

Engineering Trust in Algorithmic Lending: Explainable, Fairness-Aware Credit Risk Modeling for Financial Inclusion

Rezoan Sultan¹ and Marwan Hossain²

¹Department of Electrical and Computer Engineering

Morgan State University, Baltimore, MD, USA

resul2@morgan.edu

²Department of Electrical and Electronic Engineering

European University of Bangladesh, Dhaka, Bangladesh

marwan@eub.edu.bd

<https://doi.org/10.68414/VJLR7181>

Abstract

Consumer credit decisions are increasingly influenced by machine learning, but its use creates a major conflict for financial inclusion because models that are geared for predictive accuracy may encode and magnify inequalities against the very groups that inclusion-oriented lending aims to assist. This study asks whether predictive performance alone is a sufficient basis for trust in algorithmic lending. Using the UCI Default of Credit Card Clients dataset (30,000 borrowers in Taiwan), we train and compare logistic regression, random forest and gradient-boosted decision tree (XGBoost) classifiers and subject the strongest model to SHapley Additive exPlanations (SHAP) interpretation, a demographic fairness audit across sex, education and age, a feature-ablation test and a post-processing mitigation. XGBoost attains a ROC-AUC of 0.775, on par with established benchmarks, confirming that raw accuracy is not the differentiating contribution. The audit reveals moderate but real disparities in false-positive rates, with the youngest and least-educated applicants most often wrongly flagged; we also show that naively including a very small subgroup substantially overstates apparent disparity. The assigned credit limit is strongly stratified by education and age, yet removing it does not reduce the disparity, illustrating that “fairness through unawareness” is insufficient because the signal is redundantly encoded in correlated features. Group-aware decision thresholds, a straightforward post-processing solution, lower the education false-positive gap from 0.082 to 0.005 at the expense of 0.4 percentage points of accuracy. We suggest a layered evaluation approach and contend that explainability, meticulous fairness auditing and mitigation are necessary conditions for reliable, inclusion-focused credit AI.

Keywords: credit scoring, explainable AI, SHAP, algorithmic fairness, financial inclusion, trustworthy machine learning.

Introduction

Financial technology has moved credit decisioning from human underwriting toward automated, data-driven scoring. Lenders increasingly rely on machine learning (ML) models to estimate the probability that an applicant will default and these models now mediate access to credit cards, consumer loans and microfinance products worldwide. For proponents of financial inclusion, this shift carries genuine promise: data-driven models can, in principle, extend credit to applicants whom traditional scorecards overlook, including the underbanked and those with thin credit files (Berg et al., 2020). Yet the same flexibility that allows ML models to uncover predictive structure can also reproduce and even amplify, historical disparities in who receives credit and on what terms (Fuster et al., 2022).

This tension sits at the heart of the present study and of the conference theme of engineering trust. A model that achieves high predictive accuracy is not, by that fact alone, trustworthy. Trust in a lending model depends on whether its decisions can be explained, whether the features it relies upon are legitimate and whether the errors it makes are distributed equitably across the population it serves. When a model is opaque, an applicant who is denied credit cannot know why, a lender cannot defend the decision to a regulator and a community cannot tell whether the system is fair. Accuracy answers none of these questions.

We examine these issues on a widely used, transparent benchmark: the UCI Default of Credit Card Clients dataset. Rather than chasing marginal accuracy gains, we treat accuracy as a baseline and ask what an explainability analysis and a demographic fairness audit reveal about the trustworthiness of a representative model. Specifically, we address three research questions:

RQ1. How accurately do standard ML models predict default on this benchmark and does model choice materially change performance?

RQ2. Which features drive the predictions of the strongest model and do they raise concerns for financial inclusion?

RQ3. Are the model's predictive errors distributed equitably across demographic groups defined by sex, education and age?

This paper's main contribution is not any particular technique, each of which is proven, but rather their integration into a single, repeatable evaluation framework that links mitigation, fairness, accuracy and explanation to a single choice and reveals what each phase would overlook on its own. This paper makes four contributions. First, it provides a reproducible comparison of logistic regression, random forest and gradient-boosted models on a public benchmark, situating their accuracy relative to the literature. Second, it conducts a careful demographic fairness audit across three protected attributes, showing that disparities are real but moderate and that the careless inclusion of a small subgroup can substantially overstate them. Third, it tests the common intuition that removing an access-correlated feature reduces bias and finds that it does not direct evidence that fairness through unawareness is insufficient on this data. Fourth, it demonstrates a simple, effective mitigation (group-aware thresholds) and synthesizes the analyses into a layered evaluation framework that treats explainability, auditing and mitigation as prerequisites, rather than afterthoughts, for trustworthy credit AI.

Literature Review

Machine Learning for Credit Scoring

Statistical classification has underpinned consumer credit scoring for decades, beginning with linear and logistic scorecards valued for their transparency and regulatory acceptability (Hand & Henley, 1997). Yeh and Lien (2009), who tested several data-mining methods for default prediction and established baseline performance that later research has reviewed, introduced the benchmark dataset used here. The advent of ensemble methods and in particular gradient-boosted decision trees (Chen & Guestrin, 2016), improved predictive accuracy substantially but at the cost of interpretability, motivating the explainability and fairness literatures discussed below.

Explainable Machine Learning

As high-stakes decisions increasingly rely on opaque models, explainable AI (XAI) has emerged to render their behavior intelligible (Arrieta et al., 2020). Among post-hoc methods, SHapley Additive exPlanations (Lundberg & Lee, 2017) provide a theoretically grounded, locally accurate attribution of a prediction to its input features and have become a de facto standard for model interpretation. Practitioners caution, however, that post-hoc explanations can be incomplete or misleading and some argue that interpretable models should be preferred outright for consequential decisions (Rudin, 2019). Explanations are frequently used internally

for debugging rather than to tell impacted individuals, according to studies of real-world deployment (Bhatt et al., 2020), highlighting the fact that explainability is essential but insufficient for confidence.

Algorithmic Fairness

What constitutes a fair automated choice is formalized in a parallel body of literature. While (Hardt et al. 2016) presented equalized odds, which necessitates equivalent true-positive and false-positive rates across protected groups, earlier work offered individual-level fairness through similarity constraints (Dwork et al., 2012). A comprehensive assessment by (Mehrabi et al. 2021) lists the sources of bias that emerge across the ML pipeline, while (Barocas and Selbst 2016) place these ideas within anti-discrimination law and the doctrine of disproportionate impact. A recurring lesson is that demographic parity and error-rate criteria can clash and no single statistic fully represents justice.

FinTech and Financial Inclusion

With demonstrated effects on stability and inclusion (Ozili, 2018) and an increasing role for fintech intermediaries in reaching underrepresented borrowers (Philippon, 2019), digital finance has increased access to formal financial services. (Berg et al. 2020) show that even simple digital-footprint variables can rival bureau scores, suggesting new pathways to credit for thin-file applicants. (Yet Fuster et al. 2022) demonstrate, using United States mortgage data, that the move from traditional models to machine learning can widen disparities, leaving some minority borrowers less likely to benefit. This study fills this gap by combining accuracy, explainability and a fairness audit on a single transparent benchmark. When taken as a whole, this body of work demonstrates that the inclusion effects of credit ML are ambiguous and must be measured rather than assumed.

Recent Work and Positioning of This Study

The current study is best appreciated considering the rapidly expanding body of work that explicitly incorporates explainability and fairness into credit rating. Japinye and Adedugbe (2025) create an explainability-first system that combines multi-criteria fairness monitoring across many public loan datasets, including the benchmark employed here, with calibrated gradient-boosting models with SHAP and LIME explanations. Similarly, Akavaram (2025) assesses bias mitigation at the pre-processing, in-processing and post-processing stages and incorporates explanations based on SHAP and LIME with demographic-parity, equal-opportunity

and equalized-odds measures. In particular, fairness in machine-learning credit lending is examined by Babaei and Giudici (2024). In the context of this, the study's contribution is purposefully limited and applied: instead of suggesting new explainability or mitigation mechanisms, it employs a transparent, repeatable audit to make a single argument concrete: that group-level error disparities can be concealed by aggregate accuracy, that prior access inequality can be encoded by reliance on the assigned credit limit and that these failures primarily affect the populations that financial inclusion initiatives seek to serve. Rather than being a methodological advancement, the audit and the layered evaluation system are presented as a transparent, teachable example of reliable lending practices. The end-to-end integration of auditing, mechanism testing through ablation, robustness across thresholds and mitigation applied transparently to a single benchmark, along with an explicit demonstration that both careless auditing and naive feature removal can mislead, is what sets this study apart from that work rather than a novel approach.

Methodology

Dataset

We use the UCI Default of Credit Card Clients dataset, comprising 30,000 credit-card holders in Taiwan observed from April to September 2005. Each record contains 23 explanatory variables: a credit limit (LIMIT_BAL), demographic attributes (sex, education, marital status, age), six months of repayment status (PAY_1 through PAY_6), six monthly bill amounts (BILL_AMT1–BILL_AMT6) and six-monthly payment amounts (PAY_AMT1–PAY_AMT6). The binary target indicates whether the client defaulted on the following month's payment. The dataset is moderately imbalanced, with a default rate of 22.1%.

Preprocessing

Following common practice for this dataset, undocumented categorical codes were consolidated: education values of 0, 5 and 6 were mapped to an “other” category and marital-status code 0 was merged into “other.” No records were dropped. To maintain the default rate, the data were divided into 80% training and 20% test sets (24,000 and 6,000 records) using stratified sampling. Inverse-frequency weighting (a positive-class weight for the boosted model and balanced class weights for the linear and random-forest models) was used to alleviate class imbalance during training.

Models

Three classifiers covering the range of interpretability and flexibility were trained. Fitted on standardized features, logistic regression provides a clear baseline. A bagged ensemble is represented as a random forest with 400 trees and a maximum depth of 8. The most popular boosted ensemble in modern credit modeling is an XGBoost gradient-boosted model (400 trees, maximum depth 4, learning rate 0.05, subsample and column-sample fractions of 0.8). Instead of maximizing any one score, hyperparameters were kept fixed at standard levels to maintain the comparison's transparency and reproducibility.

Evaluation Metrics

Predictive performance is reported using the area under the ROC curve (ROC-AUC), the area under the precision–recall curve (PR-AUC), the Kolmogorov–Smirnov statistic (KS) favored in credit practice and accuracy and F1 at a 0.50 decision threshold. ROC-AUC and PR-AUC are threshold-independent and therefore primary; accuracy and F1 contextualize behavior at the operating threshold used for the fairness audit.

Explainability

We use SHAP's precise tree estimator to interpret the strongest model. The distribution of per-instance attributions is analyzed to describe the direction and strength of each feature's influence and global feature relevance is summed up as the mean absolute SHAP value throughout the test set.

Fairness Audit

We examine the best model for three protected attributes: age (binned into five bands), sex and education. We calculate the flagged rate (the percentage expected to default), the true-positive rate (TPR) and the false-positive rate (FPR) for each group at the 0.50 level. Two disparity measurements are derived from these. The difference between the maximum and minimum highlighted rates across groups is known as the demographic-parity difference. The maximum minus minimum TPR (the uneven advantage of accurate risk detection) and the maximum minus minimum FPR (the unequal burden of being incorrectly detected) are equalized-odds gaps (Hardt et al., 2016). Greater disparities suggest more unfair treatment.

Robustness, Ablation and Mitigation

Three further analyses probe the audit. First, a threshold sweep recomputes the education disparity as the decision threshold varies from 0.15 to 0.85, testing whether the gaps depend on the 0.50 operating point. Second, an ablation retraines the model without the assigned credit limit to test whether that single access-correlated feature drives the disparity. Third, a post-processing mitigation assigns each education group a decision threshold chosen so that its false-positive rate matches the overall rate and we report the resulting change in the disparity gap and in overall accuracy.

Results

Predictive Performance

Table 1 reports test-set performance. The two ensembles are essentially tied and clearly outperform logistic regression; XGBoost attains the highest ROC-AUC (0.775) and PR-AUC (0.554). These values are consistent with previously published results on this benchmark, indicating that the model is representative rather than unusually strong or weak. Because the ensembles do not differ meaningfully, model selection on accuracy alone is largely indeterminate and motivating the explainability and fairness analyses that follow.

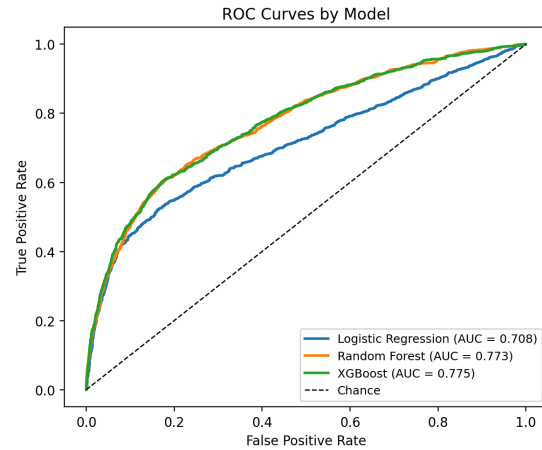
Table 1

Test-set predictive performance of the three classifiers ($n = 6,000$).

Model	ROC-AUC	PR-AUC	KS	Accuracy	F1
Logistic Regression	0.708	0.491	0.359	0.680	0.461
Random Forest	0.774	0.550	0.428	0.783	0.543
XGBoost	0.775	0.554	0.431	0.763	0.537

Figure 1

ROC curves for the three classifiers; the two ensembles nearly overlap and dominate logistic regression.



Feature Attributions

Figure 2 ranks feature by mean absolute SHAP value for the XGBoost model. The most recent repayment status (PAY_1) dominates by a wide margin, followed by the assigned credit limit (LIMIT_BAL) and the most recent bill amount. Figure 3 shows the corresponding distribution of attributions, confirming that recent delinquency pushes predictions toward default while higher credit limits push them away. The prominence of LIMIT_BAL is notable: because the assigned credit limit is itself an outcome of prior credit access, a model that leans heavily on it risks encoding past access inequalities into present-day decisions which is an inclusion concern examined in the Discussion.

Figure 2

Global feature importance (mean $|SHAP\ value|$) for the XGBoost model.

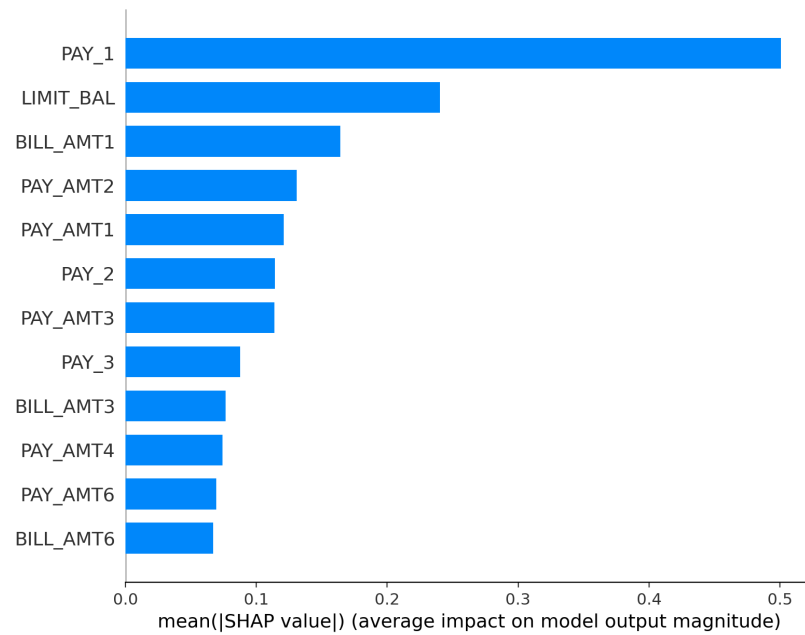
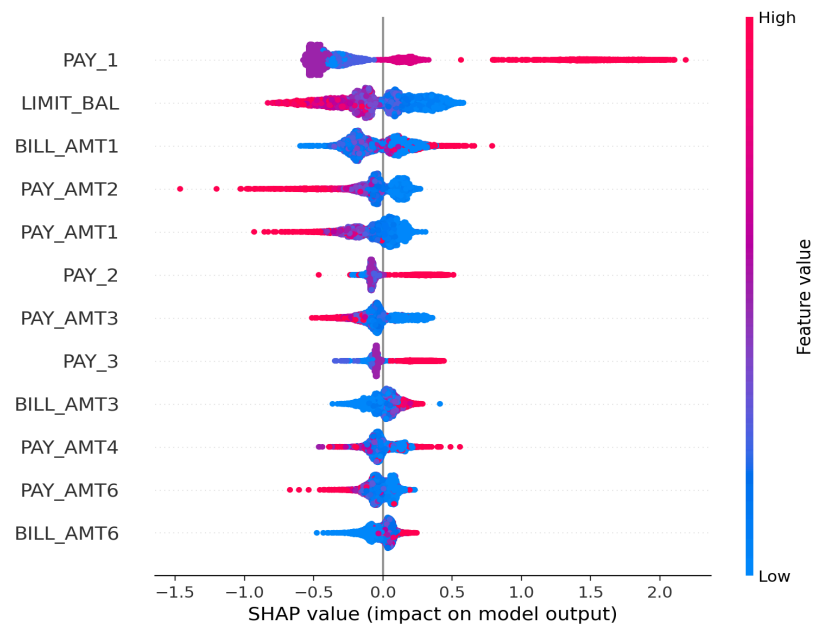


Figure 3

SHAP summary (beeswarm) showing the direction and magnitude of feature effects.



Fairness Audit

Tables 2 through 4 report group-level error rates and disparity measures; all disparity gaps are computed over groups of reliable size, excluding the small education “other” category ($n = 82$), whose rates are too noisy to interpret. Across sex (Table 2), disparities are modest, with a false-positive gap of 0.053. Across education (Table 3, three substantive groups), the demographic-parity, true-positive and false-positive gaps are 0.091, 0.079 and 0.082, respectively; high-school-educated applicants are wrongly flagged more often than graduate-educated applicants (FPR 0.243 vs. 0.161). Across age (Table 4), a clear gradient appears: the youngest borrowers (under 25) face the highest flagged rate (0.381) and the highest false-positive rate (0.265), giving the largest reliable false-positive gap, 0.098. It is worth noting that computing the education gaps over all four groups inflates them two- to fivefold (for example, a false-positive gap of 0.204 rather than 0.082) purely because of the 82-record subgroup and a concrete illustration of how naive auditing can mislead. Figure 4 visualizes the group error rates.

Table 2

Fairness audit by sex (XGBoost, 0.50 threshold).

Group	n	Flagged	TPR	FPR	Default
Male	2,402	0.323	0.635	0.229	0.234
Female	3,598	0.269	0.612	0.176	0.213

Demographic-parity difference 0.054; equalized-odds TPR gap 0.023; FPR gap 0.053.

Table 3*Fairness audit by education (XGBoost, 0.50 threshold).*

Group	n	Flagged	TPR	FPR	Default
Graduate school	2,130	0.241	0.583	0.161	0.190
University	2,774	0.321	0.662	0.215	0.236
High school	1,014	0.332	0.590	0.243	0.257
Other	82	0.049	0.200	0.039	0.061

Gaps over the three substantive groups: demographic-parity 0.091; equalized-odds TPR gap 0.079; FPR gap 0.082. The “other” group (n = 82) is shown for completeness but excluded from gap computation.

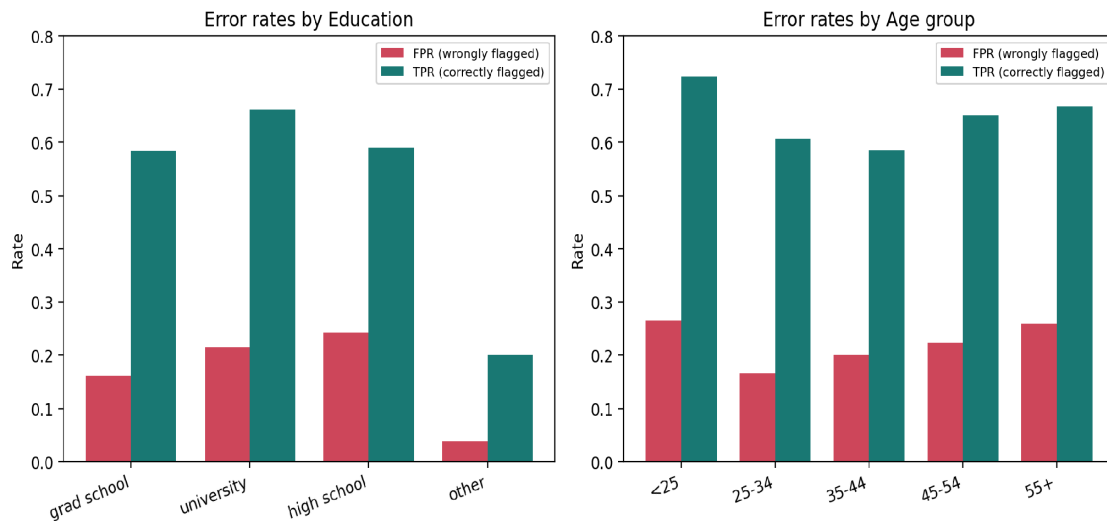
Table 4*Fairness audit by age band (XGBoost, 0.50 threshold).*

Group	n	Flagged	TPR	FPR	Default
Under 25	541	0.381	0.723	0.265	0.253
25–34	2,518	0.255	0.607	0.167	0.201
35–44	1,855	0.287	0.585	0.200	0.225
45–54	859	0.329	0.651	0.224	0.247
55 and over	227	0.357	0.667	0.260	0.238

Demographic-parity difference 0.126; equalized-odds TPR gap 0.138; FPR gap 0.098.

Figure 4

False-positive and true-positive rates by education and age group.



Larger gaps in bar height across groups indicate larger equalized-odds violations; the false-positive (red) bars show the unequal burden of being wrongly flagged. The small “other” education group is shown but excluded from gap computation.

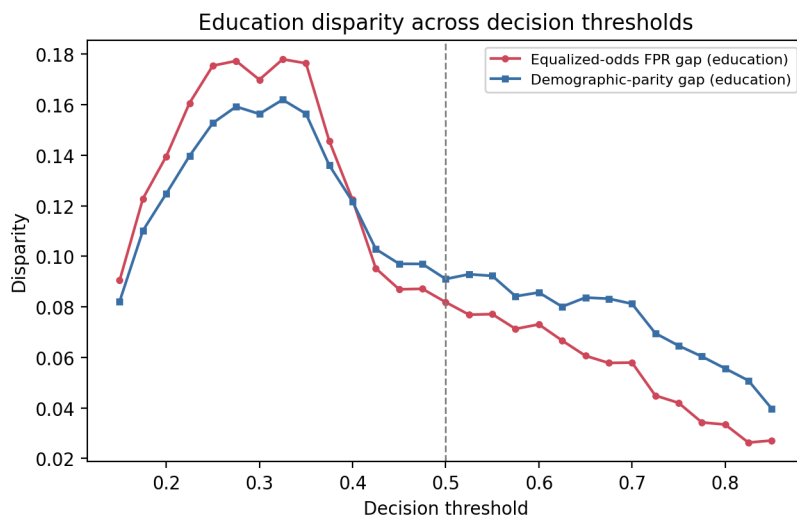
Robustness, Mechanism and Mitigation

Threshold Robustness

Because the audit above fixes the decision threshold at 0.50, we test whether the disparities are an artifact of that choice. Figure 5 sweeps the threshold from 0.15 to 0.85 and recomputes the education disparity at each point. The false-positive gap is not eliminated at any operating point and it ranges from roughly 0.03 to 0.18 and remains near 0.08 at the conventional 0.50 threshold; so, the disparity is a property of the model rather than of a particular cutoff.

Figure 5

Education disparity across decision thresholds; the dashed line marks the 0.50 threshold used elsewhere.



The Role of the Assigned Credit Limit

SHAP ranked the assigned credit limit (LIMIT_BAL) among the most influential features, raising the concern that the model predicts creditworthiness partly from prior credit access. Table 5 shows that the limit is indeed strongly stratified by education and age. To test whether it causes disparity, we retrained the model without it (Table 6). Removing the limit leaves both accuracy and the education false-positive gap essentially unchanged, indicating that the disparity is redundantly encoded in correlated behavioral features rather than driven by this one variable.

Table 5

Mean assigned credit limit (NT\$) by group, showing stratification by protected attributes.

Dimension	Group	Mean limit (NT\$)
Education	Graduate school	215,085
	University	145,753
	High school	128,795
Age	Under 25	64,177
	35–44	191,303
	55 and over	154,802

Table 6

Ablation test: removing LIMIT_BAL changes neither accuracy nor the disparity.

Model	ROC-AUC	Education FPR gap
Full (with LIMIT_BAL)	0.775	0.082
Ablated (no LIMIT_BAL)	0.771	0.078

Mitigating Disparity with Group-Aware Thresholds

Because removing the proxy does not help, we intervene at the decision stage instead. We select an education-group-specific threshold so that each group’s false-positive rate matches the overall rate, a standard post-processing approach to equalized odds. Table 7 reports the effect: the education false-positive gap falls from 0.082 to 0.005, while overall accuracy declines only from 0.763 to 0.759 and F1 from 0.537 to 0.535. Figure 6 shows the per-group false-positive rates converging after mitigation. The disparity is therefore largely correctable at negligible cost.

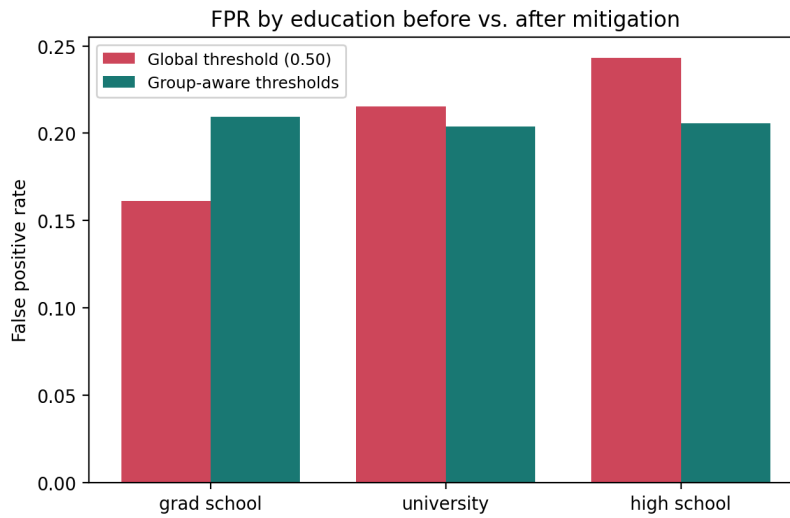
Table 7

Effect of group-aware thresholds: the disparity nearly vanishes at minimal accuracy cost.

Policy	Education FPR gap	Accuracy	F1
Global threshold (0.50)	0.082	0.763	0.537
Group-aware thresholds	0.005	0.759	0.535

Figure 6

Education false-positive rates before (global threshold) and after (group-aware thresholds) mitigation; the rates converge.



Discussion

The main conclusion is that trustworthiness and accuracy are not the same. The fairness audit reveals that the deployed model treats demographic groups differently, particularly across age and education, even though the two ensemble models are statistically identical on all accuracy metrics. None of this would be seen to a lender using a model based solely on ROC-AUC. And disparate error rates at the group level were concealed by predictive parity at the aggregate level, demonstrating that accuracy is a required but insufficient foundation for confidence.

The explainability and ablation analyses together refine the inclusion concern. The assigned credit limit ranks among the most influential features and is strongly stratified by protected attributes: graduate-educated clients carry roughly NT\$215,000 in limit versus NT\$129,000 for high-school-educated clients and applicants under 25 average about NT\$64,000 versus NT\$191,000 in mid-career. In that sense the limit plainly encodes prior, unequally distributed access to credit. However, excluding it from the model little affects the education false-positive gap (0.082 to 0.078) or accuracy (ROC-AUC 0.775 to 0.771). As a result, the discrepancy is redundantly transmitted across associated behavioral characteristics like repayment history rather than being conveyed by a single feature. This is the well-known lesson that bias cannot be eliminated by justice via ignorance, just removing a sensitive proxy and thus encourages intervention at the choice stage as opposed to the feature level.

Despite being moderate, the differences in error rates are nonetheless significant. In practice, a false positive that being mistakenly identified as a possible defaulter; translates into lower limits, higher prices, or credit denial. The audit reveals that among the groups targeted by financial inclusion programs, younger and less educated applicants bear a disproportionate amount of this burden. Therefore, even when the differences are not very large, a model that is accurate overall may still treat groups differently in distribution.

These results imply that explainability, careful fairness auditing and mitigation should be standard, reported components of any credit model intended for inclusive lending. We propose a layered evaluation framework in which a model advances through five gates before deployment: (1) predictive adequacy, established with threshold-independent metrics; (2) explanation, in which the drivers of predictions are inspected for legitimacy and proxy risk; (3) fairness, in which disparity gaps are measured across protected attributes of reliable size; (4) mitigation, in which residual disparities are reduced—here, group-aware thresholds cut the education false-positive gap from 0.082 to 0.005 for 0.4 points of accuracy; and (5) governance, in which any remaining disparities are documented, justified and monitored over time. Accuracy clears only the first gate.

A warning of the mitigation is crucial. Because group-aware thresholds apply distinct decision rules to each educational group, the protected property directly influences the decision, which minimizes discrepancy. This is ethically and legally questionable in many lending contexts: anti-discrimination laws like the U.S. Equal Credit Opportunity Act typically forbid

using protected or proxy attributes as the foundation for an individual credit decision, so an intervention that seems reasonable based on one statistical metric may be illegal disparate treatment. We therefore present group-aware thresholds as a diagnostic that quantifies how much of the disparity is correctable at the decision stage, not as a deployment recommendation. Group-neutral options, such as a single threshold chosen to balance error rates, reject-option or randomized post-processing that is independent of group membership, or in-processing techniques that punish disparity during training, are preferred when disparity needs to be minimized. Group-specific regulations would need to be implemented with clear legal and regulatory justification that is relevant to the jurisdiction.

Implications, Limitations and Future Research

The findings carry a direct implication for practice: institutions that deploy credit models for inclusive lending should treat explanation, fairness auditing, and mitigation as standard reporting requirements alongside accuracy, and should measure disparity only over demographic groups of reliable size.

Several limitations qualify these findings. The analysis uses a single dataset from one market and period (Taiwan, 2005), so the specific magnitudes may not generalize. Most importantly, the sample is made up of current credit card holders rather than thin-file or unbanked applicants. Therefore, the study shows a mechanism that is pertinent to financial inclusion: how disparate error rates and access-correlated features can disadvantage marginalized groups. However, it does not directly demonstrate how these models behave on the underprivileged populations that inclusion efforts aim to reach. Rather than being scientific results regarding the unbanked, claims about inclusion should be interpreted as inspiring and insightful. The protected attributes available are coarse and the “other” education group is too small for reliable inference. The fairness metrics are computed at a single 0.50 threshold; a full treatment would sweep thresholds and consider intersectional groups. We measure disparities but do not apply mitigation methods and we make no causal claims. Each of these points marks a direction for future work rather than a flaw in the central argument, which concerns the evaluation methodology itself.

Conclusion

Predictive accuracy is not a sufficient foundation for trust in algorithmic lending, according to this study. A representative gradient-boosted model matched published accuracy on

a conventional credit-default benchmark but produced uneven mistake rates across age and education which disparities that simple auditing overstated and detailed auditing revealed to be real but low. While group-aware decision criteria almost completely removed the education gap at a minimal accuracy cost on this benchmark, eliminating the most access-correlated feature did not assist; nevertheless, as mentioned, that fix is illustrative rather than deployable. Trust, we argue, must be engineered through explanation, careful auditing and mitigation, not inferred from accuracy. The layered framework proposed here offers lenders and regulators a concrete path toward credit AI that advances, rather than quietly erodes, financial inclusion. Future research will assess other data signals and richer, intersectional audits, as well as expand the analysis to various datasets and emerging-market, thin-file situations where financial inclusion is particularly important.

References

- Akavaram, S. (2025). Responsible AI in financial decision systems: Fairness and explainability in credit scoring models. *Journal of International Crisis and Risk Communication Research*, 571–580. <https://doi.org/10.63278/jicrcr.vi.3624>.
- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Babaei, G., & Giudici, P. (2024). How fair is machine learning in credit lending? Quality and Reliability Engineering International, 40(6), 3452–3464. <https://doi.org/10.1002/qre.3579>
- Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104(3), 671–732. <https://doi.org/10.15779/Z38BG31>
- Berg, T., Burg, V., Gombović, A., & Puri, M. (2020). On the rise of FinTechs: Credit scoring using digital footprints. *The Review of Financial Studies*, 33(7), 2845–2897. <https://doi.org/10.1093/rfs/hhz099>
- Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M. F., & Eckersley, P. (2020). Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 648–657). Association for Computing Machinery. <https://doi.org/10.1145/3351095.3375624>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. <https://doi.org/10.1145/2939672.2939785>
- 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* (pp. 214–226). Association for Computing Machinery. <https://doi.org/10.1145/2090236.2090255>

- Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. *The Journal of Finance*, 77(1), 5–47. <https://doi.org/10.1111/jofi.13090>
- Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A*, 160(3), 523–541.
- Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems* (Vol. 29, pp. 3315–3323).
- Japinye, A. O., & Adedugbe, A. A. (2025). Explainable AI for credit scoring with SHAP-calibrated ensembles: A multi-market evaluation on public lending data. *SSR Journal of Artificial Intelligence*, 2(3), 5–24. <https://doi.org/10.5281/zenodo.17155174>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765–4774).
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1–35. <https://doi.org/10.1145/3457607>
- Ozili, P. K. (2018). Impact of digital finance on financial inclusion and stability. *Borsa Istanbul Review*, 18(4), 329–340. <https://doi.org/10.1016/j.bir.2017.12.003>
- Philippon, T. (2019). On fintech and financial inclusion (NBER Working Paper No. 26330). National Bureau of Economic Research. <https://doi.org/10.3386/w26330>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- Yeh, I.-C., & Lien, C.-H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2), 2473–2480. <https://doi.org/10.1016/j.eswa.2007.12.020>

Appendix A: Reproducibility Details

Data: UCI Default of Credit Card Clients (30,000 records, 23 features, binary default target).

Split: 80/20 stratified (random seed 42). Models and key hyperparameters:

- Logistic Regression: standardized features, balanced class weights, max 2,000 iterations.
- Random Forest: 400 trees, max depth 8, balanced class weights.
- XGBoost: 400 trees, max depth 4, learning rate 0.05, subsample 0.8, colsample_bytree 0.8, positive-class weight set to the negative/positive ratio.

Explainability: SHAP TreeExplainer (exact). Fairness metrics computed at a 0.50 decision threshold. Education categories 0/5/6 mapped to “other”; marital-status 0 mapped to “other.”