Auditing Fairness and Explainability in AI-Assisted Loan-Restructuring Recommendations: A Governance Framework for Community Banks, Credit Unions, and CDFIs

Jensen Alexandru[1*], William Stan [2]
[1] University of Bucharest, ROMANIA
[2] Eötvös Loránd University, HUNGARY

ABSTRACT

AI-assisted engines that flag financially distressed small-business borrowers and recommend loan restructuring can help community lenders intervene earlier, but they also create new channels through which unequal treatment can arise. A recent multi-institution validation study demonstrated that such an engine can perform consistently across US community banks, credit unions and Community Development Financial Institution (CDFI) portfolios [1]. Building on that work, this article develops a fairness and explainability audit framework for restructuring-recommendation engines. Fairness risk is not confined to the risk score: it arises at five points, namely the data, the decision of who is flagged, the remedies that are offered, the remedies that officers grant, and the outcomes that follow. For each checkpoint we define audit metrics, discuss how group membership can be measured lawfully, and propose a three-tier explanation design serving validators, credit officers and borrowers. A synthetic demonstration with 20,000 simulated borrowers illustrates two failure modes an audit should detect. We then implement the engine's two stages on public repayment data for 24,000 borrowers and run the audit at checkpoints C0 to C2. Error-rate differences by sex and age band stayed within the example triggers (largest true-positive-rate gap 4.0 points; flag-rate ratios 0.82 and 0.81). However, 8 of 25 group-by-score-band calibration cells breached the 2-point trigger, and group-specific thresholds tuned on small validation groups overshot their targets on new data. Adding a proxy variable to the ranking rule reduced rate-reduction offers for every group rather than for one, which a group-gap test alone would miss. Reason codes were stable across refits (93 percent agreement). Checkpoints C3 and C4 need officer decisions and post-modification outcomes, which no public data provide. We compare mitigation options, set out a quarterly audit cycle, and map the framework to US supervisory and consumer-protection expectations.

Keywords: Algorithmic Fairness; Explainable AI; Loan Restructuring; Small-Business Lending; Community Banks; Credit Unions; CDFIs; Model Audit; Fair Lending; Model Risk Management

Introduction

Community lenders increasingly have access to machine-learning tools that promise earlier detection of borrower distress. A multi-institution validation study [1] has shown that an AI-assisted engine combining early warning with loan-restructuring recommendations can be validated across community banks, credit unions and CDFIs together. That evidence lowers one important barrier to adoption. It does not, however, settle whether the engine's recommendations treat different groups of borrowers fairly, or whether its reasoning can be explained to the officers who act on it and the borrowers who are affected by it.

These questions are pressing for two reasons. First, the institutions in question serve many minority-owned, women-owned and low-income-area businesses, and CDFIs exist specifically to serve borrowers that mainstream credit markets underserve. Second, a restructuring engine makes more consequential choices

than a conventional scorecard. It decides not only who receives attention but also what kind of relief is proposed, and the relief offered can materially change a firm's chances of survival. Evidence from credit markets shows that machine-learning models can widen disparities between groups even when they improve accuracy overall [2], and that algorithmic lending can reduce but not eliminate differences in pricing [3].

This article develops an audit framework for restructuring-recommendation engines of the kind validated in [1]. It makes four contributions:

- A map of five fairness checkpoints across the engine's pipeline, showing why an audit of the risk score alone is insufficient.
- A set of audit metrics and a discussion of how group membership can be measured within US fair-lending rules.
- A three-tier explanation design serving model validators, credit officers and borrowers.
- A synthetic demonstration of two common failure modes, an audit of an implemented engine on public credit data at checkpoints C0 to C2, a comparison of mitigation options, and a quarterly audit cycle with defined governance roles.

Background and Related Work

Fairness in Machine-Learning Credit Models

The algorithmic-fairness literature offers several formal criteria. Demographic parity compares the rate of a decision across groups; equality of opportunity compares true-positive rates among those who truly have the outcome of interest [4]; calibration requires that a given score mean the same probability in each group. Influential results show that when base rates differ, these criteria generally cannot all be satisfied at once [5][6]. Institutions must therefore choose which criteria matter most for a particular decision and justify that choice [7]. Surveys of the field catalogue the many sources of bias, from unrepresentative samples to proxies for protected characteristics [8]. In credit specifically, thinner and noisier data for some groups can by itself produce disparities, even when the model is otherwise unbiased [9].

Explainability

Post-hoc methods such as LIME and SHAP attribute individual predictions to input features [10][11]. Counterfactual explanations instead tell an affected person what would have had to be different for the decision to change [12], which is often more useful to a borrower than a list of feature weights. Critics argue that for high-stakes decisions, inherently interpretable models should be preferred over post-hoc explanations of opaque ones [13], and that explanation quality should be evaluated rigorously rather than assumed [14].

The Legal and Supervisory Setting

In the United States, the Equal Credit Opportunity Act prohibits discrimination in any aspect of a credit transaction, including the modification of existing credit [15]. The Consumer Financial Protection Bureau has made clear that creditors using complex algorithms must still provide specific and accurate reasons for adverse action [16], and its Section 1071 rule establishes the collection of demographic data for small-business credit applications [17]. Supervisory model risk guidance expects independent validation and effective challenge of models used in decision-making [18], while the NIST AI Risk Management Framework identifies fairness and explainability among the characteristics of trustworthy AI [19].

The Reference Engine

The engine validated in [1] unites distress detection and restructuring recommendation and was assessed across three institution types. Its cross-institution design is what makes a fairness audit both necessary and feasible: necessary, because the borrower populations of community banks, credit unions and CDFIs differ markedly; feasible, because a common engine allows common audit metrics to be applied and compared across institutions. The present framework is designed to sit alongside an implementation of that engine, and the Audit of an Implemented Engine section below applies it to one.

Where Fairness Risk Arises

Most fairness audits of credit models examine a single score. A restructuring engine, however, makes a chain of decisions, and unfairness can enter at any link. Figure 1 shows the five stages of the pipeline and the audit checkpoint attached to each, labelled C0 to C4.

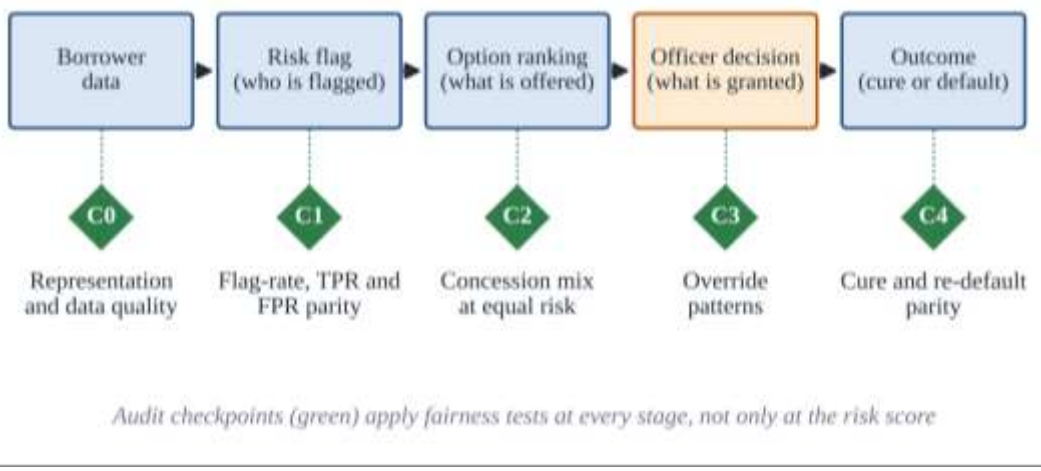

Figure 1. Fairness audit checkpoints across the restructuring-recommendation pipeline [authors' design].

Table 1 describes the characteristic risk at each checkpoint and an example of the harm it can cause. Two features of this pipeline deserve emphasis. First, being flagged is ambiguous: it can bring earlier help, but it can also bring closer scrutiny, tighter covenants or reduced credit lines. An audit must therefore examine both who is missed and who is flagged unnecessarily. Second, disparities in remedies can arise even when flagging is fair, because the option-ranking stage may draw on variables that correlate with group membership.

Table 1. Fairness risks and example harms at each audit checkpoint.

Checkpoint	Characteristic risk	Example harm	Primary test
C0 Data	Under-representation or thinner files for some groups [9]	Noisier risk estimates for those borrowers	Coverage and missing-data rates by group
C1 Risk flag	Unequal error rates between groups [4]	Distress missed for one group; unnecessary scrutiny for another	TPR and FPR by group
C2 Option ranking	Proxy variables steering remedies [8]	One group offered extensions while another is offered rate relief	Option mix by group at equal risk

Checkpoint	Characteristic risk	Example harm	Primary test
C3 Officer decision	Human overrides that reintroduce bias	Recommendations accepted for one group, overridden for another	Override rates and reasons by group
C4 Outcome	Remedies that work unequally	Higher re-default after modification for one group	Cure and re-default rates by group

Audit Metrics

Each checkpoint needs a small number of metrics with pre-set trigger levels. Table 2 proposes a core set. The equal-opportunity difference at checkpoint C1 is defined as follows, with an analogous difference for false-positive rates:

$$EOD = TPR(B) - TPR(A), \quad TPR = \text{flagged distressed} \div \text{all distressed}$$

Because flagging can be either beneficial or burdensome, we recommend reporting both the true-positive and false-positive differences rather than a single parity statistic. For the ratio of flag rates, institutions sometimes borrow the four-fifths heuristic from employment-selection guidance [20]; we treat it only as a screening signal that prompts investigation, not as a legal test for credit decisions. Trigger levels in Table 2 are examples that each institution should set and document in advance.

Table 2. Core audit metrics, checkpoints and example trigger levels.

Metric	Definition	Checkpoint	Example trigger
Coverage gap	Difference in the share of borrowers with complete core data	C0	More than 5 percentage points
Equal-opportunity difference	Difference in true-positive rates between groups [4]	C1	More than 5 percentage points
False-positive difference	Difference in false-positive (unnecessary flag) rates	C1	More than 3 percentage points
Flag-rate ratio	Lower group's flag rate divided by higher group's rate [20]	C1	Below 0.8 (screening only)
Within-group calibration	Observed distress rate per score band, by group	C1	Gap above 2 points in any band
Concession-mix difference	Difference in each option's share, stratified by risk band	C2	More than 10 percentage points
Override disparity	Difference in override rates and reason categories	C3	More than 10 percentage points
Re-default difference	Difference in 12-month re-default after modification	C4	More than 5 percentage points

Measuring Group Membership

A fairness audit requires knowledge of group membership, which lenders often do not hold. US law generally restricts the collection of protected characteristics for non-mortgage credit, but the Section 1071 rule creates an exception for covered small-business applications [17]. Where self-reported data are unavailable,

statistical proxies such as Bayesian Improved Surname Geocoding, which combines surname and geography, are widely used to estimate disparities in aggregate [21]. Table 3 compares the main options.

Table 3. Options for measuring group membership in fairness audits.

Source	Strengths	Limitations
Section 1071 self-reported data [17]	Direct and legally sanctioned for covered applications	Covers new applications, not the existing loan book; responses may be missing
Surname and geography proxy [21]	Available for the existing portfolio; established in aggregate analysis	Probabilistic; not suitable for individual decisions; weaker for business owners
Census-tract income and minority share	Simple; aligns with community-development mandates	Measures places, not people; risks ecological error
Ownership attributes in loan files	Already held for many CDFI and certified-business programs	Inconsistent recording; selective coverage

Two safeguards apply whatever source is used. Group membership should be held in a separate, access-controlled dataset used only for auditing, never as a model input. And results based on proxies should be reported as estimates with uncertainty, so that a disparity near a trigger level prompts closer review rather than an automatic conclusion.

A Three-Tier Explanation Design

An explanation is only useful if it fits its reader. Figure 2 sets out a three-tier design. Validators and examiners need global evidence of how the model behaves and how stable it is. Credit officers need borrower-level reason codes, the ranked options with projected repayment capacity, and a counterfactual statement of what would change the recommendation [12]. Borrowers need plain-language principal reasons and concrete steps they can take, which is also what adverse-action rules require when a restructuring request is denied or granted on less favorable terms than requested [16].

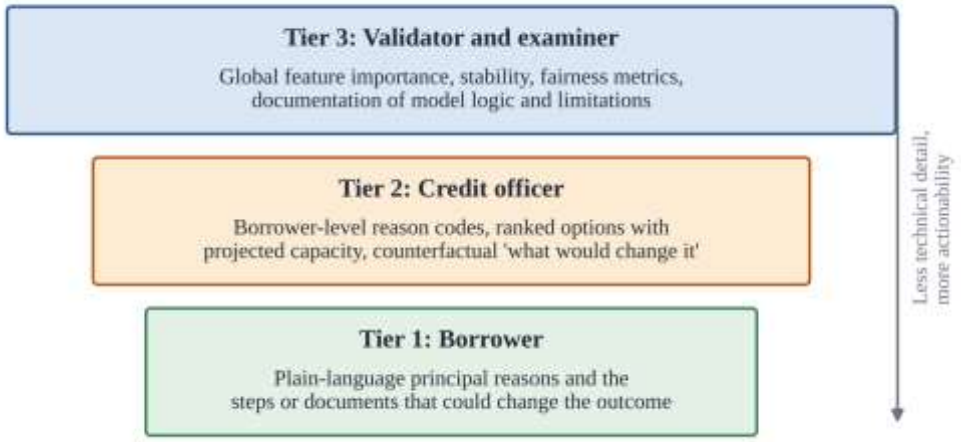

Figure 2. Three-tier explanation design for restructuring recommendations.

Explanations should themselves be audited. A practical check is whether reason codes differ systematically between groups that receive the same recommendation; if one group is told more often that its file is incomplete, the data gap identified at checkpoint C0 may be surfacing in borrower communications.

Synthetic Demonstration

To show how the audit metrics behave, we simulated a portfolio of 20,000 small-business borrowers. The simulation is purely synthetic: it uses synthetic data, does not reproduce the engine in [1], and its figures are not estimating of any real portfolio. Table 4 summarizes its design. Group B, 30 percent of borrowers, was given the same true distress rate as Group A but noisier data, representing thinner credit files. A geographic proxy correlated with group membership was included in the option-ranking rule to represent an inadvertent steering variable.

Table 4. Design of the synthetic demonstration (simulated data).

Element	Specification
Portfolio	20,000 simulated borrowers; Group A 70 percent, Group B 30 percent
True distress	Logistic function of latent financial health; about 17 percent in both groups
Risk score	Latent health plus noise; noise standard deviation 0.55 for Group A and 1.10 for Group B
Flag rule	Top 12 percent of scores portfolio-wide; mitigation uses group-specific thresholds equalizing true-positive rates [4]
Option rule	Rate reduction when a payment-burden signal is high; otherwise term extension or deferral; geographic proxy optionally added
Software	Python with a fixed random seed for reproducibility

Figure 3 shows the first failure mode. With a single portfolio-wide threshold, Group B's noisier scores spread its borrowers toward both tails, so it was flagged more often: 16.3 percent against 10.2 percent for Group A. Its detection rate was slightly higher, but its false-flag rate was 11.6 percent against 5.1 percent, meaning that non-distressed Group B borrowers were more than twice as likely to face unnecessary scrutiny. Group-specific thresholds chosen to equalize true-positive rates at 35.7 percent narrowed the false-flag gap to 10.0 against 5.6 percent but did not close it.

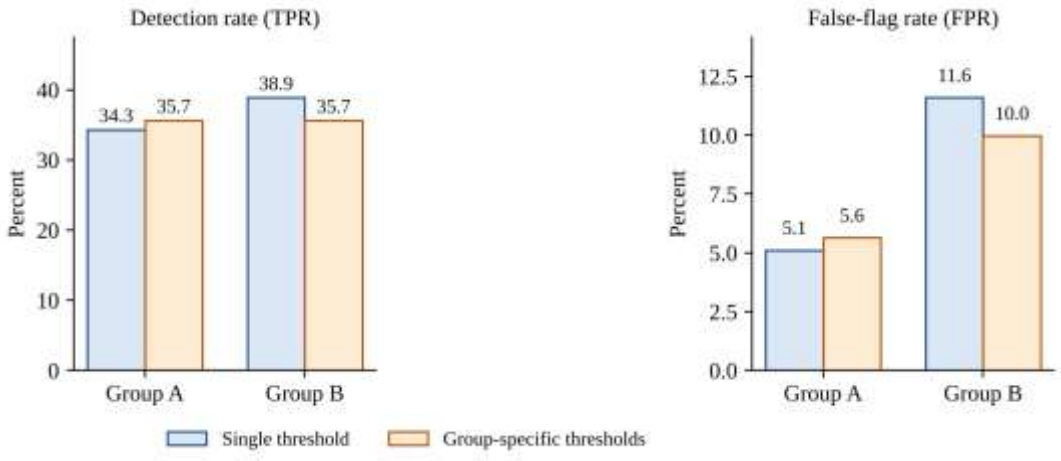

Figure 3. Detection and false-flag rates by group before and after threshold adjustment (synthetic data).

This result illustrates the incompatibility findings in practice [5][6]: equalizing one error rate does not equalize the other when one group's data are noisier. The more durable remedy is to improve data quality for the thinner-file group, consistent with evidence that noise itself drives credit disparities [9]. Table 5 reports the flag-rate figures that accompany Figure 3.

Table 5. Flag rates and flag-rate ratio in the synthetic demonstration.

Measure	Group A	Group B	Ratio A ÷ B
Simulated borrowers	14,034	5,966	–
True distress rate	17.4%	17.1%	–
Flag rate, single threshold	10.2%	16.3%	0.63
Flag rate, group-specific thresholds	10.9%	14.4%	0.76

Figure 4 shows the second failure mode, at checkpoint C2. With the geographic proxy in the option-ranking rule, 35 percent of flagged Group A borrowers were recommended a rate reduction against 11 percent of Group B borrowers, who were steered toward term extensions and deferrals. Removing the proxy brought the shares to 27 and 28 percent. A flag-level audit alone would not have detected this disparity, which is why the concession-mix test belongs in every audit.

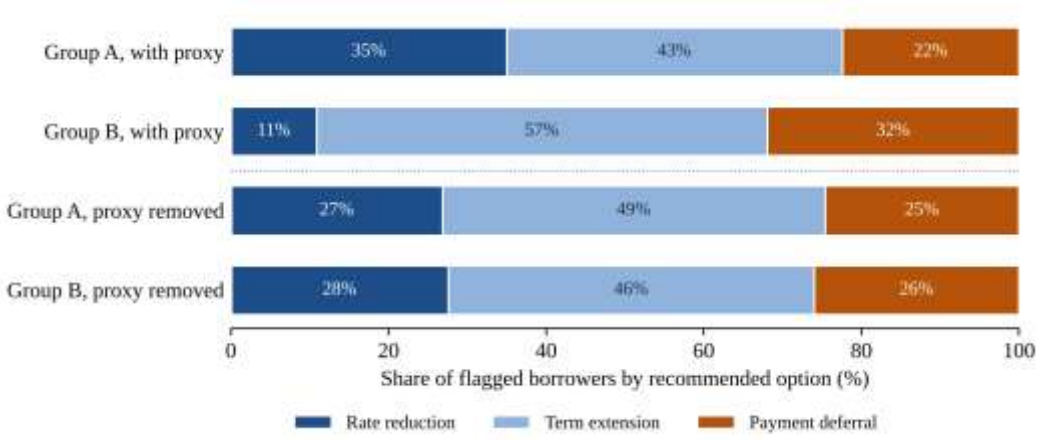

Figure 4. Recommended restructuring options by group with and without a geographic proxy (synthetic data).

Audit of an Implemented Engine on Public Credit Data

The synthetic demonstration shows how the metrics behave by construction. To see how they behave on real borrowers, we implemented the two-stage logic of the reference engine and audited it. No public small-business dataset records repayment histories, group characteristics and restructuring outcomes together, so we used the Taiwan credit-card default data[24], which record sex and age for 24,000 borrowers along with six months of repayment status, balances and payments. These are consumer card accounts, and sex and age are the only group attributes available, so the audit tests the framework's procedures rather than producing estimates for any US small-business portfolio.

Stage one was a LightGBM distress model[25] trained on 56 percent of borrowers (14 percent for validation, 30 percent held out). It used repayment status, balances, payments, the credit limit and derived utilisation and arrears features, and excluded sex, age, marital status and education. On held-out borrowers it reached an AUC of 0.779 with an expected calibration error of 0.013. The watch list was the top 12 percent of scores.

Stage two ranked five restructuring options by expected value to the lender, using stated assumptions for option effects, with suitability adjusted by observable signals such as recent arrears and utilisation. Each flagged borrower received three reason codes derived from SHAP attributions[11].

Checkpoints C0 and C1

At checkpoint C0, data coverage was similar across groups: thin files (three or more months with no statement balance) ranged from 8.8 to 11.2 percent, within the 5-point coverage trigger. Table 6 reports the C1 metrics at the common threshold.

Table 6. Checkpoint C1 metrics for the implemented engine on held-out borrowers (common watch-list threshold, top 12% of scores).

Attribute	Group	Borrowers	Default rate	Flag rate	TPR	FPR	ECE
Sex	Female	4,354	20.9%	10.6%	0.346	0.043	0.016
Sex	Male	2,846	23.9%	12.9%	0.386	0.048	0.023
Age band	30-45	3,770	20.3%	10.5%	0.349	0.042	0.019
Age band	<30	2,285	23.1%	12.6%	0.383	0.048	0.010
Age band	>45	1,145	26.3%	12.9%	0.362	0.046	0.033

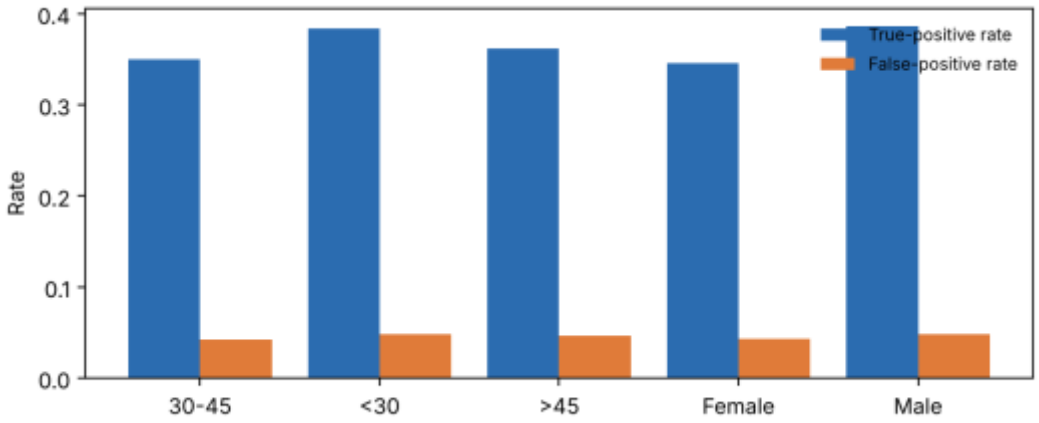

Figure 5. True-positive and false-positive rates of the implemented engine by group at the common threshold

Error-rate differences stayed within the example triggers in Table 2. The equal-opportunity difference between women and men was -4.0 points, and the largest true-positive gap between age bands was 3.4 points. False-positive differences were below 0.6 points, and flag-rate ratios were 0.82 by sex and 0.81 by age, just above the 0.8 screening level. Within-group calibration was the weaker point. Of the 25 group-by-score-quintile cells, 8 differed from the predicted rate by more than the 2-point trigger. The largest gap was +6.5 points for >45 borrowers in score quintile 5 (observed 61.9 percent, predicted 55.5 percent). Overall parity, in other words, coexisted with group-specific miscalibration in the highest-risk band, the band from which the watch list is drawn.

The mitigation in the synthetic demonstration, group-specific thresholds chosen to equalise true-positive rates, behaved differently on real data. Thresholds were set on the validation sample to bring each age band

to the portfolio's true-positive rate of 0.363; on held-out borrowers, the rate for borrowers over 45 rose to 0.429 and their false-positive rate to 0.068, up from 0.046. With only about 520 validation borrowers in that band, the tuned threshold was noisy, and the correction overshot. Group thresholds therefore need larger tuning samples or shrinkage toward the common threshold, a caution to add to the comparison in the Mitigation Options section.

Checkpoint C2 and the Proxy Test

A proxy screen asked how well the model's own inputs predict group membership: a classifier using only those inputs separated women from men with an AUC of 0.59 and borrowers under 30 from others with 0.64. Repayment and balance features therefore carry some group information even with protected attributes excluded. For checkpoint C2, the option mix was compared by group, first with the ranking rule as specified and then with a deliberately introduced proxy: graduate education, which lowered the suitability of rate reduction and raised that of term extension. Without the proxy, rate reduction was recommended to 19.7 percent of flagged women and 17.2 percent of flagged men. With the proxy, the shares fell to 13.4 and 12.0 percent, and for borrowers over 45 from 16.9 to 12.8 percent.

Unlike the synthetic geographic proxy, the education proxy reduced rate-reduction offers for every group by a similar amount, so the concession-mix difference test in Table 2 stayed well below its 10-point trigger. The lesson for audit design is that a between-group comparison can pass while a change to the ranking rule shifts every borrower toward a less generous remedy. The C2 audit should therefore also compare the concession mix before and after each rule change, not only across groups at a single point in time.

Explanations and Untested Checkpoints

For the three-tier explanation design, the principal reason code for a flagged borrower was recent repayment status in 84 percent of cases and the number of months in arrears in 15 percent. Across five bootstrap refits of the model, the principal reason code agreed with the original in 93 percent of cases on average (minimum 86 percent), which meets the stability that adverse-action notices require. Checkpoints C3 (officer overrides) and C4 (post-modification outcomes) could not be audited because no public dataset records officer decisions or what happens after a modification. They must be measured during the assisted-pilot stage at each institution.

Mitigation Options

When a trigger is breached, the audit team must diagnose the cause before choosing a remedy. Mitigations fall into four families, compared in Table 7. Pre-processing changes the training data, for example by reweighting under-represented groups [22]; in-processing adds fairness constraints to model training; post-processing adjusts decision thresholds [4]; and operational measures change data collection or human review. Inherently interpretable models may also reduce the need for some mitigations by making proxies easier to spot [13].

Table 7. Mitigation options and their trade-offs.

Family	Example	Advantages	Trade-offs
Pre-processing	Reweighting or resampling training data [22]	Model-agnostic; addresses representation	May not fix proxy effects
In-processing	Fairness constraint during training	Directly targets chosen metric	Requires retraining and revalidation

Family	Example	Advantages	Trade-offs
Post-processing	Group-specific thresholds [4]	Fast; no retraining	Uses group data at decision time; legal review needed
Feature review	Remove or constrain proxy variables	Addresses steering at C2	May reduce accuracy slightly
Operational	Targeted data collection; second review of overrides	Addresses root causes	Slower; adds staff cost

Post-processing deserves particular caution. Using group membership at the point of decision raises legal questions under fair-lending law, and institutions should obtain legal review before adopting it. In many community-lending settings, feature review and better data collection will be both more effective and easier to defend.

Governance and the Audit Cycle

Audits work only if they recur and someone acts on them. Figure 6 proposes a quarterly cycle: decisions and outcomes are extracted, metrics computed and compared with pre-set triggers, root causes investigated, and any change re-validated and documented before release. Each report goes to the board risk committee.



Figure 6. Quarterly fairness and explainability audit cycle

Responsibilities should follow the three-lines model familiar to supervisors. Table 8 assigns them. Where the engine is licensed from a vendor, contracts should give the institution access to the data and documentation needed to run these audits, consistent with third-party risk guidance [23].

Table 8. Governance roles in the audit framework.

Role	Responsibilities	Key outputs
First line: lending and credit	Records decisions and override reasons; acts on findings	Complete decision logs; remediation actions

Role	Responsibilities	Key outputs
Second line: risk and compliance	Runs quarterly audit; sets triggers; reviews explanations	Audit report; trigger breaches; recommendations
Model validation	Independent validation of the engine and of any mitigation [18]	Validation report; approval for release
Third line: internal audit	Tests that the audit process itself operates as designed	Assurance opinion
Board risk committee	Approves fairness criteria and trigger levels; reviews reports	Minutes and approvals

Conclusion

The multi-institution validation in [1] shows that an AI-assisted early-intervention and restructuring engine can work across community banks, credit unions and CDFIs. This article addresses the next requirement for responsible adoption: demonstrating that the engine's recommendations are fair and explainable. Fairness risk arises at five checkpoints, from data through outcomes, and the synthetic demonstration shows that disparities in false flags and in the remedies offered can persist even when overall detection looks balanced. Applied to an implemented engine on public credit data, the audit found error-rate parity within its triggers but miscalibration in the highest-risk band for some groups, threshold corrections that overshot on small samples, and a proxy that disadvantaged every group equally and so escaped a between-group test. Clear metrics, lawful measurement of group membership, tiered explanations, considered mitigation and a governed quarterly cycle together give community lenders a practical way to adopt validated restructuring engines while protecting the borrowers they exist to serve.

References

[1] Karim, N. B., & Elias, D. (2022). Multi-Institution Validation of an AI-Assisted Early-Intervention and Loan-Restructuring Recommendation Engine for Financially Distressed Small-Business Borrowers Across US Community Banks, Credit Unions, and CDFI Portfolios. *The Computertech*, 107-120.

[2] 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.

[3] Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. *Journal of Financial Economics, 143*(1), 30–56.

[4] 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).

[5] 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* (pp. 43:1–43:23). Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

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

[7] Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT Press.

[8] 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), Article 115.

[9] Blattner, L., & Nelson, S. (2021). *How costly is noise? Data and disparities in consumer credit* (arXiv preprint arXiv:2105.07554).

[10] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM.

[11] 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).

[12] Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 841–887.

[13] 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.

[14] Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning* (arXiv preprint arXiv:1702.08608).

[15] Equal Credit Opportunity Act, 15 U.S.C. § 1691 et seq. (1974).

[16] Consumer Financial Protection Bureau. (2022). *Consumer Financial Protection Circular 2022-03: Adverse action notification requirements in connection with credit decisions based on complex algorithms*.

[17] Consumer Financial Protection Bureau. (2023). Small business lending under the Equal Credit Opportunity Act (Regulation B). *Federal Register, 88*, 35150.

[18] Board of Governors of the Federal Reserve System, & Office of the Comptroller of the Currency. (2011). *Supervisory guidance on model risk management* (SR Letter 11-7; OCC Bulletin 2011-12).

[19] National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). U.S. Department of Commerce.

[20] Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). Uniform guidelines on employee selection procedures. *29 C.F.R. Part 1607*.

[21] Elliott, M. N., Morrison, P. A., Fremont, A., McCaffrey, D. F., Pantoja, P., & Lurie, N. (2009). Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology, 9*(2), 69–83.

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

[23] Board of Governors of the Federal Reserve System, Federal Deposit Insurance Corporation, & Office of the Comptroller of the Currency. (2023). Interagency guidance on third-party relationships: Risk management. *Federal Register, 88*, 37920.

[24] 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.

[25] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30* (pp. 3146-3154).