



Scoring Thin-File Borrowers Using Mobile Transaction Data Without Compromising Fairness

Chioma Mengistu^{1*}, Divya Le², Aibek Nasser³

¹Department of Business, Andes Graduate Institute of Business, Jordan

²Department of Business, Transylvanian University Centre for Business, Guyana

³Department of Business, Netravati University Centre for Business, Tanzania

Article Info

Volume 1, Issue 1, 2026

doi: [10.62715/ijbfm.2026.v1.i1.pp21-26](https://doi.org/10.62715/ijbfm.2026.v1.i1.pp21-26)

Received: 03 October 2025

Accepted: 11 December 2025

Published: 07 February 2026

Abstract

Lenders cannot score applicants who have no credit history, and in many markets that describes most adults. Records of mobile money and telephone use offer an alternative basis for assessment, and published studies show that such records predict repayment. The same records also reflect sex, age, location and income, so that a model trained on them can reproduce disadvantage even when protected characteristics are withheld from it. This article proposes a nine-step procedure for building a credit score for thin-file borrowers from mobile transaction data in which fairness is examined at each stage instead of being checked once at the end. The steps cover the definition of the population and outcome, consent and data minimisation, feature construction with screening for proxies, model fitting against a transparent benchmark, treatment of rejected applicants, the choice of a fairness criterion, a group-wise audit, mitigation, and explanation with monitoring after release. The procedure draws on the literature on credit scoring and on algorithmic fairness, including the result that the common fairness criteria cannot all be met together, which obliges the lender to choose and to justify its choice. The procedure is a proposal. It has not been validated on lending data here, and the article states what would have to be demonstrated before it is relied upon.

Keywords: Credit scoring, Thin-file borrowers, Mobile money data, Algorithmic fairness, Alternative data, Financial inclusion

© 2026 Chioma Mengistu et al. This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

1. Introduction

A credit bureau can say nothing about a person who has never borrowed. Such applicants, described in the industry as thin-file or no-file, are refused or are offered credit on terms set for the worst case. The statistical methods of credit scoring are mature (Hand and Henley, 1997), but they need inputs, and the conventional inputs are missing for exactly the borrowers whom inclusion policy most wants to reach.

Mobile telephones generate a substitute. Where mobile money is common, a person's wallet records deposits, withdrawals, transfers, bill payments and airtime purchases, often over several years. Björkegren and Grissen (2020) showed that features computed from telephone use predicted repayment

* Corresponding author: Chioma Mengistu, Department of Business, Andes Graduate Institute of Business, Jordan. E-mail: chioma.mengistu@andes.edu

© 2026 Chioma Mengistu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

among borrowers with no formal financial history, and [Berg et al. \(2020\)](#) found that simple digital footprints carried as much information about default as a bureau score. Call records can also be used to predict wealth ([Blumenstock et al., 2015](#)), and that finding is the source of the difficulty. Data rich enough to reveal creditworthiness are rich enough to reveal who a person is.

This article proposes a procedure for developing a score for thin-file applicants from mobile transaction data with fairness built into each step. It is addressed to the model developers and risk managers of a lender or mobile money provider. The procedure is assembled from published methods and is offered as a proposal: it has not been applied to lending data for this article and no performance figures are claimed for it.

2. Background

Two literatures meet in this problem. One concerns what alternative data can do for prediction, and the other what it means for a score to be fair.

2.1. What can go wrong

[Barocas and Selbst \(2016\)](#) catalogue the ways in which a model built from data can disadvantage a protected group without any intention to do so: the outcome may be defined in a way that favours one group, the training sample may under-represent another, the labels may carry past prejudice, and innocuous features may stand in for the protected attribute. All four arise with mobile data. Women in many markets own fewer handsets and transact through a relative's account. Rural users have fewer agents nearby and so fewer cash-in events. A feature such as the share of transactions made late at night may separate occupations, and through them sex or ethnicity, more than it separates good payers from bad.

Evidence from mortgage lending shows that these concerns are real. [Fuster et al. \(2022\)](#) found that moving from a logistic model to a more flexible machine learning model changed predicted default risk unevenly across racial groups, and [Bartlett et al. \(2022\)](#) found that algorithmic lenders still charged minority borrowers more, though by a smaller margin than face-to-face lenders. Removing the protected attribute from the inputs does not solve the problem, since flexible models reconstruct it from correlated features.

2.2. Definitions of fairness and their conflict

Three families of criteria dominate the literature. Demographic parity asks that approval rates be equal across groups. Equality of opportunity asks that applicants who would repay have the same chance of approval whatever their group, and its stricter form, equalised odds, adds the same requirement for those who would default ([Hardt et al., 2016](#)). Calibration within groups asks that a given score mean the same probability of default for every group.

These criteria cannot all hold at once. [Kleinberg et al. \(2017\)](#) and [Chouldechova \(2017\)](#) proved that when default rates differ between groups, a score that is calibrated within each group must produce unequal error rates, except in degenerate cases. A lender therefore has to select a criterion, and the selection is a judgement about which kind of error matters most. [Kozodoi et al. \(2022\)](#) examined the criteria in the setting of credit scoring and argued that a single well-chosen criterion can usually be approached at a moderate cost in profit.

3. The Proposed Procedure

The procedure has nine steps, listed in Table 1 with the fairness safeguard attached to each. They follow the ordinary sequence of scorecard development, so that a team can adopt them without reorganising its work.

3.1. Steps 1 and 2: Scope, consent and data

The developer first states who is to be scored, for which product, and what counts as default. A definition based on days past due should be checked for groups whose income is seasonal, since a farmer paid at

harvest may be late by the calendar and reliable over the season. Consent must be specific to credit assessment, and the data taken should be limited to transaction metadata: amounts, times, types and counterparties in hashed form. Message content and contact lists are excluded. Protected attributes are needed to test fairness even though they are not used for scoring. They should be collected with a separate explanation, stored apart from the modelling data and made available only to those who run the audit.

Table 1: The nine steps of the proposed procedure and the fairness safeguard attached to each		
Step	Development task	Fairness safeguard
1. Scope	Define the population, product and default outcome	Check that the outcome definition does not favour a group
2. Consent and data	Obtain consent; fix the fields and history used	Collect protected attributes separately for audit only
3. Features	Build behavioural features from transactions	Screen each feature for association with protected groups
4. Model	Fit a scorecard benchmark and a boosted model	Keep the simpler model unless the gain is material
5. Rejects	Address applicants with no observed outcome	Examine who was excluded from the training sample
6. Criterion	Choose and record the fairness criterion	Decision approved by risk and compliance functions
7. Audit	Measure approval, error and calibration by group	Report intervals; flag small groups
8. Mitigation	Apply the mildest effective intervention	Re-audit after each change
9. Release	Explain decisions; monitor after launch	Track outcomes by group over time

3.2. Step 3: Features and proxy screening

Features should describe financial behaviour that has a plausible connection with repayment: the regularity of inflows, the ratio of outflows to inflows, the lowest balance in a month, the frequency of bill payment, the length of account history. Each candidate is then tested for its association with the protected attributes. A feature that predicts group membership strongly and adds little to the prediction of default once other features are present is removed. A feature that does both is retained only with a recorded business justification. Features that describe the applicant's contacts, and not the applicant, are best left out, since they score people by the company they keep.

3.3. Steps 4 and 5: Model fitting and rejected applicants

Two models are fitted. The first is a logistic scorecard, whose coefficients can be read and defended. The second is a gradient-boosted tree model of the kind implemented in XGBoost (Chen and Guestrin, 2016). Benchmarking studies find that such ensembles outperform logistic regression in credit scoring, though the margin varies (Lessmann et al., 2015). The flexible model is adopted only if its advantage on held-out data is large enough to matter commercially, because flexibility is also what allows a model to rebuild protected attributes from proxies. Outcomes are observed only for applicants who were granted credit in the past, and those applicants were selected by an earlier rule. Reject inference techniques try to correct for this, with limited success (Crook and Banasik, 2004). Where the lender can afford it, the sounder remedy is to approve a small random sample below the cut-off and observe the result.

3.4. Steps 6 and 7: Choosing a criterion and auditing

Because the criteria conflict, the choice among them is made explicitly and recorded before the audit

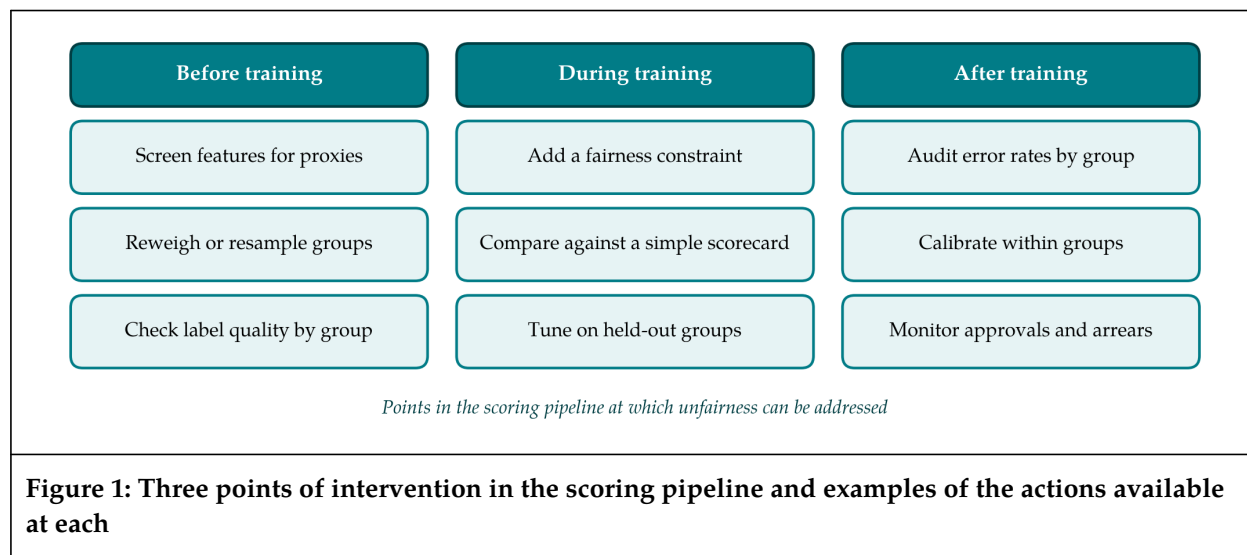
results are seen. For a product intended to widen access, equality of opportunity has a clear rationale: it asks that creditworthy applicants in every group be equally likely to be approved. Calibration within groups should be checked in every case, since a score that overstates risk for one group overcharges it. The audit then computes, for each group and for intersections where numbers allow, the approval rate, the true and false positive rates and a calibration curve, each with a confidence interval. Small groups produce wide intervals, and the audit should say so and avoid reporting a difference as absent merely because it is not statistically detectable.

3.5. Step 8: Mitigation

If the audit shows a disparity beyond a tolerance set in advance, the developer intervenes at one of the three points shown in Figure 1. Before training, the sample can be reweighted so that group and outcome are independent (Kamiran and Calders, 2012). During training, a constraint can be imposed through the reductions method of Agarwal et al. (2018). After training, cut-offs can be adjusted by group to equalise the chosen error rate (Hardt et al., 2016). The last is the most direct and also the most legally sensitive, because it treats applicants differently by group on the face of the rule, and in some jurisdictions that is prohibited. The procedure therefore prefers the earliest and mildest intervention that brings the disparity within tolerance, and repeats the audit after each change.

3.6. Step 9: Explanation and monitoring

An applicant who is refused is entitled in many jurisdictions to the principal reasons, and reasons are also what allow an error to be challenged. For the scorecard they follow from the points lost on each characteristic. For the boosted model, additive attributions such as SHAP values provide a comparable breakdown (Lundberg and Lee, 2017), which should be translated into statements the applicant can act upon. After release the lender tracks approvals, arrears and calibration by group at regular intervals. Mobile money behaviour shifts with tariffs, agent coverage and seasons, so a model that was fair at launch may drift.



4. Fitting the Procedure to Regulation and Practice

The procedure is compatible with the main regulatory approaches and does not replace legal advice on any of them. Laws against discrimination in lending generally prohibit both different treatment and unjustified adverse effect, and the proxy screening and audit steps produce the evidence that either test requires. The European Union's Artificial Intelligence Act classes systems used to evaluate the creditworthiness of natural persons as high-risk and attaches duties of data governance, documentation and human oversight (European Parliament and Council, 2024). The records created at each step map onto those duties. In markets without such legislation, the same records serve the lender's own model risk governance. A small lender can run the whole procedure with open software, and the principal cost is the

collection of protected attributes for audit under proper safeguards.

5. Limitations and What Remains to Be Shown

The procedure is untested, and several of its premises may fail in practice. Protected attributes may be unavailable or unlawful to collect, in which case the audit depends on inferred group membership and inherits its errors. Mobile data describe only those who use mobile money, so that the people with the thinnest files of all remain unscored. The fairness criteria considered here are statistical and concern groups; they do not guarantee that any individual is treated justly, and they say nothing about whether the loan itself benefits the borrower.

Validation would require a lender's data and would need to establish three things: that the screened feature set predicts repayment out of time and not merely in the development sample, that the chosen mitigation reduces the measured disparity on new applicants, and that the cost in predictive accuracy is one the lender can bear. Until such evidence exists, the procedure should be treated as a structured checklist and not as a tested method.

6. Conclusion

Mobile transaction records can make thin-file borrowers visible to lenders, and they bring with them the risk that a score will sort applicants by who they are. The nine-step procedure proposed here places a fairness safeguard beside each ordinary task of score development, requires an explicit choice among criteria that cannot all be satisfied, and prefers transparent models and mild interventions. It is offered for testing by lenders and researchers who hold the data needed to judge it.

Conflict of interest

To be declared by the author(s) before publication.

Funding

To be declared by the author(s) before publication.

References

- Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J. and Wallach, H. (2018). [A reductions approach to fair classification](#). *Proceedings of the 35th International Conference on Machine Learning*, PMLR 80: 60-69.
- Barocas, S. and Selbst, A.D. (2016). [Big data's disparate impact](#). *California Law Review*, 104(3): 671-732.
- Bartlett, R., Morse, A., Stanton, R. and 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>
- Berg, T., Burg, V., Gombović, A. and 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>
- Björkegren, D. and Grissen, D. (2020). [Behavior revealed in mobile phone usage predicts credit repayment](#). *The World Bank Economic Review*, 34(3): 618-634. <https://doi.org/10.1093/wber/lhz006>
- Blumenstock, J., Cadamuro, G. and On, R. (2015). [Predicting poverty and wealth from mobile phone metadata](#). *Science*, 350(6264): 1073-1076. <https://doi.org/10.1126/science.aac4420>
- Chen, T. and Guestrin, C. (2016). [XGBoost: A scalable tree boosting system](#). *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. <https://doi.org/10.1145/2939672.2939785>
- 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>
- Crook, J. and Banasik, J. (2004). [Does reject inference really improve the performance of application scoring models?](#) *Journal of Banking & Finance*, 28(4): 857-874.
- European Parliament and Council (2024). [Regulation \(EU\) 2024/1689 laying down harmonised rules on](#)

- artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L series, 12 July 2024.
- Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T. and Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. *The Journal of Finance*, 77(1): 5-47. <https://doi.org/10.1111/jofi.13090>
- Hand, D.J. and Henley, W.E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160(3): 523-541. <https://doi.org/10.1111/j.1467-985X.1997.00078.x>
- Hardt, M., Price, E. and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29.
- Kamiran, F. and Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1): 1-33. <https://doi.org/10.1007/s10115-011-0463-8>
- Kleinberg, J., Mullainathan, S. and Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. *Proceedings of the 8th Innovations in Theoretical Computer Science Conference*, LIPIcs 67: 43:1-43:23.
- Kozodoi, N., Jacob, J. and 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>
- Lessmann, S., Baesens, B., Seow, H.-V. and Thomas, L.C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1): 124-136. <https://doi.org/10.1016/j.ejor.2015.05.030>
- Lundberg, S.M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.