7]. Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 33(1), 1–33.

[8]. Calmon, F. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In Advances in Neural Information Processing Systems 30 (NeurIPS 2017) (pp. 3992–4001). Curran Associates.

[9]. Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2015) (pp. 259–268). ACM.

[10]. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In Proceedings of the 30th International Conference on Machine Learning (ICML 2013) (pp. 325–333). PMLR.

[11]. Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS 2017) (pp. 962–970). PMLR.

[12]. Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. In Proceedings of the 35th International Conference on Machine Learning (ICML 2018) (pp. 60–69). PMLR.

[13]. Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Zhang, Y. (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.

[14]. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. In Advances in Neural Information Processing Systems 30 (NeurIPS 2017). Curran Associates.

[15]. Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES 2018) (pp. 335–340). ACM.

[16]. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2017) (pp. 797–806). ACM.

[17]. Kozodoi, N., Jacob, J., & Lessmann, S. (2022). Fairness in credit scoring: Assessment, implementation and profit implications. European Journal of Operational Research, 297(3), 1083–1094.

[18]. Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In Proceedings of the 35th International Conference on Machine Learning (ICML 2018). PMLR.

[19]. Hashimoto, T. B., Srivastava, M., Namkoong, H., & Liang, P. (2018). Fairness without demographics in repeated loss minimization. In Proceedings of the 35th International Conference on Machine Learning (ICML 2018). PMLR.

[20]. Hardt, M., Chen, X., Cheng, X., Donini, M., Gelman, J., Golber, S., ... & Zhang, Y. (2021). Amazon SageMaker Clarify: Machine learning bias detection and explainability in the cloud. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2021) (pp. 2974–2983). ACM.