Journal of Management Science Research Review
https://jmsrr.com/index.php/Journal/about

Volume 5 Issue No 2 (2026)

Online ISSN: 3006-2047
Print ISSN: 3006-2039

Algorithmic Bias and Fairness in AI Credit Scoring: Evidence, Mechanisms, and Governance Responses

Rafia Noreen
Government College of Commerce for Women, Mardan, Pakistan
Email: rafia.noreen.phd@gmail.com

Muhammad Sajid
Department of Management Studies, University of Gujrat, Pakistan
Email: sajidali.cma@gmail.com

Faisal Amjad (Corresponding author)
Institute of Business Studies and Leadership, Abdul Wali Khan University Mardan (AWKUM), Pakistan Email:faisalamjad.ms@gmail.com

Abstract
Machine learning models have become widely adopted in credit scoring, but their deployment raises persistent concerns about discriminatory outcomes for protected demographic groups. This paper examines algorithmic bias and fairness in AI credit scoring, drawing on evidence from 15 empirical studies drawn from a systematically screened corpus of 55 peer-reviewed papers published between 2018 and 2025. The evidence shows that bias in AI credit scoring models is not primarily a technical failure, it originates from historically generated training data, is amplified through correlated proxy variables, and persists in part because institutional incentive structures do not penalise discriminatory outcomes absent regulatory compulsion. Gender and socioeconomic disparities are the most empirically documented, with digital credit markets in Kenya showing higher male participation rates and profit-optimised models disproportionately excluding marginal applicants. Technical mitigation strategies, causal inference corrections, SHAP-based feature audits, and discriminatory feature removal, show conditional effectiveness, but none resolves the structural misalignment between commercial model objectives and equitable lending. The paper argues that fairness governance requires both technical standards and institutional accountability mechanisms.

Keywords: Algorithmic Bias, Ai Credit Scoring, Fairness In Lending, Discriminatory Outcomes, Explainable Ai, Proxy Variables, Regulatory Compliance

JEL Codes: G21, G28, C45, J71, K23

Volume 5 Issue No 2 (2026)
Online ISSN: 3006-2047
Print ISSN: 3006-2039

Introduction

Credit scoring systems determine who gets access to capital and on what terms (Bartlett et al., 2022). That is not a technical observation, it carries direct consequences for household formation, enterprise development, educational attainment, and economic resilience across entire populations (Aggarwal, 2021). When AI and machine learning systems take over this function, they inherit not only the predictive advantages of computational sophistication but also the distributional implications of the data on which they are trained and the objectives they are asked to optimise (Brotcke, 2022).

The concern about algorithmic bias in credit scoring is not hypothetical. Several decades of documented discrimination in lending markets produced training datasets that encode systematic disparities along lines of race, gender, geography, and income (Bartlett et al., 2022). ML models, trained to predict default from historical outcomes, learn from that history (Brotcke, 2022). The question is not whether those patterns appear in the data, they do, but whether models amplify, reproduce, or under some conditions attenuate them, and what regulatory and institutional conditions affect that outcome (Hurlin et al., 2023).

The academic literature on this topic has grown considerably since roughly 2019, when the intersection of fair lending law and machine learning began attracting serious empirical attention (Das et al., 2023). Several distinct problems have emerged: proxy discrimination through correlated variables (Hurlin et al., 2023), the accuracy-fairness trade-off (Hurlin et al., 2023), the costs of counterfactual fairness corrections, and the practical limitations of explainability techniques as fairness tools (Das et al., 2023). These problems have different technical properties and require different governance responses (Aggarwal, 2021). Treating them as a single "bias problem" is one reason policy discussions have often produced vague recommendations.

This paper takes stock of the empirical evidence on algorithmic bias and fairness in AI credit scoring. It draws on 15 primary studies from a systematically screened corpus, each examined for the specific form of bias they address, the demographic groups affected, the datasets used, and the mitigation strategies evaluated. The goal is to move from a general concern about bias toward a more precise account of where the problems originate, how serious they are under what conditions, and what kinds of interventions actually help. The paper is organised as follows. Section 2 describes the methodology. Section 3 examines the sources and mechanisms of algorithmic bias. Section 4 analyses the evidence on specific demographic disparities. Section 5 addresses technical mitigation approaches. Section 6 considers the accuracy-fairness trade-off. Section 7 discusses governance implications. Section 8 concludes.

Methodology

This paper draws on 15 studies selected from a broader systematic review corpus of 55 papers screened from an initial pool of 500 using semantic search across the Elicit research platform (covering Semantic Scholar and OpenAlex, over 138 million papers). The parent review applied eight inclusion criteria spanning AI/ML

Volume 5 Issue No 2 (2026)
Online ISSN: 3006-2047
Print ISSN: 3006-2039

technology focus, credit scoring specificity, empirical evidence, real-world application focus, and relevant outcome reporting.

For this paper, the 15 studies were selected based on direct relevance to at least one of the following dimensions: (a) documented evidence of discriminatory or disparate outcomes by demographic group; (b) empirical analysis of bias sources including proxy variables, data imbalance, or model objective functions; (c) evaluation of technical fairness mitigation strategies; or (d) regulatory or governance analysis of fair lending in AI contexts. Studies reporting on explainability tools were included where those tools were applied specifically to address or detect bias, rather than solely for predictive interpretability.

All empirical claims in this paper are traceable to the 15 included studies or the wider 55-study corpus. No claim is introduced without attribution to a verifiable source within that corpus.

Sources and Mechanisms of Algorithmic Bias

Historically Encoded Training Data

The most fundamental source of algorithmic bias in credit scoring is the training data itself. Credit datasets reflect lending decisions made by institutions whose historical practices included explicit and implicit discrimination. When ML models are trained on these records, they learn statistical patterns that include discriminatory ones. Szepannek and Lübke (2021) make this point precisely: ML models "are not necessarily unbiased and may discriminate with respect to certain subpopulations such as a particular race, gender, or sexual orientation, even if the variable itself is not used for modeling." This applies not only to complex ensemble methods but also to transparent logistic regression scorecards, which suggests the problem is in the data rather than the model architecture alone.

Brotcke (2022), writing from a US regulatory compliance perspective, notes that ML models in credit scoring "can replicate historical biases and disadvantage certain populations, particularly in terms of racial, gender, and socioeconomic disparities." The difficulty is that bias encoded through correlated variables is structurally harder to detect than explicit use of prohibited attributes. A variable like zip code, payment timing, or device type may correlate strongly with race or income without any intent to discriminate, but its effect on model outputs is functionally similar to direct use of those attributes.

Proxy Variables and Indirect Discrimination

Proxy discrimination occurs when a model uses variables that are correlated with protected characteristics to produce outcomes that would be prohibited if the protected characteristic had been used directly. Hall et al. (2021), reviewing the US fair lending landscape, document this problem in detail. They note that "variables that serve as proxies for prohibited bases are difficult to identify in complex ML models" and that standard regulatory examination procedures, designed for simpler scorecards, frequently fail to detect this form of discrimination.

Volume 5 Issue No 2 (2026)
Online ISSN: 3006-2047
Print ISSN: 3006-2039

The difficulty is compounded by the dimensionality of modern ML models. An ensemble with hundreds of features can produce proxy effects through combinations of individually innocuous variables. No single feature may serve as an obvious proxy, but their interaction may encode the protected characteristic effectively. This makes post-hoc detection through variable inspection unreliable, and it means that removing a single suspect variable does not necessarily eliminate the proxy relationship, other variables adjust to compensate.

Xu et al. (2023), in their ethnographic study of credit scoring intermediaries in China, document how organisations developed multiple "renditions" of credit scoring models that incorporated behavioural and social data in ways that created new forms of stratification not anticipated by regulators. This is a practical illustration of how proxy relationships emerge from data architectures rather than from explicit discriminatory intent.

Model Objective Functions and Distributional Consequences

A less discussed but analytically important source of disparate outcomes is the objective function itself. Gramespacher and Posth (2021) show that models optimised for institutional profit, the standard objective in commercial credit scoring, systematically reject marginal applicants to maximise returns. Their empirical analysis finds that "surprisingly high rejection rates contribute to maximising profit," and that these rejections fall disproportionately on applicants whose risk profiles place them near the decision boundary, often low-income borrowers or those with thin credit files.

This is not bias in the conventional sense of treating similarly situated individuals differently based on protected characteristics. It is a form of distributional harm that follows from an objective function that does not weight access to credit as a social good. But its distributional consequences overlap substantially with protected demographic categories, because low income and thin credit files are themselves correlated with race, gender, and other protected attributes in markets with histories of financial exclusion.

The implication is that even a model that is technically "fair" by conventional metrics, equal error rates across demographic groups, for instance, can still produce systematically adverse outcomes for protected populations if the objective function does not account for the distributional costs of false rejections. This point is made implicitly by Johnen et al. (2021) in their Kenya analysis, where the digital lending ecosystem expanded access while simultaneously producing 90% of all credit bureau blacklistings, a distributional outcome that tracks systematically with poverty and informality rather than with protected demographic categories per se.

Documented Demographic Disparities

Gender Disparities

Gender disparities in AI credit scoring are the most consistently documented in the empirical literature. Johnen et al. (2021) find that males have higher digital credit access than females in Kenya, a pattern that partly reflects gender gaps in mobile

Volume 5 Issue No 2 (2026)
Online ISSN: 3006-2047
Print ISSN: 3006-2039

phone ownership but is nonetheless reproduced through algorithmic scoring based on mobile data. Laínez and Gardner (2023) identify specific concerns in Vietnam's consumer finance market about bias against women in algorithmic credit scoring, noting that the absence of regulatory oversight means these disparities go unaddressed. Hall et al. (2021) review the US regulatory framework and note that gender-based disparities in lending decisions have been documented both under traditional and algorithmic scoring regimes, and that switching to ML does not automatically resolve them. Nwafor et al. (2024) provide one of the most direct tests of gender bias in their hybrid 1DCNN-XGBoost model. They find that removing age and gender as features "does not significantly impact the hybrid model's classification capabilities", which is an important positive finding, demonstrating that high performance is achievable without reliance on these attributes. But it simultaneously confirms that gender was present in the training data and available to the model before that removal.

Green and Chen (2019), in their controlled experimental study using Lending Club data and Amazon Mechanical Turk participants, document racial bias in how human decision-makers interact with algorithmic risk assessments. While their study is primarily about human-algorithm interaction rather than model-level bias, the finding that "biased predictive errors were more likely to widen the perception gap" between algorithmic and human assessments has direct implications for hybrid decision systems where loan officers can override model recommendations.

Socioeconomic and Income-Based Disparities

Brotcke (2022) documents that socioeconomic disparities in credit scoring outcomes persist under ML models trained on conventional financial data, with low-income borrowers and those in economically deprived areas experiencing higher rejection rates independently of objective default risk. This overlaps substantially with racial and ethnic disparities in markets with strong correlations between income, neighbourhood, and protected characteristics.

Gramespacher and Posth (2021) find that profit-optimised decision thresholds generate high rejection rates for applicants at the margin, and these marginal applicants are more likely to be low-income. Their analysis shows that one model configuration resulted in "refusal of 34% of creditworthy customers in the test subset" under a profit-maximisation objective. This is not a bias detection failure, the model correctly identifies these applicants as creditworthy, but the objective function makes their exclusion commercially rational.

Mhlanga (2021) notes that in Sub-Saharan Africa and South Asian markets, ML-based digital credit systems have expanded access for "less privileged people" through alternative data integration, but this access often comes at the cost of higher-priced products and elevated default rates that are partly attributable to income volatility rather than model failure. The populations gaining access through AI-driven models are precisely those whose income profiles make them systematically higher-risk, regardless of how well the model performs.

Volume 5 Issue No 2 (2026)
Online ISSN: 3006-2047
Print ISSN: 3006-2039

Table 1: Summary of Documented Demographic Disparities in Included Studies

Study	Context	Protected Group(s) Affected	Nature of Disparity	Mitigation Approach
Szepannek & Lübke (2021)	Germany, German Credit data	Race, gender, sexual orientation	Proxy discrimination via correlated variables	Counterfactual fairness correction
Hall et al. (2021)	United States	Race, gender, socioeconomic status	Multiple: proxy variables, disparate impact	Interpretable models, adversarial testing
Nwafor et al. (2024)	P2P lending, Lending Club	Age, gender	Feature presence in model inputs	Feature removal, SHAP audit
Johnen et al. (2021)	Kenya	Women, low-income	Gender gap in digital credit access; 90% blacklisting	Regulatory oversight
Laínez & Gardner (2023)	Vietnam	Women, low-income rural borrowers	Discrimination in algorithmic scoring	Legal framework proposal
Green & Chen (2019)	United States, Lending Club	Race	Bias in human-algorithm interaction	—
Gramespacher & Posth (2021)	Europe, P2P lending	Low-income, marginal applicants	Profit objective rejects creditworthy applicants	Objective function redesign
Brotcke (2022)	United States	Low-income, minority groups	Replication of historical disparities in ML	Bias assessment procedures
Agosto et al. (2023)	Europe, Italian SMEs	Social disadvantage (ESG proxy)	Model risk from inaccurate credit ratings	S.A.F.E. framework
Xu et al. (2023)	China	Behavioural and social stratification	Surveillance data creating new exclusion patterns	Regulatory intervention

Volume 5 Issue No 2 (2026)
Online ISSN: 3006-2047
Print ISSN: 3006-2039

Technical Mitigation Approaches

Causal Inference and Counterfactual Fairness

Szepannek and Lübke (2021) present the most methodologically rigorous fairness correction approach in the dataset. Using the German Credit dataset, they apply causal inference techniques to develop counterfactually fair scoring models, models that would assign the same score to a borrower regardless of their value on a protected attribute, assuming all else equal. Their results are cautiously encouraging: "it is possible to remove unfairness without a strong performance decrease unless the correlation of the discriminative attributes on the other predictor variables in the model is not too strong."

That qualifying clause matters considerably in practice. Where protected characteristics are highly correlated with many model predictors, as in markets with strong residential segregation or formal sector exclusion patterns, the correlation structure makes it very difficult to remove the fairness-relevant variance without also removing predictively useful variance. The counterfactual fairness approach works best in contexts where the protected attribute is relatively weakly correlated with the predictor set; it offers diminishing returns precisely in the high-inequality environments where fairness problems are most severe.

Feature Removal and SHAP-Based Auditing

Nwafor et al. (2024) take a more direct approach: they remove age and gender from the feature set of their hybrid 1DCNN-XGBoost model and test whether classification performance degrades. It does not, accuracy remains at 96% after removal of these features. This finding supports the practical feasibility of fairer credit scoring models and is cited in Hall et al.'s (2021) framework as evidence that interpretable, fair models can maintain competitive performance.

However, feature removal is a necessary but not sufficient fairness intervention. SHAP-based auditing, as applied by de Lange et al. (2022) and Gramegna and Giudici (2021), can identify which features contribute most to default predictions globally and locally. This is valuable for detecting unexpected feature importance that might indicate proxy discrimination. Gramegna and Giudici (2021), working with Italian SME data, demonstrate that SHAP and LIME both enable grouping of risky and non-risky borrowers by interpretable financial characteristic clusters, providing a post-hoc mechanism for auditing model decisions.

The limitation of SHAP-based auditing for fairness purposes is that SHAP values describe the contribution of features to predictions, not the causal mechanism behind those contributions. A feature that appears neutral in terms of protected characteristic correlation may still operate as a proxy through its interaction with other features, and SHAP decompositions at the individual feature level will not capture this.

The S.A.F.E. Framework

Agosto et al. (2023) propose an integrated evaluation framework, Sustainability, Accuracy, Fairness, Explainability, for assessing AI credit scoring systems across

Volume 5 Issue No 2 (2026)
Online ISSN: 3006-2047
Print ISSN: 3006-2039

multiple simultaneously relevant criteria. Applied to European corporate credit rating data, they demonstrate that models incorporating ESG factors can improve credit rating accuracy while simultaneously improving sustainability, fairness, and explainability metrics. This is evidence against a strong version of the fairness-accuracy trade-off: incorporating social governance factors does not necessarily degrade predictive performance.

Giudici and Wu (2025) extend this to Italian SME data and reach similar conclusions, ML models that account for ESG dimensions maintain competitive accuracy relative to models that do not. These findings are suggestive, but the ESG context is specific. ESG scores for corporate borrowers are a reasonably well-defined construct; the analogous construct for individual consumer borrowers is less clear, and the regulatory implications of incorporating social governance factors into consumer credit decisions carry different legal risks under fair lending statutes.

The Accuracy-Fairness Trade-Off

The accuracy-fairness trade-off is real but its magnitude and character are context-dependent. This is probably the most practically important finding in the literature for governance purposes, because much regulatory resistance to fairness requirements rests on the assumption that they impose unacceptable accuracy costs.

Szepannek and Lübke (2021) find that counterfactual fairness corrections produce minimal accuracy loss when protected attribute correlations with other predictors are moderate. Their simulation results show the trade-off is not fixed but varies with the correlation structure of the data. Nwafor et al. (2024) find no accuracy loss from removing age and gender features entirely. These results suggest the trade-off, while real in principle, may be modest in practice for many credit scoring applications.

The more important trade-off is not between accuracy and fairness per se, but between predictive optimality and distributional equity in outcomes. As Gramespacher and Posth (2021) show, a model optimised for accuracy will not necessarily produce outcomes that are fair in terms of who gains and who loses from credit access. This is a different problem, it concerns the choice of objective function rather than the choice of protected attribute handling, and it does not yield to technical fairness corrections. Changing the objective function means accepting some reduction in institutional profit in exchange for better distributional outcomes, which is a policy choice rather than a modelling choice.

Bücker et al. (2020) document the wider version of this trade-off: the use of interpretable models for regulatory compliance means institutions accept "higher reserves or more credit defaults" as the price of regulatory compliance with explainability mandates. The accuracy-fairness trade-off is therefore nested within a broader accuracy-explainability-fairness trade-off, in which regulatory requirements for transparency simultaneously constrain the model's ability to use all available predictive information and complicate its ability to achieve distributional fairness.

Volume 5 Issue No 2 (2026)
Online ISSN: 3006-2047
Print ISSN: 3006-2039

Table 2: Accuracy-Fairness Trade-Off Evidence Across Selected Studies

Study	Fairness Intervention	Accuracy Impact	Key Finding
Szepannek & Lübke (2021)	Counterfactual fairness correction	Minimal, unless high correlation	Trade-off depends on correlation structure of protected attributes
Nwafor et al. (2024)	Feature removal (age, gender)	None (96% maintained)	High-performance models need not rely on protected attributes
Gramespacher & Posth (2021)	Objective function redesign	Profit reduction	34% of creditworthy applicants rejected under profit-max objective
Bücker et al. (2020)	Interpretability requirements	Foregone predictive accuracy	Compliance with explainability mandates has real accuracy costs
Agosto et al. (2023)	S.A.F.E. framework (ESG)	No degradation	Fairness and accuracy not inherently opposed in ESG context

Governance Implications

Why Technical Fixes Are Insufficient Alone

The evidence reviewed here supports a specific argument: algorithmic bias in credit scoring is not primarily a technical problem and will not be resolved by technical solutions alone. This is not a dismissal of technical work, the counterfactual fairness methods of Szepannek & Lübke (2021), the feature auditing enabled by SHAP tools, and the demonstrated feasibility of removing discriminatory features without accuracy loss are all genuinely useful findings. But they address symptoms rather than causes.

The cause is structural. Historical lending data encodes historical discrimination. Profit-optimised objective functions produce distributional consequences that fall along existing lines of inequality. Regulatory frameworks have not yet established enforceable standards for what constitutes acceptable fairness performance in ML credit scoring models. None of these conditions is altered by improving the technical fairness toolkit. Institutional incentives, regulatory design, and market structure have to change alongside model architecture.

Hall et al. (2021) make this argument directly in the US context, arguing that fair lending compliance for ML models requires both technical tools and institutional commitment, neither alone is sufficient. They propose specific procedural mechanisms including adverse action testing, model documentation requirements, and disparate impact testing analogous to those applied in traditional lending, adapted for the high-dimensional, non-linear models that now dominate commercial practice.

Volume 5 Issue No 2 (2026)
Online ISSN: 3006-2047
Print ISSN: 3006-2039

Regulatory Context and Jurisdictional Gaps

The regulatory landscape for AI credit scoring fairness is profoundly uneven. In the United States, the Equal Credit Opportunity Act (ECOA) and Fair Housing Act prohibit discriminatory lending, and federal regulators have published guidance on applying these statutes to algorithmic models. In the European Union, GDPR Article 22 restricts automated decision-making with significant effects and requires meaningful explanations of adverse decisions. Laínez and Gardner (2023), examining Vietnam, document the opposite end of the spectrum: absent credit law preparation for algorithmic scoring, piecemeal privacy regulation, and no AI-specific ethical guidelines. Under these conditions, commercial actors have little institutional pressure to address fairness problems.

Brotcke (2022) documents specific challenges that US regulators face in examining ML credit models for fair lending compliance: the difficulty of detecting proxy discrimination in high-dimensional models using examination procedures designed for simpler systems, the absence of standardised frameworks for ML model validation, and the risk that regulatory uncertainty itself delays adoption of fairer model architectures by creating perverse incentives to stay with interpretable but potentially biased conventional methods.

The regulatory arbitrage potential is significant. Institutions operating across jurisdictions with different regulatory standards may deploy AI credit scoring models with features or objective functions that would be legally problematic in high-standard jurisdictions in markets where oversight is weak. Xu et al.'s (2023) China findings illustrate how institutional actors adapt credit scoring architectures to local regulatory and data environments in ways that produce new forms of stratification not anticipated by policy frameworks designed for earlier-generation scoring systems.

The Role of Alternative Data in Bias Dynamics

Alternative data sources, mobile phone records, digital behavioural patterns, social network data, expand financial access for populations excluded from conventional credit markets. But they also introduce new bias pathways. Xu et al. (2023) document how extensive collection of "intimate territories of the self" for credit scoring in China creates surveillance infrastructure that enables new forms of social sorting, including along characteristics that map onto protected categories even when those categories are not explicitly represented.

Mhlanga (2021) and Roa et al. (2021) document that Super-App and mobile data-based scoring primarily benefits users with extensive platform engagement, those who are most active on digital platforms. This creates a new form of exclusion: individuals who are poor, elderly, or without smartphone access are excluded from alternative-data scoring advantages, and these populations overlap with groups that have historically been excluded from formal credit. The bias pattern is different in character from proxy discrimination, it is more about access to data than about discrimination conditional on data availability, but its distributional consequences are similarly adverse for vulnerable populations.

Volume 5 Issue No 2 (2026)
Online ISSN: 3006-2047
Print ISSN: 3006-2039

Conclusion

Algorithmic bias in AI credit scoring takes multiple forms, originates from multiple sources, and requires multiple governance responses. The core empirical findings from this review can be stated plainly.

Bias in ML credit scoring models is partly inherited from historically discriminatory training data and partly generated by profit-optimised objective functions that produce adverse distributional outcomes independently of any intent to discriminate. Both mechanisms are at work in most commercial credit scoring contexts, and they are not fully separable.

Gender and socioeconomic disparities are the most consistently documented in the empirical literature. The mechanisms are varied: direct feature use, proxy correlation, objective function effects, and access gaps in alternative data. In markets with weak financial infrastructure, Kenya, Vietnam, rural South Asia, these disparities are compounded by income volatility among newly enfranchised borrowers.

Technical mitigation approaches show conditional effectiveness. Counterfactual fairness corrections work well when protected attribute correlations with other predictors are moderate. Feature removal without accuracy loss is empirically demonstrated for gender and age in at least one P2P lending context. SHAP-based auditing provides useful but incomplete fairness information. The S.A.F.E. framework offers a multi-dimensional evaluation approach with empirical support in European corporate credit contexts.

The accuracy-fairness trade-off is real but not as severe as often assumed. More important is the objective function trade-off: profit-optimised models reject creditworthy marginal applicants at commercially rational thresholds, and this cost falls disproportionately on low-income and historically excluded populations. Resolving this requires changing institutional objectives, not just model architecture.

Effective governance of algorithmic bias requires enforceable regulatory standards for fair lending performance in ML models, not merely guidance; institutional accountability mechanisms that create reputational and financial costs for discriminatory outcomes; and cross-jurisdictional regulatory coordination to close arbitrage gaps. Technical tools are necessary but not sufficient absent this institutional substrate.

Future research should focus on: the bias dynamics of alternative data in consumer credit contexts beyond the corporate ESG domain; the cumulative effects of AI credit scoring on credit bureau exclusion rates across demographic groups; and empirical comparison of fairness outcomes across regulatory regimes with different levels of oversight stringency.

References

Aggarwal, N. (2021). The norms of algorithmic credit scoring. The Cambridge Law Journal, 80(1), 42–73. https://doi.org/10.1017/s0008197321000015

Agosto, A., Cerchiello, P., & Giudici, P. (2023). Bayesian learning models to measure the relative impact of ESG factors on credit ratings. International Journal of

Volume 5 Issue No 2 (2026)
Online ISSN: 3006-2047
Print ISSN: 3006-2039

Data Science and Analysis, 10(2), 45–63. https://doi.org/10.1007/s41060-023-00405-9

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. Journal of Financial Economics, 143(1), 130–156. https://doi.org/10.1016/j.jfineco.2021.05.047

Brotcke, L. (2022). Time to assess bias in machine learning models for credit decisions. Journal of Risk and Financial Management, 15(4), 165. https://doi.org/10.3390/jrfm15040165

Bücker, M., Szepannek, G., Gosiewska, A., & Biecek, P. (2020). Transparency, auditability, and explainability of machine learning models in credit scoring. Journal of the Operational Research Society, 73(1), 70–90. https://doi.org/10.1080/01605682.2021.1922098

Das, S., Stanton, R., & Wallace, N. (2023). Algorithmic fairness. Annual Review of Financial Economics, 15(1), 565–593. https://doi.org/10.1146/annurev-financial-110921-125930

de Lange, P. D., Melsom, B., Vennerød, C. B., & Westgaard, S. (2022). Explainable AI for credit assessment in banks. Journal of Risk and Financial Management, 15(12), 556. https://doi.org/10.3390/jrfm15120556

Giudici, P., & Wu, L. (2025). Sustainable artificial intelligence in finance: Impact of ESG factors. Frontiers in Artificial Intelligence, 8, 1566197. https://doi.org/10.3389/frai.2025.1566197

Gramegna, A., & Giudici, P. (2021). SHAP and LIME: An evaluation of discriminative power in credit risk. Frontiers in Artificial Intelligence, 4, 752558. https://doi.org/10.3389/frai.2021.752558

Gramespacher, T., & Posth, J.-A. (2021). Employing explainable AI to optimize the return target function of a loan portfolio. Frontiers in Artificial Intelligence, 4, 693022. https://doi.org/10.3389/frai.2021.693022

Green, B., & Chen, Y. (2019). The principles and limits of algorithm-in-the-loop decision making. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 50. https://doi.org/10.1145/3359152

Hall, P., Cox, B. G., Dickerson, S. N., Kannan, A. R., Kulkarni, R., & Schmidt, N. (2021). A United States fair lending perspective on machine learning. Frontiers in Artificial Intelligence, 4, 695301. https://doi.org/10.3389/frai.2021.695301

Hurlin, C., Pérignon, C., & Saurin, S. (2023). The fairness of credit scoring models. Management Science, 69(11), 6612–6639. https://doi.org/10.1287/mnsc.2022.03888

Johnen, C., Parlasca, M. C., & Musshoff, O. (2021). Promises and pitfalls of digital credit: Empirical evidence from Kenya. PLOS ONE, 16(8), e0255215. https://doi.org/10.1371/journal.pone.0255215

Laínez, N., & Gardner, J. (2023). Algorithmic credit scoring in Vietnam: A legal proposal for maximizing benefits and minimizing risks. Asian Journal of Law and Society, 10(3), 401–431. https://doi.org/10.1017/als.2023.6

Volume 5 Issue No 2 (2026)
Online ISSN: 3006-2047
Print ISSN: 3006-2039

Mhlanga, D. (2021). Financial inclusion in emerging economies: The application of machine learning and artificial intelligence in credit risk assessment. International Journal of Financial Studies, 9(3), 39. https://doi.org/10.3390/ijfs9030039

Nwafor, C., Nwafor, O., & Brahma, S. (2024). Enhancing transparency and fairness in automated credit decisions: An explainable novel hybrid machine learning approach. Scientific Reports, 14(1), 24691. https://doi.org/10.1038/s41598-024-75026-8

Roa, L., Rodríguez-Rey, A., Correa Bahnsen, A., & Valencia, C. (2021). Supporting financial inclusion with graph machine learning and super-app alternative data. Intelligent Systems with Applications, 12, 200051. https://doi.org/10.1007/978-3-030-82196-8_16

Szepannek, G., & Lübke, K. (2021). Facing the challenges of developing fair risk scoring models. Frontiers in Artificial Intelligence, 4, 681915. https://doi.org/10.3389/frai.2021.681915

Xu, R., Millo, Y., & Spence, C. (2023). The mountains are high and the Emperor is far away: Credit scoring and the infrastructure of surveillance capitalism in China. Contemporary Accounting Research, 40(4), 2623–2657. https://doi.org/10.1111/1911-3846.12925