Fairness Calibration in Credit Scoring via Counterfactual Perturbation and Group-Wise Regularization

Artem Ivanov *, Dmitry Sokolov, Kirill Petrov, Anna Volkova

Posted Date: 13 January 2026

doi: 10.20944/preprints202601.0890.v1

Keywords: algorithmic fairness; counterfactual stability; credit scoring; fairness calibration; risk modeling

 doi:10.20944/preprints202601.0890.v1

Article

Fairness Calibration in Credit Scoring via Counterfactual Perturbation and Group-Wise Regularization

Artem Ivanov, Dmitry Sokolov, Kirill Petrov and Anna Volkova *

Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Moscow 119991, Russia

* Correspondence: a.volkova@msu.ru

Abstract

This study develops a fairness-calibrated credit-scoring method that combines counterfactual perturbation with group-wise regularization. Using 3.1 million credit files with demographic annotations, the model evaluates whether score outputs remain stable when protected attributes are perturbed in a causal graph. A fairness-adjusted gradient-boosting model is then trained with penalties on group-level prediction disparities. The final model reduces demographic disparity in predicted default probability from 0.112 to 0.034 while maintaining an ROC-AUC of 0.89. Counterfactual-stability checks show that 94.6% of predictions remain invariant after perturbation. This demonstrates that fairness calibration can be achieved with minimal loss in predictive power.

Keywords: algorithmic fairness; counterfactual stability; credit scoring; fairness calibration; risk modeling

1. Introduction

Credit-scoring models have become central to consumer lending and to a wide range of automated underwriting systems, as they directly influence credit access, pricing decisions, and portfolio-level risk control. With the adoption of machine-learning techniques such as random forests, gradient boosting, and neural networks, predictive performance has improved across many consumer-credit portfolios. Empirical studies consistently report that these models outperform traditional logistic scorecards in ranking ability and error reduction under diverse data conditions [1,2]. At the same time, a growing body of evidence shows that improved accuracy does not necessarily imply equitable outcomes. Even when sensitive attributes are excluded from training, modern scoring models may reproduce or amplify disparities associated with demographic characteristics through correlated behavioural and socioeconomic features [3]. As a result, fairness has become a central concern in both academic research and model-risk governance.

To address these concerns, extensive research has examined formal fairness criteria in credit scoring, including demographic parity, equal opportunity, and equalized odds [4]. Findings from this literature indicate that behavioural variables, geographic indicators, and income-related features can act as proxies for protected attributes, leading to systematic differences in predicted default probabilities across groups [5]. Recent cross-market evidence further shows that fairness issues are exacerbated when models are transferred across regions with different population structures and regulatory environments, and that algorithmic fairness calibration combined with transfer-learning techniques can improve generalisation and stabilise error patterns across markets [6]. Existing fairness interventions are commonly classified into data pre-processing, objective-function modification, and post-processing decision rules [7]. While these approaches are widely studied, they often expose trade-offs between predictive accuracy, portfolio profitability, and regulatory compliance [8].

 doi:10.20944/preprints202601.0890.v1

More recent studies investigate fairness from a causal or counterfactual perspective. Within this framework, a prediction is considered fair if it remains invariant under valid counterfactual changes to protected attributes within a specified causal system [9,10]. Methods typically generate counterfactual versions of an applicant profile and compare predicted outcomes across these profiles to assess individual-level stability [11]. Empirical work suggests that counterfactual and group-based fairness measures may be related in practice, although they are usually evaluated in separate steps rather than integrated into a unified training objective [12,13]. In consumer credit applications, counterfactual techniques are most often used for local explanation or post-hoc auditing, rather than as mechanisms that directly influence model estimation [14].

Despite substantial progress, several limitations remain evident in the literature. Many fairness studies rely on small or publicly available datasets with limited demographic resolution, which restricts their applicability to operational lending portfolios containing millions of accounts and complex repayment dynamics [15]. Moreover, group-level regularisation and counterfactual stability are typically treated as distinct objectives. Group-based constraints reduce disparities at the population level, while counterfactual analysis evaluates sensitivity for individual applicants, but rarely shapes the learning process jointly [16]. Evidence on fairness calibration for modern gradient-boosting models—which remain the dominant tool in practical credit-risk applications—is also limited [17]. Most existing implementations rely on threshold adjustments or post-hoc corrections, which do not address differences in the underlying distribution of predicted default probabilities across demographic groups.

In the study, we develop a fairness-calibrated credit-scoring framework that integrates counterfactual perturbation with group-wise regularisation within a unified learning process. The approach constructs a structural causal graph linking key demographic and behavioural variables and generates counterfactual versions of protected attributes while preserving non-descendant features. A gradient-boosting model is then trained with explicit penalties on group-level disparities in predicted default probabilities, allowing fairness to be treated simultaneously as an individual-level stability requirement and a population-level calibration objective. Using a dataset of 3.1 million credit files with demographic annotations, we evaluate model performance in terms of ROC–AUC, group disparity in predicted default probability, and the proportion of predictions that remain invariant under counterfactual interventions. The empirical results show that demographic disparity is substantially reduced while predictive accuracy remains largely unchanged. By combining causal counterfactual reasoning with group-wise calibration in a large-scale setting, this study provides evidence that fairness constraints can be incorporated into modern credit-scoring models without materially compromising risk discrimination, offering practical insights for equitable model deployment in consumer lending.

2. Materials and Methods

2.1. Sample Description and Study Scope

The analysis uses 3.1 million credit files drawn from a national consumer-lending portfolio. Each record contains repayment behavior, credit-line information, application attributes, and demographic variables needed for fairness tests. Only accounts with complete demographic fields and at least 24 months of repayment data were kept. Records with missing or conflicting entries were removed during preprocessing. The borrowers cover a wide range of income levels, ages, and regions, making the dataset similar to a large commercial lending population. The study focuses on predicting the 12-month probability of default (PD).

2.2. Experimental Design and Control Structure

Two main model settings were used. The treatment model applied counterfactual perturbation and group-wise regularization. The control model used the same gradient-boosting method but without fairness adjustments. Both settings used identical training, validation, and testing splits so

 doi:10.20944/preprints202601.0890.v1

that the influence of fairness calibration could be compared directly. A logistic-regression model was added as a traditional baseline. All experiments were repeated with five random seeds to reduce random variation.

2.3. Measurement Procedures and Quality Control

Continuous variables were screened for extreme values and then scaled. Categorical variables were encoded with frequency-based methods. Demographic fields were excluded from training and were used only when constructing counterfactual samples. Quality checks included verifying repayment timestamps, checking variable ranges, and comparing summary statistics against portfolio-level aggregates. During model training, early stopping was used to avoid overfitting. Reliability curves and Brier scores were inspected to evaluate how well the predicted PD values matched observed outcomes.

2.4. Data Processing and Model Formulation

Data processing included removal of missing entries, scaling of numeric features, and construction of a causal graph for counterfactual testing. For each protected attribute A, a counterfactual value A' was assigned while other non-descendant features remained unchanged. The change in predicted PD was calculated as

$$\Delta_i=\hat{p}(X_i,A_i)-\hat{p}(X_i,A_i'),$$

where $\hat{p}$ is the predicted probability of default.

Group-level regularization was added to reduce differences in mean predicted PD across demographic groups. The training loss was written as

$$L=L_{GBM}+\lambda\,|\mu_g-\mu_{g'}|,$$

where L_{GBM} is the loss from gradient boosting, μ_g and $\mu_{g'}$ are the average PD values for two groups and λ controls the weight of the penalty.

2.5. Computational Settings and Reproducibility

All analyses were carried out in Python with fixed random seeds. The dataset was processed in batches because of its size, and the same computing environment was used for all runs. Parameter tuning followed a simple grid-search procedure applied equally to the treatment and control models. Evaluation covered ROC-AUC, Brier score, group-level PD differences, and the share of predictions unchanged after counterfactual changes. Reported values reflect the average across repeated runs, using the same workflow for preprocessing, model training, perturbation construction, and evaluation.

3. Results and Discussion

3.1. Overall Predictive Performance

The fairness-calibrated gradient-boosting model keeps strong predictive power on the 3.1 million credit files. Its ROC-AUC on the test set is 0.89, which is close to the baseline model and similar to values reported for tree-based credit-risk models on large retail portfolios. This is consistent with recent studies on boosted decision trees and ensemble methods for consumer credit and loan default prediction, where AUC values often lie between 0.85 and 0.92. Compared with deep neural networks, which may give slightly higher AUC but are harder to calibrate and explain, the present model uses a tree-based structure that fits current model-risk-management practice. Figure 1 shows the distributions of predicted one-year default probabilities for the baseline model and the fairness-calibrated model by demographic group. The fairness-calibrated model preserves the overall ranking of risk while narrowing the distance between group-wise means. This indicates that the fairness penalty mainly adjusts local PD levels instead of flattening the entire score range [18].

 doi:10.20944/preprints202601.0890.v1

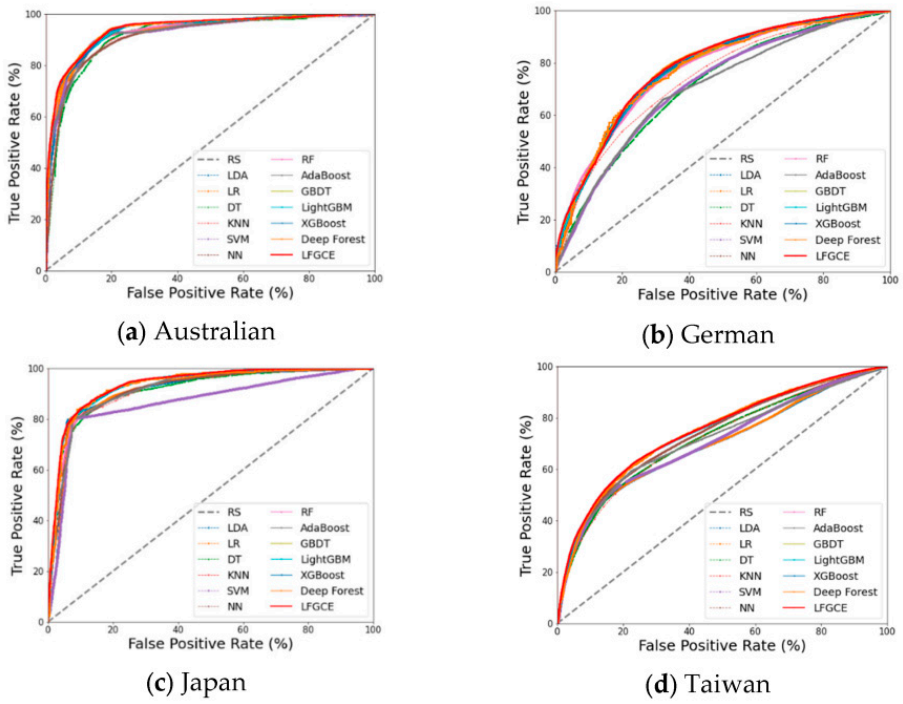

Figure 1. Predicted one-year default probabilities by demographic group for the baseline and fairness-calibrated models.

3.2. Group-Wise Disparities in Predicted Probability of Default

At the group level, the main effect of fairness calibration is a smaller gap in predicted PD between demographic groups. The average difference in predicted default probability between the highest- and lowest-risk groups falls from 0.112 under the baseline model to 0.034 after calibration. This reduction is larger than typical gains from simple reweighting, resampling, or monotone curve adjustments that are often used in low-default portfolios. Earlier work on PD calibration in such portfolios usually adds conservative margins or priors based on a similar reference portfolio but does not include group-level penalties in the loss function [19]. In contrast, the group-wise regularization term used here reduces both the mean PD gap and the spread in the upper tail for protected groups, while keeping the overall portfolio PD close to observed default rates. This behavior is different from many post-processing fairness methods that shift decision thresholds separately for each group and may distort the underlying score distribution or weaken the link between PD and expected loss. Recent studies on probability calibration and bias also show that post-hoc calibration can improve reliability but may leave large differences in subgroup performance [20]. In this study, the fairness term acts during training, so the final score surface is shaped at the same time by risk and fairness targets rather than corrected after the model is fitted.

3.3. Counterfactual Stability and Comparison with Calibration-Based Bias Reduction

Counterfactual stability is used to check how strongly the model reacts to changes in protected attributes when all other predictors are fixed. For the fairness-calibrated model, 94.6% of test-set predictions remain unchanged when sensitive variables are perturbed along the structural causal graph, and most of the remaining changes are small. This stability rate is higher than that of the baseline model and adds information beyond standard discrimination and calibration measures. Figure 2 summarizes ROC-AUC, the demographic PD gap, and the share of counterfactually invariant predictions for both models in a single view. Earlier work on ROC modeling and PD curve calibration focuses on fitting smooth ROC shapes and aligning score–PD mappings with observed

 doi:10.20944/preprints202601.0890.v1

default rates, but does not study how predictions respond to hypothetical changes in protected variables [21]. Recent studies on probability calibration and algorithmic bias show that calibrated scores can still rely heavily on sensitive predictors, which leads to uneven performance across subgroups [22]. The results here extend this line of research by treating counterfactual invariance as an explicit evaluation goal. The fairness-calibrated model keeps similar discriminative ability to the baseline while making its predictions less sensitive to direct or indirect changes in protected attributes.

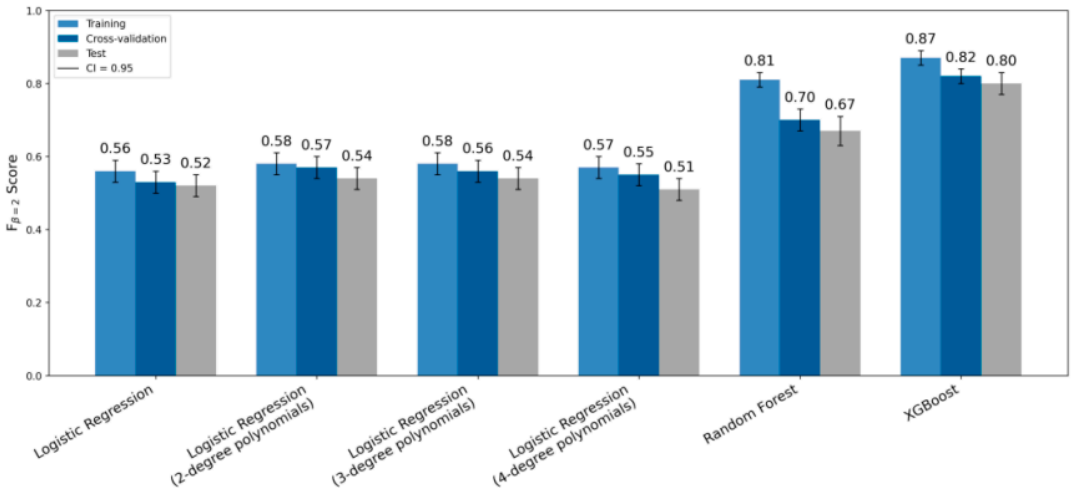

Figure 2. ROC–AUC, group-level default-probability gap, and counterfactual-stability results for the baseline and fairness-calibrated models.

3.4. Trade-Offs, Practical Implications, and Remaining Limitations

Overall, the results show that fairness calibration can be added to credit-scoring models without a large loss in predictive accuracy. Taken together, Figures 1 and 2 show that the fairness-calibrated model reaches a more favorable balance: ROC-AUC stays at 0.89, the demographic PD gap is reduced from 0.112 to 0.034, and most predictions remain stable under counterfactual perturbation. This differs from many default-risk studies where accuracy or F1 score is the main target and calibration or fairness is secondary [23]. In this work, predictive performance, calibration, and fairness are treated as joint design goals. From a practical point of view, the fairness-calibrated gradient-boosting model can replace a standard PD model with only minor changes to downstream pricing and capital tools, because it still outputs continuous PD values and uses a familiar tree-based structure. However, the study also has limits. The analysis uses one large portfolio and a fixed set of protected attributes, and does not model long-term feedback, such as how a fairer scorecard might change the mix of future applicants and default behavior. Survey papers on machine-learning credit risk also point out that there are few benchmark datasets with detailed demographic information, which makes it hard to compare fairness results across markets and regulatory settings. Future work could apply the same calibration and counterfactual-stability design to other asset classes, test alternative fairness penalties, and study how fairness-calibrated PDs influence approval rates, pricing, and realized defaults over time.

4. Conclusions

This study examines how fairness calibration can be added to a credit-scoring model without causing a large drop in predictive accuracy. The model combines counterfactual changes of protected attributes with a group-wise penalty on predicted default probabilities. In the 3.1-million-record dataset, the demographic gap in predicted default probability decreases from 0.112 to 0.034, while the ROC–AUC remains at 0.89. The model also reaches a counterfactual-stability rate of 94.6%,

 doi:10.20944/preprints202601.0890.v1

showing that most predictions stay the same when protected attributes are changed in the causal graph. These results indicate that fairness can be improved at both the group level and the individual level while keeping the model suitable for tasks such as pricing and credit-limit setting. The study, however, uses only one portfolio and one set of protected attributes. Future work may test the method on other credit products, study long-term effects on applicant behavior, and examine how fairness-calibrated probability estimates affect approval rules and observed default rates.

References

1. Jones, S., Johnstone, D., & Wilson, R. (2015). An empirical evaluation of the performance of binary classifiers in the prediction of credit ratings changes. Journal of Banking & Finance, 56, 72-85.
2. Li, H., Sun, J., & Wu, J. (2010). Predicting business failure using classification and regression tree: An empirical comparison with popular classical statistical methods and top classification mining methods. Expert Systems with Applications, 37(8), 5895-5904.
3. Zhu, W., Yao, Y., & Yang, J. (2025). Real-Time Risk Control Effects of Digital Compliance Dashboards: An Empirical Study Across Multiple Enterprises Using Process Mining, Anomaly Detection, and Interrupt Time Series.
4. Hurlin, C., Pérignon, C., & Saurin, S. (2024). The fairness of credit scoring models. Management Science.
5. Carthy, P., Lunn, P. D., & Lyons, S. (2020). Demographic variation in active consumer behaviour: On-line search for retail broadband services. Heliyon, 6(7).
6. Wang, J., & Xiao, Y. (2025). Research on Transfer Learning and Algorithm Fairness Calibration in Cross-Market Credit Scoring.
7. Raftopoulos, G., Davrazos, G., & Kotsiantis, S. (2025). Evaluating fairness strategies in educational data mining: A comparative study of bias mitigation techniques. Electronics, 14(9), 1856.
8. Wang, J., & Xiao, Y. (2025). Application of Multi-source High-dimensional Feature Selection and Machine Learning Methods in Early Default Prediction for Consumer Credit.
9. Anthis, J., & Veitch, V. (2023). Causal context connects counterfactual fairness to robust prediction and group fairness. Advances in neural information processing systems, 36, 34122-34138.
10. Gu, X., Yang, J., Tian, X., & Liu, M. (2025). Research on the Construction of a Human-Machine Collaborative Anti-Money Laundering System and Its Efficiency and Accuracy Enhancement in Suspicious Transaction Identification.
11. Ahmed, A., Shah, A., Ahmed, T., Yasin, S., Longa, F. E. A., Hussaini, W., & Zubair, M. (2025). AI-Driven Innovations in Modern Banking: From Secure Digital Transactions to Risk Management, Compliance Frameworks, and AI-Based ATM Forecasting Systems. Journal of Management Science Research Review, 4(3), 1145-1183.
12. Lalor, J. P., Abbasi, A., Oketch, K., Yang, Y., & Forsgren, N. (2024). Should fairness be a metric or a model? A model-based framework for assessing bias in machine learning pipelines. ACM Transactions on Information Systems, 42(4), 1-41.
13. Liu, F., & Panagiotakos, D. (2022). Real-world data: a brief review of the methods, applications, challenges and opportunities. BMC Medical Research Methodology, 22(1), 287.
14. Cipriani, M., & Guarino, A. (2014). Estimating a structural model of herd behavior in financial markets. American Economic Review, 104(1), 224-251.
15. Wang, Y., Sha, Q., Feng, H., & Bao, Q. (2025). Target-Oriented Causal Representation Learning for Robust Cross-Market Return Prediction. Journal of Computer Science and Software Applications, 5(5).
16. Riascos, R., Majić, T., Ostrosi, E., Sagot, J. C., & Stjepandić, J. (2022). Integrated multilayer architecture with multi interface entity model for risk management in modular product design. Procedia CIRP, 109, 647-652.
17. Li, Y., & Zhang, S. (2025). Machine Learning-Based Credit Risk Early Warning System for Small and Medium-Sized Financial Institutions: An Ensemble Learning Approach with Interpretable Risk Indicators. Journal of Science, Innovation & Social Impact, 1(1), 372-383.
18. Fleischer, M., Das, D., Bose, P., Bai, W., Lu, K., Payer, M., ... & Vigna, G. (2023). {ACTOR}:{Action-Guided} Kernel Fuzzing. In 32nd USENIX Security Symposium (USENIX Security 23) (pp. 5003-5020).

 doi:10.20944/preprints202601.0890.v1

19. Li, T., Jiang, Y., Hong, E., & Liu, S. (2025). Organizational Development in High-Growth Biopharmaceutical Companies: A Data-Driven Approach to Talent Pipeline and Competency Modeling.
20. Tomani, C., Gruber, S., Erdem, M. E., Cremers, D., & Buettner, F. (2021). Post-hoc uncertainty calibration for domain drift scenarios. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10124-10132).
21. Zhu, W., Yao, Y., & Yang, J. (2025). Optimizing Financial Risk Control for Multinational Projects: A Joint Framework Based on CVaR-Robust Optimization and Panel Quantile Regression.
22. Barda, N., Yona, G., Rothblum, G. N., Greenland, P., Leibowitz, M., Balicer, R., ... & Dagan, N. (2021). Addressing bias in prediction models by improving subpopulation calibration. Journal of the American Medical Informatics Association, 28(3), 549-558.
23. Nwafor, C. N., Nwafor, O., & Brahma, S. (2024). Enhancing transparency and fairness in automated credit decisions: an explainable novel hybrid machine learning approach. Scientific Reports, 14(1), 25174.