International Journal of Engineering Research & Management Technology
(Peer-Reviewed, Open Access, Fully Refereed International Journal) ISSN: 2348-4039
Volume 11, Issue-6 November-December- 2024 Impact Factor: 7.09
Email: editor@ijermt.in www.ijermt.in

A Fairness Audit of Behavioural Credit Scoring Models in Indian Fintech Lending

T Lokendran
Research Scholar, NIILM University, Kaithal, Haryana, India
Dr. Kratika Saxena
Professor, NIILM University, Kaithal, Haryana

Abstract

Purpose: Machine learning models trained on alternative behavioural data are being promoted as a route to credit inclusion for borrowers without credit histories. This study asks whether such models are also fair: whether they reduce disparities across gender, occupation, and geography relative to conventional scoring, whether behavioural features act as proxies for protected characteristics, and what governance safeguards stakeholders consider necessary.

Design/methodology/approach: A convergent mixed-methods design is applied to Southern India. The quantitative strand audits five model families trained on 12,847 anonymised thin-file borrower records from Bengaluru, Chennai, and Hyderabad, using disparate impact ratios, equalized-odds tests, and SHAP-based proxy detection with feature-removal sensitivity analysis. The qualitative strand analyses twenty-two semi-structured interviews with credit risk professionals, fintech practitioners, regulators, and data ethicists using reflexive thematic analysis.

Findings: Behavioural models substantially reduce disparate impact: the gender ratio rises from 0.73 under conventional scoring to 0.88 under gradient boosting, and the employment ratio from 0.63 to 0.84, so all comparisons pass the four-fifths rule. Equalized odds are not achieved: significant true-positive-rate gaps persist for women (0.044, $p = 0.032$) and daily wage workers (0.085, $p = 0.004$). The smartphone value index acts as a proxy for gender and income; removing it improves the gender ratio to 0.91 at a cost of 0.008 in AUC-ROC. Experts identify privacy, consent, and explainability as the binding constraints on responsible deployment.

Originality/value: The study offers one of the first multi-metric fairness audits of behavioural credit scoring in a developing economy, shows that inclusion and fairness are distinct outcomes, and proposes an operational fairness audit framework for Indian lenders and regulators.

Keywords: algorithmic fairness; credit scoring; alternative data; disparate impact; equalized odds; explainable AI; India

JEL classification: G21; G28; D63; C53; O33

1. Introduction

Credit is allocated increasingly by algorithms. Fintech lenders, non-banking financial companies, and banks now use machine learning models that draw on alternative data — telephone usage, digital payments, utility bills, and device characteristics — to score applicants who lack a credit bureau history (Frost et al., 2019; Cornelli et al., 2023). The promise is inclusion: by observing the everyday economic behaviour of people the bureau cannot see, lenders can extend credit to the self-employed, gig workers, migrants, and women who have been excluded from formal finance (Berg et al., 2020; Agarwal et al., 2020). Evidence from Southern India supports this promise: a companion study shows that behavioural machine learning roughly doubles the share of thin-file applicants who can be approved at an unchanged default rate (Lokendran & Saxena, n.d.).

Inclusion, however, is not the same as fairness. A model can approve more applicants overall while treating some groups systematically worse than others. Behavioural data are generated within social structures marked by unequal access to smartphones, data plans, digital payments, and stable housing (GSMA, 2024; Gabor & Brooks, 2017). Features derived from such data may therefore encode gender, income, or occupation even when those characteristics are excluded from the model — the problem of proxy discrimination (Barocas & Selbst, 2016). Evidence from the United States shows that machine learning can raise accuracy while widening some disparities (Fuster et al., 2022) and that algorithmic lenders discriminate less than human lenders but do not eliminate discrimination (Bartlett et al., 2022). Whether these patterns hold for behavioural scoring in a developing economy is largely unknown.

The question is pressing in India. The Reserve Bank of India's digital lending guidelines regulate data collection by lending apps (Reserve Bank of India, 2022), and the Digital Personal Data Protection Act, 2023 imposes consent and purpose-limitation obligations (Government of India, 2023), but neither sets standards for testing the fairness of credit algorithms. NITI Aayog's (2021) principles for responsible artificial intelligence emphasise equality and non-discrimination without specifying how they should be measured in lending. Lenders and regulators thus lack both evidence and tools.

This study addresses these gaps with a convergent mixed-methods design. Quantitatively, it audits five model families on 12,847 thin-file borrower records from Bengaluru, Chennai, and Hyderabad along three dimensions: outcome fairness (disparate impact), error-rate fairness (equalized odds), and mechanism (proxy detection with SHAP values). Qualitatively, it analyses twenty-two expert interviews to understand the institutional conditions for responsible deployment. The study makes three contributions: it provides one of the first multi-metric fairness audits of behavioural credit scoring in a developing economy; it shows empirically that improvements in inclusion and in outcome fairness do

not guarantee error-rate fairness; and it proposes an operational fairness audit framework grounded in both the quantitative results and stakeholder perspectives.

Section 2 reviews the literature and states the hypotheses. Section 3 describes the data and methods, Section 4 the results, and Section 5 the discussion. Section 6 proposes a fairness audit framework and policy implications, Section 7 notes limitations, and Section 8 concludes.

2. Literature Review and Hypotheses

2.1 Defining Fairness in Credit Scoring

The algorithmic fairness literature distinguishes several criteria, which are summarised in Table 1 (Mehrabi et al., 2021; Verma & Rubin, 2018). Demographic parity requires equal approval rates across groups and is operationalised in anti-discrimination practice through the disparate impact ratio and the four-fifths rule (Feldman et al., 2015). Equalized odds requires equal true positive and false positive rates across groups, so that creditworthy applicants have the same chance of approval regardless of group, and non-creditworthy applicants the same chance of rejection (Hardt et al., 2016). Calibration requires that a given predicted probability corresponds to the same observed default rate in every group, and individual fairness requires that similar individuals be treated similarly (Dwork et al., 2012).

Table 1. Fairness criteria used in the audit

Criterion	Definition	Operational test
Disparate impact (demographic parity)	Approval rate of a group divided by that of the reference group	Ratio ≥ 0.80 (four-fifths rule)
Equalized odds	Equal true positive and false positive rates across groups	χ^2 test of TPR and FPR differences
Proxy discrimination	A model input correlated with a protected attribute drives outcomes	Correlation with attribute; SHAP contribution; removal sensitivity
Calibration	Equal observed default at equal predicted probability	Group-wise calibration and Brier score

Note: in this study a 'positive' prediction means the applicant is classed as creditworthy (approved).
Sources: Feldman et al. (2015); Hardt et al. (2016); Mehrabi et al. (2021).

These criteria cannot in general be satisfied simultaneously. When base rates of default differ between groups, a classifier cannot achieve both equalized odds and equal predictive value (Chouldechova, 2017; Kleinberg et al., 2017), and enforcing fairness constraints typically carries some cost in accuracy (Corbett-Davies et al., 2017). Fairness assessment therefore requires multiple metrics and explicit choices about which criteria matter in a given context.

2.2 Alternative Data, Machine Learning, and Disparity

There are reasons to expect behavioural models to be fairer than conventional ones. Conventional variables — documented income, formal employment, collateral — are themselves unequally distributed, so scorecards built on them disadvantage informal workers and women (Demirgüç-Kunt et al., 2022). Behavioural data measure conduct rather than status, and so may reveal creditworthiness in groups that conventional variables overlook (Hurley & Adebayo, 2016). Bartlett et al. (2022) found that algorithmic lenders discriminated about a third less than face-to-face lenders in US mortgage pricing. This suggests the first hypothesis:

> ***H1:*** *Machine learning models trained on behavioural data exhibit less disparate impact — disparate impact ratios closer to parity — across gender, occupation, and city than conventional scoring.*

There are also reasons for caution. More flexible models can exploit subtle correlations between features and group membership, so that gains in accuracy are distributed unequally (Fuster et al., 2022). Behavioural data may be noisier for groups with irregular incomes or limited digital access, so that models identify creditworthy members of those groups less reliably. Equalized odds is a stricter criterion than disparate impact, and improvement on the latter need not imply improvement on the former:

> ***H2:*** *Machine learning models trained on behavioural data achieve equalized odds — statistically equal true positive and false positive rates — across gender and occupation.*

2.3 Proxy Discrimination and Explainability

Excluding protected attributes from a model does not prevent discrimination if other inputs are correlated with them (Barocas & Selbst, 2016). Explainable AI methods allow such proxies to be detected. SHAP values, grounded in cooperative game theory, attribute each prediction to individual features (Lundberg & Lee, 2017; Shapley, 1953), and have been applied to credit risk management (Bussmann et al., 2021). Kozodoi et al. (2022) show that fairness can be improved at modest profit cost through pre-, in-, and post-processing interventions; the simplest pre-processing intervention is to remove an identified proxy:

> ***H3:*** *Removing behavioural features identified as proxies for protected characteristics improves fairness metrics with only a marginal loss of predictive accuracy.*

2.4 Governance and Stakeholder Perspectives

Fairness is ultimately enacted by institutions. Scholars have called for due-process protections for scored individuals (Citron & Pasquale, 2014), argued that consent-based privacy regimes are weak when individuals cannot meaningfully evaluate data uses (Solove, 2013), and proposed counterfactual

explanations as a means of giving applicants actionable reasons (Wachter et al., 2018). Internal algorithmic audits (Raji et al., 2020) and model documentation (Mitchell et al., 2019) are emerging as governance tools, and regulators such as the Monetary Authority of Singapore (2018) have articulated principles of fairness, ethics, accountability, and transparency. Perceived legitimacy also depends on procedural fairness — whether people believe the process is fair (Tyler, 2006). Because these questions concern institutional readiness rather than measurable model properties, they are examined qualitatively through the following proposition:

> ***P1:*** *Stakeholders perceive behavioural credit scoring as creating privacy, consent, transparency, and governance demands that exceed those of conventional credit assessment.*

2.5 The Indian Regulatory Context

India's regulatory framework for credit algorithms is developing but incomplete. The Credit Information Companies (Regulation) Act, 2005 governs bureau data but does not address alternative data (Government of India, 2005). The Reserve Bank of India's guidelines on digital lending, issued after a working group documented abusive practices by lending apps, require lending through regulated entities, prohibit indiscriminate access to borrowers' phone data, and mandate explicit consent and transparent pricing (Reserve Bank of India, 2021, 2022). The Digital Personal Data Protection Act, 2023 requires consent that is free, specific, informed, and unambiguous, limits processing to stated purposes, and gives individuals rights of access, correction, and erasure (Government of India, 2023). The constitutional recognition of privacy as a fundamental right underpins these obligations (Justice K. S. Puttaswamy (Retd.) v. Union of India, 2017).

None of these instruments, however, specifies how lenders should test credit algorithms for bias, which fairness metrics should be reported, or how applicants may contest automated decisions. NITI Aayog's (2021) responsible-AI principles call for equality, inclusivity, and accountability, but remain principles rather than operational standards. By contrast, the European Union's data protection regime gives individuals rights in relation to solely automated decisions (European Union, 2016), and the Monetary Authority of Singapore (2018) has issued FEAT principles that financial institutions are expected to apply. India therefore has the legal foundations for consent and privacy but lacks fairness standards — a gap that empirical evidence of the kind presented here can help to fill.

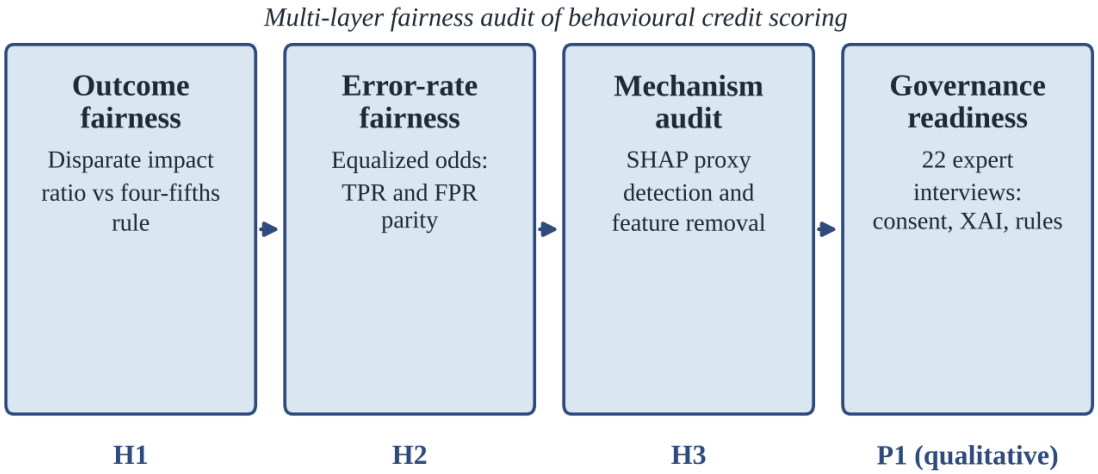


Figure 1. Multi-layer fairness audit framework and hypotheses

Source: Developed by the authors.

3. Data and Methods

3.1 Research Design

The study uses a convergent mixed-methods design in which a quantitative fairness audit and a qualitative expert study are conducted in parallel and integrated at the interpretation stage (Creswell & Plano Clark, 2018; Fetters et al., 2013). The design is appropriate because fairness has both measurable statistical dimensions and institutional dimensions that statistics cannot capture.

3.2 Quantitative Data and Models

The quantitative strand uses 12,847 anonymised records of thin-file borrowers from Bengaluru (38.2 per cent), Chennai (33.5 per cent), and Hyderabad (28.3 per cent), obtained through institutional data-access arrangements with fintech platforms and telecommunications providers. Each record links behavioural features — telecommunications regularity, UPI transaction patterns, utility payment behaviour, device characteristics, and residential stability, engineered into 84 features — to a binary outcome of default (90 or more days past due). Data were pre-anonymised and k-anonymity with $k \geq 5$ was applied (Sweeney, 2002).

Protected and audit attributes are gender, employment type, city, and income band. These attributes were used only for auditing; gender was not a model input in the behavioural models. Table 2 shows that default base rates differ across groups — a fact that, by the impossibility results noted above, makes perfect equalisation of all fairness metrics mathematically unattainable.

Table 2. Audit groups and default base rates

Attribute	Group	Share of sample	Default rate
Gender	Male	61.4%	17.6%
	Female	38.6%	19.4%
Employment	Self-employed	34.7%	15.3%
	Salaried informal	22.1%	17.8%
	Gig workers	27.3%	19.9%
	Daily wage workers	15.9%	22.4%
City	Bengaluru	38.2%	16.8%
	Chennai	33.5%	19.2%
	Hyderabad	28.3%	19.7%
All		100.0%	18.3%

Source: Authors' computation.

Five model families are audited against a conventional logistic regression trained only on traditional variables (age, gender, income band, employment type, and residential stability): a decision tree, a random forest, an XGBoost gradient boosting machine, and a feedforward neural network, all trained on behavioural features under nested cross-validation (Breiman, 2001; Chen & Guestrin, 2016; Cawley & Talbot, 2010). Their discriminative performance ranges from an AUC-ROC of 0.641 for the conventional model to 0.847 for gradient boosting; full predictive results are reported in the companion paper.

3.3 Fairness Metrics

Approval decisions are compared at a common operating point at which each model's post-approval default rate does not exceed 10 per cent. Disparate impact is the approval rate of the focal group divided by that of the reference group (men, self-employed borrowers, and Bengaluru respectively), with ratios below 0.80 indicating adverse impact (Feldman et al., 2015). Equalized odds is assessed by comparing true positive rates (the share of actual non-defaulters approved) and false positive rates (the share of actual defaulters approved) across groups, with differences tested by chi-square tests (Hardt et al., 2016). Proxy detection combines the correlation of each feature with protected attributes and its SHAP contribution to predictions (Lundberg & Lee, 2017); features identified as proxies are removed and all metrics recomputed.

3.4 Qualitative Data and Analysis

Twenty-two semi-structured interviews were conducted with seven credit risk professionals from banks and NBFCs, six fintech practitioners, five policymakers and regulators, and four data ethicists and researchers, selected through purposive and snowball sampling and guided by thematic saturation (Guest

et al., 2006; Patton, 2015). The protocol, pre-tested with three experts, covered six domains: limitations of current scoring, the potential of behavioural data, model reliability and interpretability, privacy and consent, regulatory readiness, and implementation. Interviews of 45 to 60 minutes were recorded with written consent, transcribed, and analysed using Braun and Clarke's (2006) six-phase thematic analysis with deductive and inductive codes. Inter-coder reliability was assessed on three transcripts against a Cohen's (1960) kappa target of 0.70. Table 3 summarises the respondents.

Table 3. Interview respondents

Stakeholder category	n	Focus of perspective
Credit risk professionals (banks, NBFCs)	7	Underwriting practice, model risk, explainability
Fintech practitioners	6	Alternative-data products, data integration, sandboxes
Policymakers and regulators	5	Regulation, consumer protection, systemic trust
Data ethicists and researchers	4	Bias, consent, structural inequality
Total	22	

Source: Authors' fieldwork.

3.5 Integration and Trustworthiness

The two strands were analysed independently and then integrated using a joint display in which each quantitative result is set against the qualitative evidence bearing on it, classifying the relationship as convergence, complementarity, or divergence (Fetters et al., 2013). Trustworthiness of the qualitative strand was supported by pre-testing the protocol, offering respondents the opportunity to review their transcripts, double-coding a subset of transcripts, and maintaining an audit trail of coding decisions (Lincoln & Guba, 1985). Respondents and their organisations are anonymised, and findings are reported by stakeholder category.

4. Results

4.1 Disparate Impact (H1)

Table 4 and Figure 2 report disparate impact ratios. The conventional model fails the four-fifths rule for gender (0.73) and for daily wage relative to self-employed workers (0.63). Every behavioural model improves both ratios, and gradient boosting achieves 0.88 for gender and 0.84 for employment, passing the threshold on both. City ratios are high for all behavioural models (0.89 to 0.93); the conventional model's Hyderabad–Bengaluru ratio of 0.84 is its lowest city ratio. Behavioural scoring therefore narrows disparities in approval outcomes on every dimension examined. H1 is supported.

Table 4. Disparate impact ratios by model

Comparison	Conv. LR	Decision tree	Random forest	XGBoost	Neural net
Gender (female / male)	0.73	0.82	0.86	0.88	0.85
Employment (daily wage / self-employed)	0.63	0.71	0.79	0.84	0.81
City (Chennai / Bengaluru)	0.90	0.91	0.93	0.93	0.92
City (Hyderabad / Bengaluru)	0.84	0.89	0.91	0.92	0.90

Note: values below 0.80 fail the four-fifths rule. Source: Authors' computation.

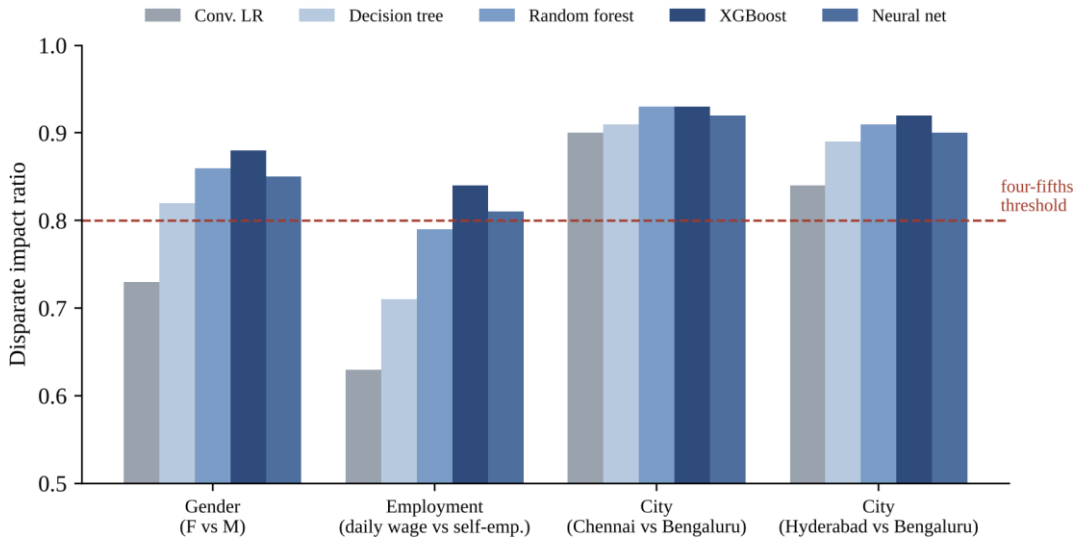


Figure 2. Disparate impact ratios across models and audit groups

Source: Table 4.

The improvement reflects the substitution of conduct for status. The conventional model relies on variables such as income band and employment type, which mechanically disadvantage daily wage workers. Behavioural features reveal discipline within that group — for example, timely bill payment, which subgroup feature-importance analysis shows to be relatively more important for daily wage workers than for other groups — so that more of its creditworthy members can be identified.

4.2 Equalized Odds (H2)

Error-rate fairness presents a different picture (Table 5, Figure 3). Under gradient boosting, the true positive rate — the probability that a borrower who will repay is approved — is 0.658 for men and 0.614 for women. The difference of 0.044 is statistically significant ($p = 0.032$). The false positive rate is 0.124 for men and 0.138 for women; that difference of 0.014 is not significant ($p = 0.187$). Across employment types the gap is larger: the true positive rate is 0.672 for self-employed borrowers but 0.587 for daily

wage workers, a difference of 0.085 (p = 0.004). Creditworthy women and creditworthy daily wage workers are therefore less likely to be recognised as such. H2 is not supported.

Table 5. Equalized-odds analysis (gradient boosting)

Comparison	Reference group	Focal group	Difference	p-value
True positive rate — gender	Men 0.658	Women 0.614	0.044	0.032
False positive rate — gender	Men 0.124	Women 0.138	0.014	0.187
True positive rate — employment	Self-employed 0.672	Daily wage 0.587	0.085	0.004

Note: TPR = share of actual non-defaulters approved; FPR = share of actual defaulters approved. χ^2 tests. Source: Authors' computation.

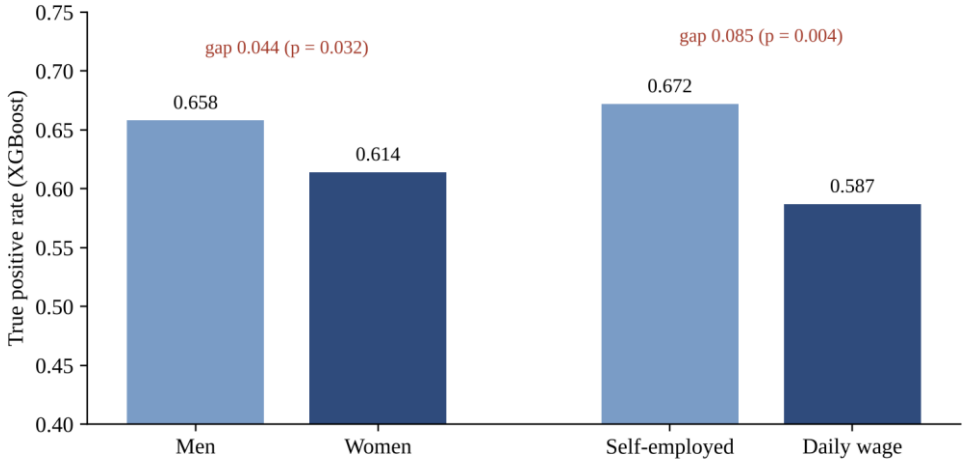


Figure 3. True positive rates by gender and occupation (gradient boosting)
Source: Table 5.

Two mechanisms are consistent with this pattern. First, base rates differ (Table 2), and fairness theory shows that equal error rates cannot generally coexist with calibration when base rates differ (Chouldechova, 2017). Second, behavioural data are less informative for groups whose digital activity is irregular or partly cash-based. Models trained on fewer or noisier signals for a group will identify its creditworthy members less reliably even if they do not penalise the group as such. Analysis of disagreement among model families supports this second mechanism: daily wage workers, borrowers under 25, and borrowers with fewer than 60 days of behavioural data are over-represented among cases on which models disagree, indicating greater assessment uncertainty for these groups.

4.3 Proxy Detection and Mitigation (H3)

SHAP analysis identified the smartphone value index — the price of the applicant's device relative to the market average — as a proxy. It correlates with gender ($r = 0.18$, $p < 0.001$) and, more strongly, with

income ($r = 0.42$, $p < 0.001$), and its SHAP contributions favour owners of more expensive devices. Borrowers with cheaper phones, who are disproportionately women, lower-income, and informally employed, therefore receive less favourable predictions partly through this feature.

Removing the smartphone value index and re-estimating gradient boosting reduces AUC-ROC by only 0.008, from 0.847 to 0.839, while raising the gender disparate impact ratio from 0.88 to 0.91 and narrowing the gender true-positive-rate gap from 0.044 to 0.029 (Table 6, Figure 4). A single targeted intervention thus delivers a meaningful fairness gain at negligible cost to accuracy, supporting H3. It does not, however, eliminate the gap, and the larger occupational gap remains, indicating that proxy removal is necessary but not sufficient.

Table 6. Effect of removing the smartphone value index (gradient boosting)

Metric	With proxy	Proxy removed	Change
AUC-ROC	0.847	0.839	–0.008
Gender disparate impact ratio	0.88	0.91	+0.03
Gender TPR gap	0.044	0.029	–0.015

Source: Authors' computation.

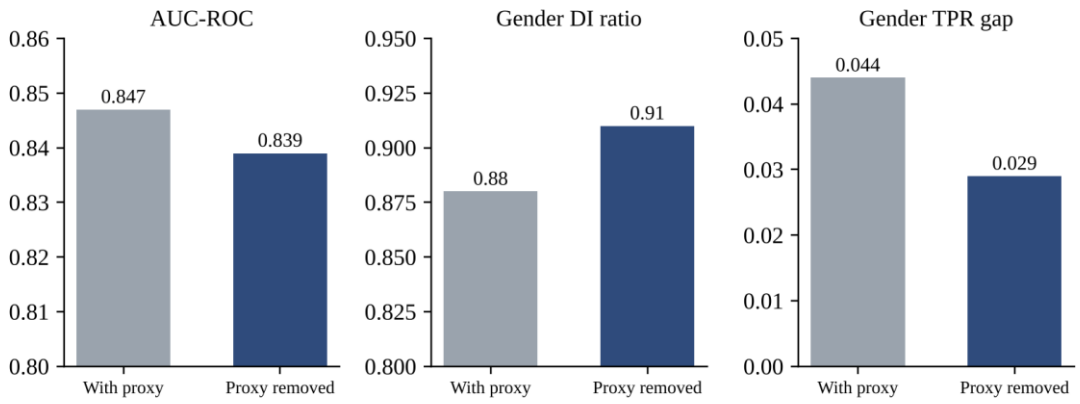


Figure 4. Accuracy–fairness effect of proxy removal

Source: Table 6.

4.4 Stakeholder Perspectives (P1)

Thematic analysis produced five principal themes (Table 7, Figure 5). Because the interview data are reported thematically, they are summarised by category of respondent rather than by direct quotation.

Table 7. Themes from the expert interviews

Theme	Summary of evidence
T1. Inadequacy of current scoring	All 22 respondents agreed that bureau-based scoring fails thin-file borrowers and that manual underwriting is inconsistent and unscalable.
T2. Behavioural data: cautious optimism	Risk and fintech respondents saw value in payment and transaction regularity; data ethicists warned that behaviour reflects structural factors (income, infrastructure, gender norms) and may encode inequality.
T3. Transparency and regulatory readiness	Regulators and risk professionals saw opaque models as hard to defend to boards, auditors, and supervisors; explainability tools and sandboxes were seen as preconditions for adoption.
T4. Privacy, consent, and trust	The most consistent concern across all categories, including 'consent fatigue', the gap between legal consent and genuine understanding, and reputational risk.
T5. Implementation safeguards	Hybrid scoring, continuous monitoring, routine fairness audits, borrower-facing explanations and contestation, and regulatory sandboxes.

Source: Authors' thematic analysis of 22 interviews.

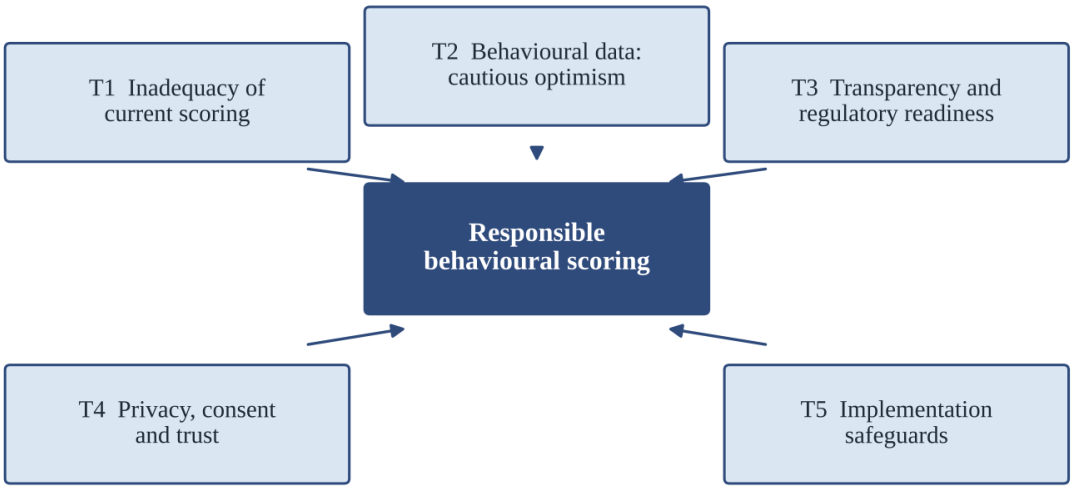


Figure 5. Thematic map of stakeholder perspectives

Source: Authors' analysis.

Secondary themes deepen these findings. Respondents described lending officers who, when shown both traditional and behavioural scores, placed greater weight on the traditional score even where the behavioural score was more accurate — an instance of algorithm aversion (Dietvorst et al., 2015). Data ethicists stressed procedural fairness: borrowers may regard behavioural scoring as unfair if they cannot understand or contest it, even when outcomes are more equitable (Tyler, 2006). Respondents also identified the absence of data standards across telecom, payment, and utility providers as a source of

noise that may affect groups unequally. Taken together, the interviews support P1: stakeholders regard behavioural scoring as creating governance demands beyond those of conventional scoring.

4.5 Fairness of the Financial Trust Score

Lenders may prefer a single composite score to a complex model, so the six-component financial trust score developed in the companion paper was also audited. Mean scores differ across groups (Table 8): men score 54.2 and women 50.1 on a 0–100 scale (difference 4.1, $p < 0.001$), and scores range from 56.8 for the self-employed to 47.3 for daily wage workers. These differences reflect lower average UPI activity and call regularity among women and greater income irregularity among daily wage workers. Crucially, however, the score's validity is similar across groups: its correlation with default is –0.39 for men and –0.43 for women, and between –0.36 and –0.44 across employment types. The score therefore measures creditworthiness equally well in each group, but groups start from different positions — so a single cut-off would approve proportionally fewer women and daily wage workers even though the score is not biased in its predictive relationship. This is a clear illustration of how fairness depends on the decision rule as well as on the model.

Table 8. Financial trust score by group

Group	Mean score (0–100)	Correlation with default
Men	54.2 (SD 17.8)	–0.39
Women	50.1 (SD 19.2)	–0.43
Self-employed	56.8	within –0.36 to –0.44
Salaried informal	53.2	within –0.36 to –0.44
Gig workers	50.7	within –0.36 to –0.44
Daily wage workers	47.3	within –0.36 to –0.44
Bengaluru / Chennai / Hyderabad	53.8 / 51.9 / 52.1	—

Note: gender difference in means $p < 0.001$ (t-test). Source: Authors' computation.

4.6 Integration of Quantitative and Qualitative Findings

Table 9 presents the joint display. The strands converge on two points: behavioural data carry real signals of creditworthiness, and those signals can encode structural inequality. They complement each other on governance, where the qualitative strand identifies institutional barriers that metrics cannot reveal. They diverge on interpretability: the most accurate model is the least transparent, whereas respondents regarded explainability as a non-negotiable condition for adoption.

Table 9. Joint display of quantitative and qualitative findings

Issue	Quantitative result	Qualitative evidence	Relationship
Value of behavioural data	AUC-ROC 0.847 vs 0.641; approval gains for all groups	Cautious optimism; payment regularity seen as meaningful	Convergence
Structural bias	TPR gaps for women and daily wage workers; device proxy	Ethicists: behaviour reflects income, infrastructure, gender norms	Convergence
Interpretability	Accuracy rises as transparency falls	Explainability regarded as essential for adoption	Divergence
Governance	Not measurable from model outputs	Consent, contestability, data standards, officer acceptance	Complementarity

Source: Authors' integration (Fetters et al., 2013).

Table 10. Summary of results

Hypothesis / proposition	Result	Key evidence
H1: Less disparate impact than conventional scoring	Supported	Gender 0.73 → 0.88; employment 0.63 → 0.84
H2: Equalized odds achieved	Not supported	TPR gaps: gender 0.044 (p = 0.032); employment 0.085 (p = 0.004)
H3: Proxy removal improves fairness at small cost	Supported	AUC –0.008; gender DI 0.88 → 0.91; TPR gap 0.044 → 0.029
P1: Heightened governance demands	Supported	Privacy and consent most consistent theme across all 22 interviews

Source: Authors' compilation.

5. Discussion

The central finding is that inclusion and fairness are related but distinct outcomes. Behavioural scoring expands access and narrows disparities in approval rates, so on the criterion most familiar to anti-discrimination law it is fairer than conventional scoring. Yet the same models remain less able to recognise creditworthy women and daily wage workers. In practical terms, a creditworthy daily wage worker has about a 59 per cent chance of approval under the best model, compared with 67 per cent for an equally creditworthy self-employed borrower. A lender or regulator that tracked only disparate impact would conclude that the problem had been solved; one that also tracked error rates would see that it had only been reduced.

This pattern mirrors Fuster et al. (2022), who found that machine learning in US mortgage markets benefits some groups more than others, and Bartlett et al. (2022), who found that algorithmic lending reduces but does not eliminate discrimination. The present study adds a developing-economy setting and

a mechanism. The persistence of error-rate gaps appears to stem less from the model penalising groups than from unequal informativeness of behavioural data: borrowers whose economic lives are partly cash-based, irregular, or recently digitised leave thinner traces, and the model is therefore less certain about them. Fairness in behavioural scoring is thus partly a problem of data infrastructure, not only of algorithms.

The proxy finding illustrates how explainability can serve fairness. A device-price variable is a plausible predictor — device value correlates with income — but it rewards wealth rather than conduct and so contradicts the rationale for behavioural scoring. SHAP made the proxy visible, and its removal cost almost nothing in accuracy. This supports the argument that fairness improvements are often available at modest cost when the right features are targeted (Kozodoi et al., 2022), and that explanation methods should be used not only to justify individual decisions but also to audit model design.

The qualitative findings show why technical fixes are insufficient. Experts were nearly unanimous that consent, transparency, and contestability determine whether behavioural scoring will be trusted, and they identified institutional obstacles — algorithm aversion among loan officers, regulatory uncertainty, and inconsistent data standards — that fairness metrics cannot reveal. The convergence of quantitative evidence of residual bias and qualitative concern about governance strengthens the conclusion that fairness must be managed continuously rather than certified once.

Which fairness criterion should govern Indian lending? Disparate impact is intuitive and aligned with the inclusion objective of public policy, but it can be satisfied by a model that still misjudges creditworthy members of disadvantaged groups. Equalized odds, and in particular equality of true positive rates — what Hardt et al. (2016) call equality of opportunity — directly expresses the idea that a borrower who will repay should have the same chance of approval whoever they are. Because the two criteria cannot both be perfectly satisfied when base rates differ (Chouldechova, 2017; Kleinberg et al., 2017), the choice is normative. We argue that, for inclusion-oriented lending, true-positive-rate parity deserves priority, with disparate impact as a complementary screen, and that the trade-off should be set by explicit policy rather than left to model defaults (Corbett-Davies et al., 2017).

The findings also illuminate a three-way tension among accuracy, fairness, and interpretability (Rudin, 2019). Gradient boosting is the most accurate and, on disparate impact, the fairest model, but the least interpretable; yet it was interpretability tools that revealed its proxy problem. This suggests that the tension is not fixed: explainability can be the means by which fairness is improved, provided it is used for auditing rather than only for justifying decisions.

6. A Fairness Audit Framework and Policy Implications

6.1 Proposed Fairness Audit Indicators

Drawing on the results, Table 11 proposes a set of indicators that lenders could report and supervisors could review before deployment and after every retraining. The thresholds are proposed starting points for discussion with regulators rather than settled standards.

Table 11. Proposed fairness audit indicators for behavioural credit scoring

Indicator	Definition	Proposed threshold
Disparate impact ratio	Approval rate of group ÷ most favoured group	≥ 0.80
TPR parity	Difference in true positive rates across groups	≤ 0.05
FPR parity	Difference in false positive rates across groups	≤ 0.05
Predictive parity	Difference in positive predictive value across groups	≤ 0.05
Calibration equity	Difference in calibration slopes across groups	Not significant ($p > 0.05$)
Proxy feature correlation	Correlation of each input with protected attributes	$r < 0.30$, or justified
SHAP disparity	Contribution of proxy features to group outcomes	No significant contribution

Source: Developed by the authors from the audit results.

6.2 Implications for Lenders

- Measure both outcome fairness and error-rate fairness; disparate impact alone can conceal residual bias.
- Screen every behavioural feature for proxy effects using correlation and SHAP analysis, and remove or justify features that reward wealth rather than conduct.
- Treat groups with thin behavioural traces as higher-uncertainty segments, using human review or additional data rather than automatic rejection.
- Provide applicants with clear reasons and counterfactual guidance, and a route to contest decisions (Wachter et al., 2018).
- Document models and audits in standard form and repeat audits after each retraining (Mitchell et al., 2019; Raji et al., 2020).

6.3 Implications for Regulators

The Reserve Bank of India could complement its digital lending guidelines with explicit expectations for fairness testing of credit algorithms, including the indicators in Table 11 and documentation of feature screening. The regulatory sandbox (Reserve Bank of India, 2019) provides a setting in which such audits can be piloted. Under the Digital Personal Data Protection Act, consent for behavioural scoring should be specific, informed, and revocable, and supported by financial and digital literacy so that it is meaningful (Government of India, 2023; National Centre for Financial Education, 2020). Fairness assessments in India should also consider dimensions such as caste and region, which are not captured by frameworks developed elsewhere, subject to lawful and ethical data collection.

6.4 Implications for Data Infrastructure

Because residual unfairness arises partly from thinner behavioural traces for some groups, technical fixes inside models are insufficient. Extending the Account Aggregator framework, with consent, to additional data relevant to credit — such as utility and rent payments — could enrich the data available for under-observed groups (Reserve Bank of India, 2016; Carrière-Swallow et al., 2021). Industry data standards for behavioural data would reduce noise arising from inconsistent provider formats, which respondents identified as a practical obstacle. Digital inclusion programmes that close the gender gap in smartphone ownership and mobile internet use (GSMA, 2024) would, over time, also reduce the informational disadvantage that drives error-rate disparities.

7. Limitations and Future Research

The audit covers gender, employment type, and city; caste, religion, and disability could not be examined because the data did not record them. Results are reported for a single operating point and may differ at other risk thresholds. The data are secondary institutional records, and fairness outcomes are estimated counterfactually rather than observed in live lending. The qualitative sample comprises experts rather than borrowers. Future research should audit behavioural models on live lending data, examine additional protected attributes and intersectional groups, evaluate in-processing and post-processing mitigation methods, and study borrowers' own perceptions of fairness and consent.

8. Conclusion

Behavioural credit scoring can make lending in India both more inclusive and, in important respects, fairer than conventional scoring: it narrows disparities in approval rates across gender, occupation, and city. But it is not yet fair on every measure. Creditworthy women and daily wage workers remain less likely to be recognised, partly because behavioural data describe their lives less fully and partly because

some features act as proxies for wealth. These residual gaps can be reduced through proxy screening and continuous auditing, but closing them will also require better data infrastructure, meaningful consent, and regulatory standards for algorithmic fairness. Inclusion is the promise of behavioural scoring; fairness is the condition that must accompany it.

Declarations

Ethics approval. The study was approved by the Institutional Ethics Committee of NIILM University, Kaithal. Behavioural data were anonymised by the providing institutions before transfer; no directly identifying information was accessible to the researchers.

Informed consent. Written informed consent was obtained from all interview participants. Secondary records were used under data-sharing agreements that permit research use of anonymised data.

Conflict of interest. The authors declare no conflict of interest.

Funding. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability. The borrower-level data were obtained from institutional partners under confidentiality agreements and cannot be shared publicly. Aggregated results and analysis code are available from the authors on reasonable request.

Author contributions (CRediT). T. Lokendran: conceptualisation, data curation, formal analysis, investigation, methodology, software, visualisation, writing – original draft. Kratika Saxena: conceptualisation, methodology, supervision, validation, writing – review and editing.

Originality and plagiarism. This manuscript is based on the first author's doctoral research at NIILM University. It has not been published or submitted elsewhere.

References

[1] Agarwal, S., Alok, S., Ghosh, P., & Gupta, S. (2020). *Financial inclusion and alternate credit scoring: Role of big data and machine learning in fintech* (SSRN Working Paper No. 3507827). https://doi.org/10.2139/ssrn.3507827

[2] Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*(3), 671–732. https://doi.org/10.15779/Z38BG31

[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. https://doi.org/10.1016/j.jfineco.2021.05.047

[4] 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

[5] Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

[6] Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics, 57*(1), 203–216. https://doi.org/10.1007/s10614-020-10042-0

[7] Carrière-Swallow, Y., Haksar, V., & Patnam, M. (2021). *India's approach to open banking: Some implications for financial inclusion* (IMF Working Paper No. 21/52). International Monetary Fund.

[8] Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research, 11*, 2079–2107.

[9] 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). ACM. https://doi.org/10.1145/2939672.2939785

[10] Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153–163. https://doi.org/10.1089/big.2016.0047

[11] Citron, D. K., & Pasquale, F. (2014). The scored society: Due process for automated predictions. *Washington Law Review, 89*(1), 1–33.

[12] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 797–806). ACM. https://doi.org/10.1145/3097983.3098095

[13] Cornelli, G., Frost, J., Gambacorta, L., Rau, P. R., Wardrop, R., & Ziegler, T. (2023). Fintech and big tech credit: Drivers of the growth of digital lending. *Journal of Banking & Finance, 148*, Article 106742. https://doi.org/10.1016/j.jbankfin.2022.106742

[14] Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE.

[15] Demirgüç-Kunt, A., Klapper, L., Singer, D., & Ansar, S. (2022). *The Global Findex Database 2021: Financial inclusion, digital payments, and resilience in the age of COVID-19*. World Bank. https://doi.org/10.1596/978-1-4648-1897-4

[16] Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126. https://doi.org/10.1037/xge0000033

[17] 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). ACM. https://doi.org/10.1145/2090236.2090255

[18] European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation). *Official Journal of the European Union, L 119*, 1–88.

[19] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 259–268). ACM. https://doi.org/10.1145/2783258.2783311

[20] Fetters, M. D., Curry, L. A., & Creswell, J. W. (2013). Achieving integration in mixed methods designs: Principles and practices. *Health Services Research, 48*(6, Pt. 2), 2134–2156. https://doi.org/10.1111/1475-6773.12117

[21] Frost, J., Gambacorta, L., Huang, Y., Shin, H. S., & Zbinden, P. (2019). BigTech and the changing structure of financial intermediation. *Economic Policy, 34*(100), 761–799. https://doi.org/10.1093/epolic/eiaa003

[22] 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

[23] Gabor, D., & Brooks, S. (2017). The digital revolution in financial inclusion: International development in the fintech era. *New Political Economy, 22*(4), 423–436. https://doi.org/10.1080/13563467.2017.1259298

[24] Government of India. (2005). *The Credit Information Companies (Regulation) Act, 2005* (Act No. 30 of 2005).

[25] Government of India. (2020). *The Code on Social Security, 2020* (Act No. 36 of 2020).

[26] Government of India. (2023). *The Digital Personal Data Protection Act, 2023* (Act No. 22 of 2023).

[27] GSMA. (2024). *The mobile gender gap report 2024*. GSMA.

[28] Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. *Field Methods, 18*(1), 59–82. https://doi.org/10.1177/1525822X05279903

[29] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29* (pp. 3315–3323). Curran Associates.

[30] Hurley, M., & Adebayo, J. (2016). Credit scoring in the era of big data. *Yale Journal of Law and Technology, 18*(1), 148–216.

[31] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)* (Leibniz International Proceedings in Informatics, Vol. 67, pp. 43:1–43:23). Schloss Dagstuhl. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

[32] Kozodoi, N., Jacob, J., & Lessmann, S. (2022). Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research, 297*(3), 1083–1094. https://doi.org/10.1016/j.ejor.2021.06.023

[33] Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. SAGE.

[34] Lokendran, T., & Saxena, K. (n.d.). *Beyond the bureau: Machine learning and behavioural data for inclusive credit scoring of thin-file borrowers in Southern India* [Manuscript in preparation]. Department of Management, NIILM University.

[35] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765–4774). Curran Associates.

[36] 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. https://doi.org/10.1145/3457607

[37] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). ACM. https://doi.org/10.1145/3287560.3287596

[38] National Centre for Financial Education. (2020). *National strategy for financial education 2020–2025*. Reserve Bank of India.

[39] Patton, M. Q. (2015). *Qualitative research & evaluation methods* (4th ed.). SAGE.

[40] Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33–44). ACM. https://doi.org/10.1145/3351095.3372873

[41] Reserve Bank of India. (2016). *Master Direction – Non-Banking Financial Company – Account Aggregator (Reserve Bank) Directions, 2016*.

[42] Reserve Bank of India. (2019). *Enabling framework for regulatory sandbox*.

[43] Reserve Bank of India. (2021). *Report of the Working Group on Digital Lending including lending through online platforms and mobile apps*.

[44] Reserve Bank of India. (2022). *Guidelines on digital lending*.

[45] 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

[46] Shapley, L. S. (1953). A value for n-person games. In H. W. Kuhn & A. W. Tucker (Eds.), *Contributions to the theory of games* (Vol. 2, pp. 307–317). Princeton University Press.

[47] Solove, D. J. (2013). Introduction: Privacy self-management and the consent dilemma. *Harvard Law Review, 126*(7), 1880–1903.

[48] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10*(5), 557–570. https://doi.org/10.1142/S0218488502001648

[49] Tyler, T. R. (2006). Psychological perspectives on legitimacy and legitimation. *Annual Review of Psychology, 57*, 375–400. https://doi.org/10.1146/annurev.psych.57.102904.190038

[50] Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness* (pp. 1–7). ACM. https://doi.org/10.1145/3194770.3194776

[51] 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.