

Thin-file, Not Signal-free: An Alternative-data Credit Score for Smallholder Farmers Built with Explainable Machine Learning

Anil Kumar Jonnalagadda^{1,2*}, Narasimha Rao MVAL¹, Ramesh Babu S³

¹Department of MBA, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India, ²School of Management, Centurion University of Technology and Management, Vizianagaram, Andhra Pradesh, India, ³Department of Management Sciences, R.V.R. & J.C. College of Engineering (Autonomous), Chowdavaram, Guntur, Andhra Pradesh, India.
*Corresponding Author's Email: janilkumar@yahoo.com

Abstract

The majority of Indian small farmers belong to the category of thin-file customers who fail to meet the requirements for the payment history needed for credit bureau ratings and thus alternative data credit rating has been recommended as a solution. The practice itself is plagued by issues with measurements, since the accuracy is exaggerated due to resampling before the train-test splitting. This study addresses both gaps. We surveyed 265 small and marginal farmers across seven districts of North Coastal Andhra Pradesh and built a two-stage model on the 4C framework, comprising 22 ordinal items across four constructs. Oversampling was confined to the training partition of each fold. Under this protocol, CatBoost reached a Receiver Operating Characteristic - Area Under the Curve (ROC-AUC) of 0.783 for credit access and 0.819 for repayment, with Matthews correlation coefficients of 0.363 and 0.474 respectively. When the same model was trained on data oversampled before splitting, it reported 0.956 for access, an inflation of 0.173 AUC units arising from protocol alone. SHapley Additive exPlanations (SHAP) ranks were then tested across ten independent resamples: crop insurance enrolment ranked first for access in every resample, with a rank standard deviation of zero, while bank distance and farm location led the repayment stage. A four-tier score constructed from score quartiles showed a 17.2 percentage-point access gap between the top and bottom bands.

Keywords: Alternative Credit Scoring, CatBoost, Explainable AI, SHAP, Thin-file Borrowers.

Introduction

Credit bureaus score the past. A borrower who has repaid a formal loan leaves a record and that record renders the borrower legible to the next lender. A borrower without such a record is not scored badly; that borrower is not scored at all. The distinction is consequential, because a poor score can be priced through a higher interest rate, whereas an absent score usually terminates the application altogether. Thin-file exclusion is therefore not a matter of unfavourable terms but of categorical unavailability and it is reproduced at every subsequent loan cycle, since a borrower who is never lent to never generates the record that would permit future lending. For Indian smallholders this absence is the binding constraint on formal agricultural finance (1, 2). Small and marginal farmers operate the majority of the country's operational holdings, yet only a minority hold a formal credit history (3). Lenders consequently fall back on land title, physical collateral and branch-level discretion, each of

which reproduces the very exclusion it is intended to price (4, 5). The result is a population that is large, economically active and yet almost entirely invisible to conventional scoring infrastructure. Alternative-data credit scoring offers a different route, replacing the missing repayment history with observable behavioural and contextual signals. Berg and colleagues demonstrated that a small set of digital footprint variables can equal or exceed bureau scores in predicting consumer default (6). A past research showed that mobile phone usage patterns predict repayment among borrowers with no formal history and Óskarsdóttir and colleagues found that call-network features add discriminatory power in financially excluded populations (7, 8). Another group reported comparable gains from alternative data in fintech lending and a recent review documents a rapidly expanding body of machine-learning applications in rural finance (9, 10). Related literature indicates that the adoption of BigTech platforms has

This is an Open Access article distributed under the terms of the Creative Commons Attribution CC BY license (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution and reproduction in any medium, provided the original work is properly cited.

(Received 16th July 2026; Accepted 11th October 2026; Published 22nd October 2026)

facilitated the availability of data for credit assessments (11). Predictive variables based on psychometric scores and usage data were effective in distinguishing between good and bad borrowers amongst people who lacked any previous history (12). Digital financial inclusion may result in differential outcomes across gender and socio-economic groups (13). Credit constraints can similarly impede productive technology uptake among poor households (14).

Two limitations recur across this literature. The first is protocol-driven optimism. Credit-access data are typically imbalanced, since most surveyed farmers have borrowed and class imbalance is conventionally corrected through minority oversampling (15). Oversampling generates synthetic observations by interpolating between existing neighbours, so the order of operations matters: if oversampling precedes partitioning, synthetic observations and the real observations from which they were derived can fall into different subsets and the model is then evaluated partly on information it has already seen. The resulting performance figure is no longer an out-of-sample estimate and the bias it introduces is both large and, in the absence of an explicit protocol statement, undetectable by the reader (16-18).

The second limitation concerns explanations that are never validated. Explainable artificial intelligence is now routine in credit scoring and justifiably so, since a score that a loan officer cannot interpret is unlikely to satisfy a regulator (19, 20). Yet SHapley Additive exPlanations (SHAP) importance is ordinarily computed from a single model fitted to a single data split. Tree-based models are sensitive to resampling and validation estimates are themselves unstable at limited sample sizes (21, 22). A ranking derived from one partition may therefore reflect the idiosyncrasies of that partition rather than any durable property of the data and a policy recommendation built on such a ranking inherits that fragility without disclosing it.

These two limitations have not, to our knowledge, been addressed jointly and neither has been examined in the population where the consequences are most acute. No existing study quantifies protocol-induced optimism on Indian smallholder field data while holding every other analytical choice constant and none reports the

stability of SHAP rankings in agricultural credit scoring. That is the gap this paper addresses.

The overarching aim of this study is to establish a defensible benchmark for alternative-data credit scoring among thin-file smallholder farmers and to determine which of its explanations are robust enough to inform policy. Three specific objectives follow:

- (a) To assess, under a leakage-free estimation protocol, how well alternative data predict smallholder credit access and repayment.
- (b) To quantify the optimism introduced by oversampling before partitioning, holding data, features, hyperparameters and random seed constant.
- (c) To identify the determinants that retain stable importance ranks across independent resamples and which are artefacts of a single partition.

The novelty of the study is threefold. It provides the first quantification, on Indian smallholder field data, of resampling-induced optimism with every other analytical choice held constant, yielding an inflation of 0.173 ROC-AUC units at the credit-access stage that is reported here as a methodological finding in its own right rather than as a caveat. It is the first application of SHAP rank-stability testing in agricultural credit scoring, demonstrating that an apparently informative predictor can carry a rank standard deviation of 4.84 across resamples and should not inform policy. And it introduces the 4C instrument, a 22-item field-administrable alternative-data schema for thin-file agricultural borrowers, together with construct-utilization diagnostics that identify which forms of capital are binding.

Methodology

Research Approach and Design

Rationale

Three linked design choices require justification. First, a cross-sectional primary survey was necessary rather than convenient. No bureau file, lender panel or administrative dataset exists for thin-file smallholder borrowers; the absence of secondary data is precisely the condition this study investigates, so primary collection was the only available route to the population of interest. Second, supervised gradient boosting was preferred to conventional econometric estimation because the object of the study is out-of-sample

discrimination rather than coefficient inference and because the ordinal 4C items relate to credit outcomes through non-linear and interactive effects that a linear specification would not recover. Penalised logistic regression is retained throughout as an interpretable benchmark, consistent with established credit-scoring benchmarking practice (23). Third, post-hoc SHAP explanation was adopted rather than reliance on an intrinsically interpretable model alone, so that the interpretive cost of the boosted model is made explicit and auditable rather than assumed away; the stability testing reported below is the mechanism by which that cost is assessed.

Study Area

Data come from structured interviews with 265 small and marginal farmers. The study area covers seven districts of North Coastal Andhra Pradesh, namely Srikakulam (18.3092485° N, 83.8932628° E), Anakapalli (17.6860° N, 83.0088° E), Vizianagaram (18.1083602° N and 83.3988236° E), Alluri Sitharama Raju (18.08779° N, 82.668° E), Parvathipuram Manyam (18.77372° N, 83.42296° E), Visakhapatnam (17.7238° N, 83.3073° E) and East Godavari (17.00234° N, 81.78985° E). The region combines rainfed agriculture, low institutional credit penetration and rising digital adoption (24). These seven districts were selected for four reasons. The zone carries among the highest concentrations of small and marginal operational holdings in the state, so the sampling frame is dense in the population of interest. It combines predominantly rainfed cultivation with below-state-average institutional credit penetra-

tion, which is the condition under which thin-file alternative scoring is relevant; districts with high formal credit penetration would yield comparatively few thin-file cases. The seven districts form a contiguous agro-climatic zone with comparable cropping calendars and broadly common lending practice, which holds constant sources of variation that would otherwise confound a multi-region sample. Finally, the zone spans a genuine infrastructure gradient, from the coastal plain to the agency areas of Alluri Sitharama Raju and Parvathipuram Manyam and this internal variation is what permits distance-to-branch effects to be identified at all. District selection was thus purposive at the zone level, while respondent selection within the zone was district-stratified random. Village-level lists of cultivators holding under two hectares formed the sampling frame and allocation was proportional to each district's share of small and marginal holdings. The breakdown of respondents across districts is provided in Table 1 below. Of the total 265 respondents, 198 hailed from Srikakulam (74.7%), 24 from Anakapalli (9.1%), 21 from Vizianagaram (7.9%) and five to seven respondents from each of the other four districts. Such disproportionate representation arises from the predominance of small and marginal farmers in this particular district relative to the total study area. Maintaining this sample allocation instead of equal distribution ensures alignment with the demographics under study. All interviews took place using Telugu-language communication facilitated by enumerators.

Table 1: District-wise Distribution of Respondents (n = 265)

District	Respondents (n)	Share (%)
Srikakulam	198	74.7
Anakapalli	24	9.1
Vizianagaram	21	7.9
Alluri Sitharama Raju	7	2.6
Parvathipuram Manyam	5	1.9
Visakhapatnam	5	1.9
East Godavari	5	1.9
Total	265	100.0

Note: n = Number of Respondents. Column percentages represent proportions based on the overall sample size and do not necessarily add up to 100.0 because of rounding errors.

Inclusion, Exclusion Criteria and Data Collection Procedure

To be eligible for the study, cultivators needed to own an operational holding not exceeding two hectares, cultivate crops in the reference

agricultural year, reside within one of the selected villages, have attained the age of 18 years and give their informed consent. Non-inclusion criteria of cultivators include failure to have documented operational holdings, incomplete responses for any of the 22 items in the tool and inability to indicate

whether they received loans. In total, 300 cultivators were approached, of whom 15 declined to participate and 20 were excluded for incompleteness or inability to report the outcome, yielding the analytic sample of 265.

Data collection proceeded as follows. Enumerators were recruited locally and trained over 7 days in the instrument, in consent procedures and in the coding of ordinal responses. The instrument was administered orally in Telugu, the respondents' first language, following translation from English and independent back-translation to verify equivalence. A pilot with 25 respondents preceded the main survey and led to 5 item rewordings. Interviews lasted 24 minutes on average and were conducted at the respondent's residence or field. A field supervisor accompanied each enumerator

team and 5% of completed schedules were back-checked by telephone. Responses were coded to their respective ordinal scales according to written anchoring rules fixed before fieldwork.

The 4C Instrument

Indicators of creditworthiness were measured using 22 variables. Ten items are scored on four points, five on three and seven dichotomously; Table 2 reports the attainable range of each construct. These variables comprise four constructs: C1, Connectivity and Digital Capital; C2, Contextual Socio-demographic Capital; C3, Cultivation and Farm Capacity; and C4, Credit and Financial Behaviour. Table 2 sets out the composition of the constructs and the utilization of their scores.

Table 2: Composition and Utilization of the 4C Constructs (n = 265)

Construct	Items	Possible Range	Observed Range	Mean (SD)	Range Utilized (%)
C1: Connectivity and Digital Capital	7	7–20	8–20	15.39 (2.69)	64.6
C2: Contextual Socio-demographic Capital	5	5–20	8–19	13.48 (2.22)	56.5
C3: Cultivation and Farm Capacity	4	4–16	6–16	11.56 (2.35)	63
C4: Credit and Financial Behaviour	6	6–17	9–16	12.14 (1.11)	55.8
Overall alternative credit score	22	22–73	39–65	52.57 (5.14)	59.9

Note: SD = Standard Deviation.

Measures in Table 2 were assessed based on ordinal scales of different granularities – ten measures using four points, five measures with three points and seven dichotomous measures. The range adopted was obtained via mean–minimum possible score ÷ maximum possible score–minimum possible score formula, eliminating the artificial imposition of zero lower bound which cannot be achieved by the tool. Adoption rates of constructs C4 (55.8%) and C2 (56.5%) reveal that Credit and Financial Behaviour and Contextual Socio-Demographic Capital represent binding constraints, respectively. In contrast, construct C1 is relatively the best-used measure at 64.6%.

Table 2 reveals that the selected farmers have high productive capabilities and connectivity levels. From an analysis of the construct utilization, Cultivation and Farm Capacity and Connectivity and Digital Capital were the best-used constructs at 63.0% and 64.6% of maximum capacity. However, the weakest construct was Credit and Financial Behaviour at 55.8%, incorporating aspects such as insurance subscription, borrowing from financial institutions and transaction

documentation. Therefore, the constraint affecting these farmers was documentary rather than productive, matching the evidentiary standards of traditional credit scoring methodologies.

Ethical Clearance

Ethical considerations in this study included securing approval from the Institutional Ethics Committee of the Koneru Lakshmaiah Education Foundation prior to any data collection, along with obtaining verbal consent from each participant involved in the project. Full documentation can be found within the Ethics Approval section.

Dependent Variables

A two-stage approach was adopted in order to separate access from repayment, since the two decisions are distinct and are taken by different parties. In the first stage, access to crop loans is modelled as a binary outcome across the full sample of 265 farmers, of whom 203 (76.6%) obtained loans and 62 (23.4%) did not. In the second stage, attention turns to repayment performance among borrowers. The original four-level repayment scale is collapsed into two groups: good repayment comprises categories 3 and 4 (n = 87, 42.9%) and poor repayment comprises

categories 1 and 2 ($n = 116$, 57.1%). The second stage therefore analyses 203 farmers.

Leakage-free Estimation Protocol

The placement of the oversampling step is central to the methodology of this paper. All 22 variables were carried into the analysis and no feature selection was performed on the full sample, since selection on the full sample is itself a form of leakage.

The protocol is as follows. Repeated stratified k-fold cross-validation is employed with five folds and six repetitions, yielding 30 model fits. To prevent overly optimistic results, oversampling is applied exclusively within each training split (18). The Synthetic Minority Over-sampling Technique is employed inside the imbalanced-learn pipeline, fitted on each training split only, so that it never observes the holdout fold and the class balance of the validation fold is left unchanged (15). Two models are fitted. Penalised logistic regression serves as an interpretable benchmark, following credit-scoring benchmarking practice (24), fitted with L2 regularisation. (23). CatBoost is the second model, configured with 400 iterations, maximum depth 4, learning rate 0.05 and L2 leaf regularisation 6 (25). Hyperparameter tuning was not performed. The penalised logistic regression was fitted with L2 regularisation; the CatBoost configuration is as stated above. Five performance measures are reported: ROC-AUC, average precision, the Matthews correlation coefficient, F1 score and accuracy. The Matthews correlation coefficient was included because it evaluates binary classification performance in a balanced manner under class imbalance and average precision because precision-recall measures are informative when assessing classifiers on imbalanced data (26, 27). Each measure is reported as a mean and standard deviation across the 30 fits. A second experiment reproduces the conventional setup, in which CatBoost is trained on data oversampled before splitting and evaluated by five-fold cross-validation. Data, features, hyperparameters and random seed remain identical; the only difference is the placement of the resampling step.

Software and Computational Environment

All statistical tests were performed using Python, Version 3.12 (Python Software Foundation, Beaverton, OR, USA). Methods: Model develop-

ment, validation and evaluation metrics were calculated using scikit-learn, Version 1.8.0 (scikit-learn contributors, open-source). Upsampling and pipeline building were performed using imbalanced-learn, Version 0.14.2 (imbalanced-learn contributors, open-source). Gradient boosting algorithms were computed using CatBoost, Version 1.2.10 (Yandex, Moscow, Russia). Explanatory analyses were conducted using SHAP, Version 0.52.0 (SHAP contributors, open-source). Gradient boosting was performed with CatBoost, Version 1.2.10 (Yandex Inc., Moscow, Russia). SHAP model interpretation used SHAP package, Version 0.52.0 (SHAP developers, open-source software). Data processing involved pandas, Version [3.0.6] and NumPy, Version [2.5.3]; graphs were plotted using Matplotlib, Version [3.11.2] (all projects under NumFOCUS sponsorship). Exact reproducibility was achieved by setting a seed value at 2026 for each of the 30 iterations in the cross-validation fit and protocol comparisons, whereas the independent SHAP samples of size 10 were obtained using seed values between 1000 and 1009.

SHAP Rank-stability Analysis

SHAP values cannot be calculated based on one estimation alone. The mean absolute SHAP value estimates the individual predictor's contribution to the prediction based on changes made to random subsets of predictors to capture interaction effects (28). The following steps were repeated ten times with independent seeds: partition the data; oversample the training partition; fit Cat Boost; and compute mean absolute SHAP values on the held-out partition using Tree Explainer (29). Each feature thereby acquires a distribution of ranks and three quantities are reported: the mean absolute SHAP value, the mean rank and the standard deviation of that rank. A rank standard deviation near zero marks a stable signal, whereas a wide rank spread marks an artefact of partitioning and such features should not inform policy.

Because the reported rank standard deviations are themselves a function of the resampling scheme, that scheme is stated here in full: each repetition draws a stratified held-out partition comprising 20% of the observations, using seeds 1000 to 1009, with the CatBoost configuration and the within-partition oversampling procedure held identical to those described above. A study reporting rank

stability without specifying the held-out fraction and the seeds does not permit its stability estimates to be reproduced and we regard this

specification as part of the reporting standard the present paper argues for.

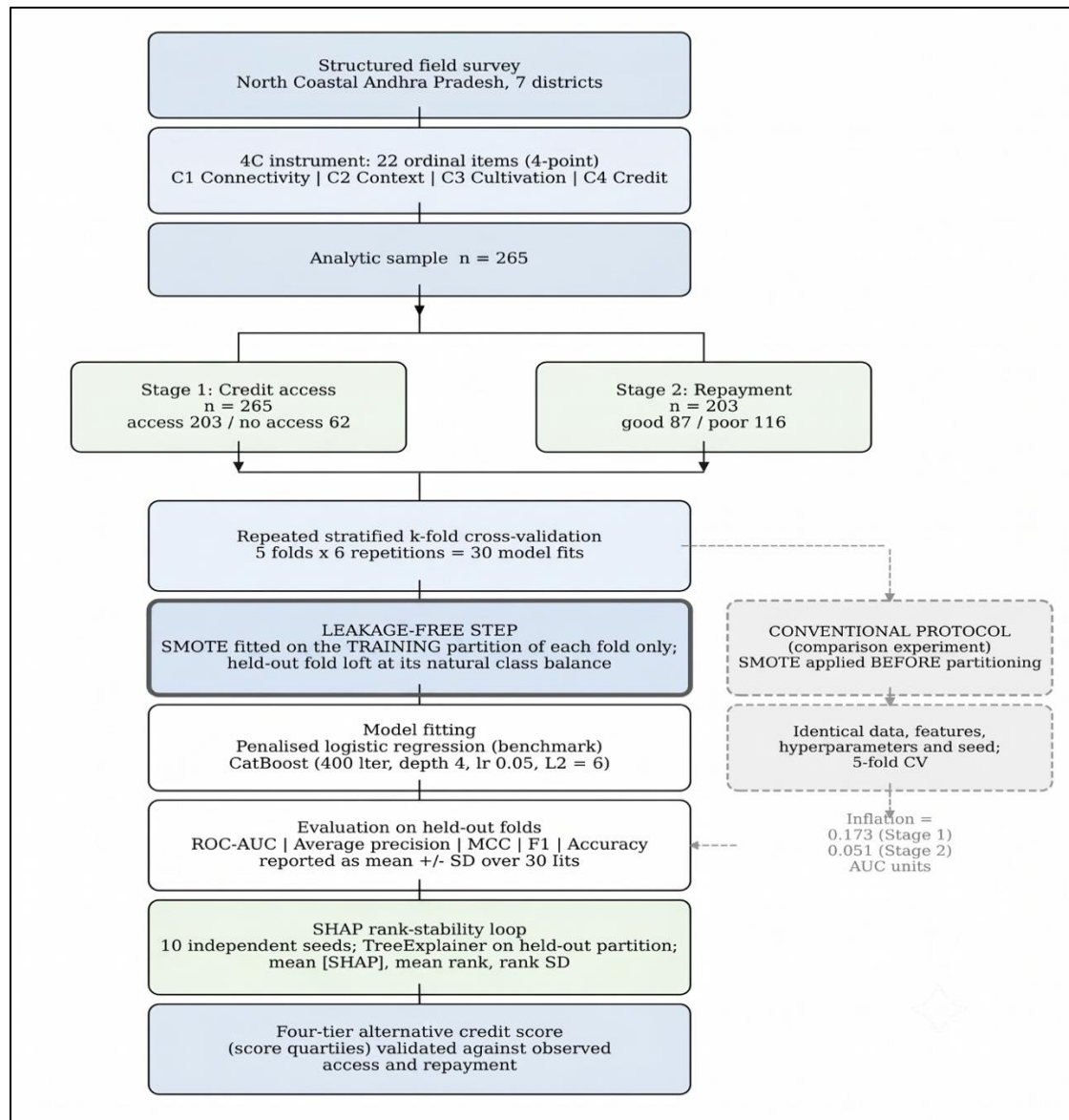


Figure 1: Experimental Design and Leakage-free Estimation Pipeline

Figure 1 should be read from top to bottom along the solid path, which is the protocol adopted in this study. The single highlighted box marks the step on which the paper turns: oversampling is fitted on the training partition of each fold and never sees the held-out fold. The dashed branch to the right reproduces the conventional arrangement, in which oversampling precedes partitioning. Every other element of the two experiments is identical, so the difference in reported performance between the two paths is attributable to that one step alone.

Results

Predictive Performance under the Corrected Protocol

Table 3 presents results for the leakage-free protocol, detailing performance metrics in each stage using cross-validation. In all metrics assessed, CatBoost outperforms penalised logistic regression as a benchmark. ROC-AUC increases from 0.686 to 0.783 and from 0.763 to 0.819 in access and repayment stages, respectively. The increase in Matthews Correlation Coefficient from 0.224 to 0.363 and from 0.373 to 0.474 suggests consistent gains regardless of class imbalance.

Notably, these figures fall substantially below what has typically been cited in prior studies. This

discrepancy motivates comparisons between protocols, presented in Table 4 subsequently.

Table 3: Cross-validated Performance Under Leakage-free Resampling (mean $\pm$ SD over 30 fits)

Stage / Model	ROC-AUC	Avg. precision	MCC	F1	Accuracy
Stage 1: Logistic Regression	0.686 $\pm$ 0.070	0.884 $\pm$ 0.036	0.224 $\pm$ 0.120	0.751 $\pm$ 0.048	0.656 $\pm$ 0.056
Stage 1: CatBoost	0.783 $\pm$ 0.062	0.923 $\pm$ 0.028	0.363 $\pm$ 0.109	0.861 $\pm$ 0.022	0.783 $\pm$ 0.033
Stage 2: Logistic Regression	0.763 $\pm$ 0.063	0.742 $\pm$ 0.066	0.373 $\pm$ 0.119	0.649 $\pm$ 0.066	0.688 $\pm$ 0.061
Stage 2: CatBoost	0.819 $\pm$ 0.059	0.805 $\pm$ 0.059	0.474 $\pm$ 0.120	0.688 $\pm$ 0.076	0.741 $\pm$ 0.056

Note: ROC-AUC = Receiver Operating Characteristic-Area Under the Curve, Avg. Precision = Average Precision, MCC = Matthews Correlation Coefficient, F1 = the Harmonic Mean Between Precision and Recall, SD = Standard Deviation, SMOTE = Synthetic Minority Over-sampling Technique, n = Sample size. Data presented are the means $\pm$ standard deviations derived from 30 runs. SMOTE was trained using the training set folds only. Sample sizes were 265 in Stage 1 and 203 in Stage 2. Majority class accuracy baselines are 0.766 for Stage 1 and 0.571 for Stage 2.

95% confidence intervals, computed as the 2.5th and 97.5th percentiles of the 30-fold-level estimates, are as follows. For CatBoost at Stage 1, ROC-AUC 0.783 [0.632, 0.877] and MCC 0.363 [0.187, 0.526]; at Stage 2, ROC-AUC 0.819 [0.704, 0.933] and MCC 0.474 [0.287, 0.679]. The intervals are wide, as would be expected at this sample size, but both exclude chance performance and the Stage 2 interval for MCC excludes values consistent with a negligible effect. Intervals are reported for CatBoost, the model on which the substantive findings rest.

Two aspects merit comment. First, Stage 2 outperforms Stage 1 on both ROC-AUC and MCC. Repayment is easier to predict than access, which is intelligible: access depends on the lender's decision, which incorporates branch policy and local practice and individual-level indicators cannot observe those factors, whereas repayment follows from the farmer's own conduct. Second, average precision at Stage 1 is 0.923 against a ROC-AUC of 0.783. It is expected that there would be

such a discrepancy since the base rate of the positive class is 76.6% and average precision needs to be understood with respect to that base rate.

Magnitude of Resampling-induced Optimism

Table 4 illustrates the effect of this protocol by evaluating both treatment groups within the experimental framework described in Fig. 1, holding all other variables constant. Within the credit-access phase, the use of oversampling prior to partitioning increased the ROC-AUC result from 0.783 to 0.956, representing an inflation value of 0.173. Within the repayment phase, the corresponding values were 0.819 to 0.870, respectively, indicating an inflation rate of 0.051. These differences arise due to disparities in imbalanced classes wherein Stage 1 contains 62 observations in the minority group and 203 in the majority group, while Stage 2 had comparable proportions of 87:116.

Table 4: CatBoost ROC-AUC under two resampling protocols

Stage	SMOTE Before Split (Conventional)	SMOTE within Training Folds	Inflation
Stage 1: Credit access	0.956	0.783	0.173
Stage 2: Repayment behaviour	0.870	0.819	0.051

Note: SMOTE = Synthetic Minority Over-sampling Technique, ROC-AUC = Receiver Operating Characteristic-Area Under Curve. The inflation represents the disparity between the two experimental procedures measured using ROC-AUC metrics. Specifications for CatBoost are kept consistent across the two sets, barring oversampling prior to vs. after splitting. Stage 1 sample size equals 265, whereas Stage 2 sample size is 203.

At Stage 1, the difference of 0.173 AUC points between the two values is not an artefact of the calculation process but a real difference between the two. While one model seems ready to be deployed, the other model performs considerably better than random but is unable to discriminate

sufficiently well for autonomous decision-making on lending.

The method consists of two stages. At Stage 1, there are only 62 observations of the minority class. By oversampling before cross-validation, new observations are produced by interpolating

observations adjacent to each other and the new observations are assigned to folds together with the original observations, so that observations in the evaluation fold will be based on observations from the training fold. The difference is larger with smaller proportion of the minority class. At Stage 2, the difference between classes becomes smaller because there are 87 versus 116 observations and the effect is reduced to 0.051.

The significance of this finding goes beyond the current data set, where numerous publications show the values of the ROC-AUC higher than 0.90 based on field data with only a couple of hundreds of samples. If the distribution of the class is unbalanced and the point of oversampling is unclear, then such results should not be treated as benchmarks.

SHAP Importance and Rank Stability

The following results are presented in Table 5 and showcase the top five predictors in each phase alongside their stability in terms of ranks across ten repeated sampling cycles. In credit-access, membership in agricultural insurance had an average absolute SHAP score of 1.642, which is approximately three times higher than the next predictor and occupied rank one across all resamples. Repayment exhibited a relatively less skewed ranking order where the top five variables ranged from 0.481 to 0.746 without clear dominance by any predictor. Mean absolute SHAP and stability rank differ in what they capture but should be interpreted jointly; for example, mobile phone recharge type had an average absolute SHAP score of 0.349 but exhibited high rank variability with a standard deviation of 4.84 and featured among the top five in only half the resampling cycles. Figures 2 and 3 visualise these metrics.

Table 5: SHAP Importance and Rank Stability across Ten Resamples

Stage / Feature	Mean SHAP	Mean rank	Rank SD	Top-5 freq.
Stage 1: Crop Insurance (C4)	1.642	1.0	0.00	1.0
Stage 1: Bank Distance (C2)	0.457	3.7	1.74	0.9
Stage 1: Savings Account Tenure (C1)	0.443	4.2	2.18	0.7
Stage 1: Share of Own Land (C3)	0.410	4.8	1.66	0.7
Stage 1: Mode of Mobile Recharge (C1)	0.349	7.5	4.84	0.5
Stage 2: Bank Distance (C2)	0.746	1.9	1.22	1.0
Stage 2: Location of the Farm (C3)	0.593	2.8	1.08	1.0
Stage 2: No. of Dependents (C4)	0.591	3.9	2.07	0.7
Stage 2: Informal Borrowing (C4)	0.534	4.0	2.28	0.7
Stage 2: Agriculture-related Training (C4)	0.481	5.2	2.09	0.6

Note: SHAP = SHapley Additive exPlanations, |SHAP| = Absolute Value of SHAP, SD = Standard Deviation, Freq. Top-5 = Percentage of the Ten Resamples Where a Predictor was Within the Top Five in Importance Ranking, C1 = Connectivity & Digital Capital, C2 = Contextual Socio-demographic Capital, C3 = Cultivation & Farm Capacity, C4 = Credit & Financial Behaviour. Ten independent random seeds were considered. SHAP estimation was carried out using Tree Explainer on out-of-sample observations alone. Rank standard deviation equal to zero implies that a predictor retained the same rank across all resamples. Mean rank values closer to zero signify higher predictor importance. Stage 1: N = 265; Stage 2: N = 203.

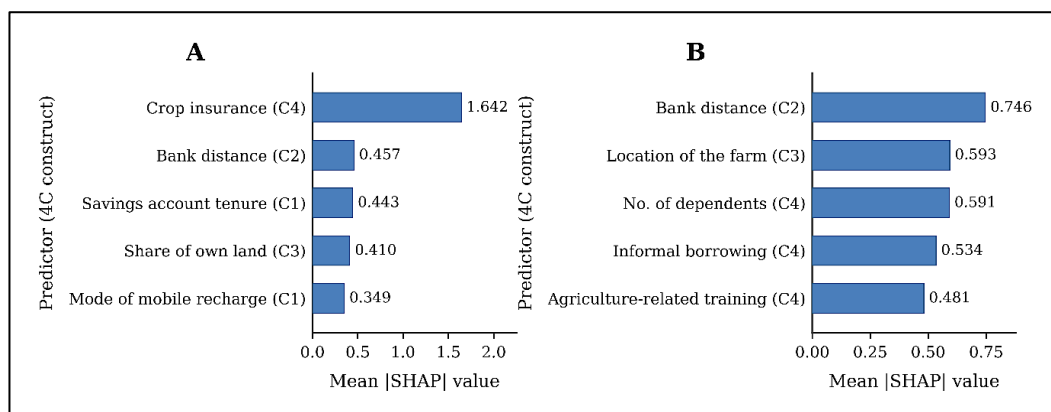


Figure 2: SHAP Importance for Credit Access and Repayment.

(A) Stage 1: Credit Access, (B) Stage 2: Repayment

Note: |SHAP| = absolute SHapley Additive exPlanations value, C1 = Connectivity and Digital Capital, C2 = Contextual Socio-demographic Capital, C3 = Cultivation and Farm Capacity, C4 = Credit and Financial Behaviour. Stage 1: n = 265; Stage 2: n = 203.

Figure 2 depicts mean absolute SHAP values for each leading predictor, capturing the average magnitude of movement in the model output induced by a particular feature. In Panel (A), we observe a pronounced level of influence exerted by crop insurance, amounting to 1.642 at the credit access phase – this figure far outweighs that

exerted by the nearest competitor, banks proximity [0.457], more than tripling its impact. By contrast, Panel (B) reveals a considerably less pronounced pattern in terms of repayment phase, wherein the top five predictors range between 0.481 and 0.746, with none exhibiting overwhelming dominance.

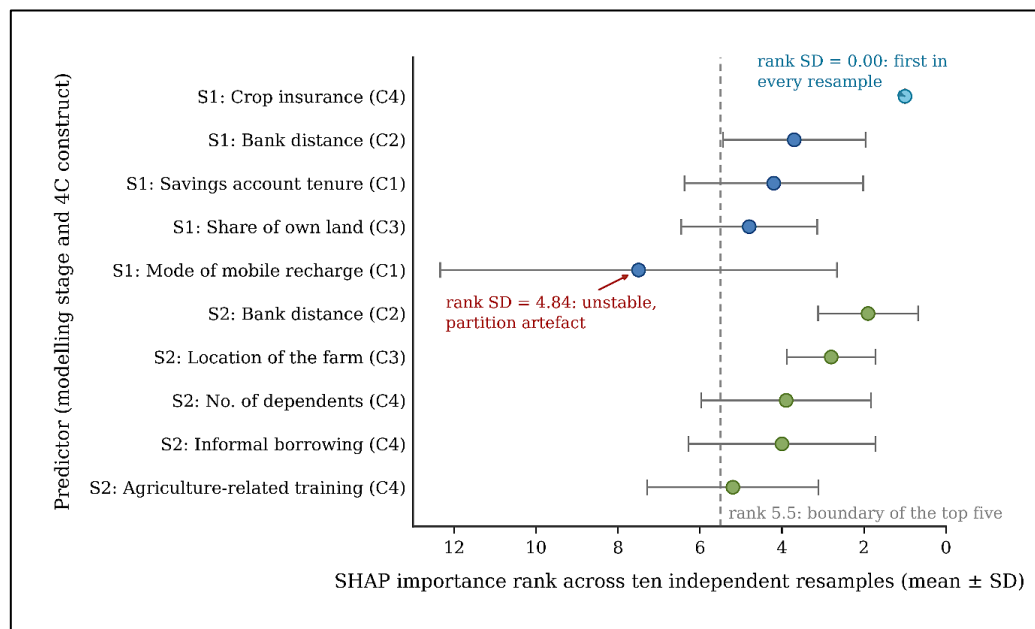


Figure 3: Rank Stability of SHAP Importance across Ten Independent Resamples

Note: S1 = Stage 1 (Credit Access), S2 = Stage 2 (Repayment), SHAP = SHapley Additive exPlanations, SD = Standard Deviation, C1 = Connectivity and Digital Capital, C2 = Contextual Socio-demographic Capital, C3 = Cultivation and Farm Capacity, C4 = Credit and Financial Behaviour. Markers show the mean importance rank across ten resamples and horizontal bars show ± 1 SD of that rank. Lower rank denotes greater importance. Stage 1: $n = 265$; Stage 2: $n = 203$.

Figure 3 represents the stability in rankings instead of the magnitude of contributions. Every point represents the average importance ranking of the predictor based on ten resampled samples. Every horizontal line represents the standard deviation of that ranking, hence representing uncertainty regarding the placement of the predictor regardless of the partition used. In this regard, short bars denote highly stable predictors since their rankings are relatively constant across partitions, while those with long bars represent unstable predictors because they have different rankings based on partitions. The dashed vertical line represents rank 5.5, the threshold dividing the top five predictors. Both figures should be considered concurrently, whereby some predictors may have a significant average absolute SHAP value in Figure 2 despite having large standard deviations in rankings in Figure 3. Mode of mobile recharge ranks among them.

Three features are stable across all resamples. Crop insurance enrolment ranks first for credit access in

all ten models, with a rank standard deviation of zero and its mean absolute SHAP value exceeds that of the next-ranked predictor by a factor of more than three, which is uncommon in field data. In terms of repayment, bank proximity and farmer location appear in the first two spots for every resample, suggesting that there is robustness in the relationships, not mere partition specificity.

The negative finding is equally important. The type of mobile phone recharge is of high importance with respect to mean absolute SHAP of 0.349; nevertheless, its rank standard deviation is 4.84 and it ranks among the top five only in half of the resamples. It can be easily identified by a single-split analysis as one of the indicators of digital capital; nevertheless, stability test proves otherwise. Coding width did not contribute to the instability. Although dichotomous, Crop insurance ranked first across all resamples with a rank standard deviation equal to zero. Differences in results emerged from the differences in the sizes of effects, not in the number of available split points.

The structure of the signalling system clearly displays a separation. Access is governed by engagement with the formal sector, represented by insurance enrolment and savings tenure and by physical distance from a branch. Repayment is governed by physical distance and farm location, which together proxy the cost of servicing a loan rather than the intention to repay. The implication for the alternative-data proposition is direct: digital connectivity measures, though central to that proposition, do not dominate either decision. On the present evidence a digital presence is a precondition for access rather than the strongest determinant of lending outcomes.

Sample-size Sensitivity

Given the relative smallness of the analytical sample, learning curve analyses were conducted by incrementally expanding the nested random subsample sizes in both phases from 30% to 100% of the total training dataset with ten replicates in each category. The results have been presented

graphically in Figure 4 below. Phase 1 demonstrates an improvement in performance measured in terms of ROC-AUC from 0.666 for 64 training instances to 0.788 when using all available training cases (approximately 212). Similarly, phase 2 shows a similar trend of increasing performance from 0.734 to 0.822 across 212 training instances. Both trends exceed the 0.500 level of randomness, but there appears to be no evidence of saturation yet.

This finding is reported in the interest of candour rather than advocacy. A curve that is still rising at the maximum available sample size indicates that discrimination is constrained by the quantity of data and that a larger sample would be expected to improve it. The estimates presented here should therefore be read as a lower bound on what the 4C instrument can achieve, not as an asymptotic ceiling. It does not, however, bear on the protocol comparison reported in Table 4, which holds sample size constant across both arms and is unaffected by the absolute level of performance.

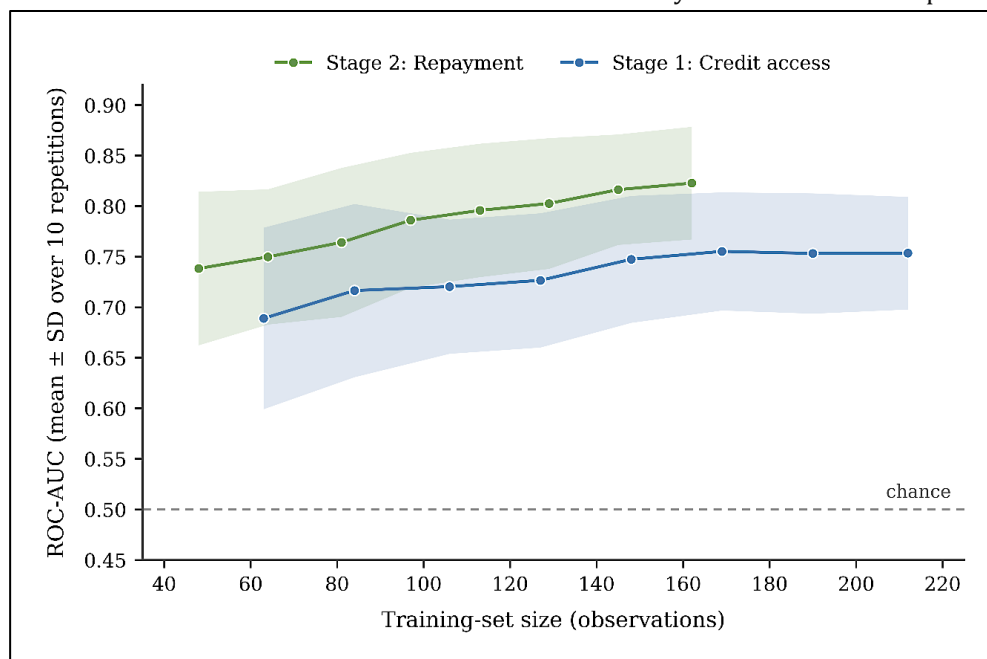


Figure 4: Learning Curves for Both Stages Under the Leakage-free Protocol

Note: ROC-AUC = receiver operating characteristic – area under the curve, SD = standard deviation. Points show the mean ROC-AUC across ten repetitions at each training-set size and shaded bands show ± 1 SD. The dashed line at 0.500 marks chance performance. Stage 1 $n = 265$; Stage 2 $n = 203$.

The Four-tier Score as a Lending Instrument

This study examined the credit score distribution using quartiles to determine loan risk classifications, where band widths differed due to ties in the discrete numerical values. Table 6

presents the classification structure of the risk bands alongside their corresponding results, while Figure 5 compares the two variables. Access to financing remains consistent throughout Tier 1 to Tier 3 at levels ranging from 76.9% to 82.8%, dropping sharply to 64.2% in Tier 4 – representing

a disparity of 17.2 percentage points relative to Tier 1. On the contrary, borrower performance exhibits an inverse trend; borrowers maintain a repayments rate of around 38% within the first

three tiers but escalate to 60.5% in Tier 4. These findings suggest that loan accessibility and borrower quality exhibit contrasting trends depending on credit scores.

Table 6: Four-tier Alternative Credit Score and Observed Outcomes (n = 265)

Tier	Score band	N	Loan access (%)	Good repayment among borrowers (%)
Tier 1 (Low risk)	57–65	59	81.4	37.5
Tier 2 (Moderate)	54–56	52	76.9	37.5
Tier 3 (Elevated)	50–53	87	82.8	38.9
Tier 4 (High risk)	39–49	67	64.2	60.5

Note: N = Number of Participants in Each Score Band, Tiers Indicate Ranking in Respect of Alternative Credit Scores and not True Risk Level. Loan Access denotes proportion of participants in each band having acquired a farm loan. Good Repayment represents proportion of participants within each band classified under Repayment Categories 3 and 4 and is based only on the sample size of 203 individuals in loan repayment status. Quartiles of scores on the 22-question test define score bands, which vary in size due to tied ranks.

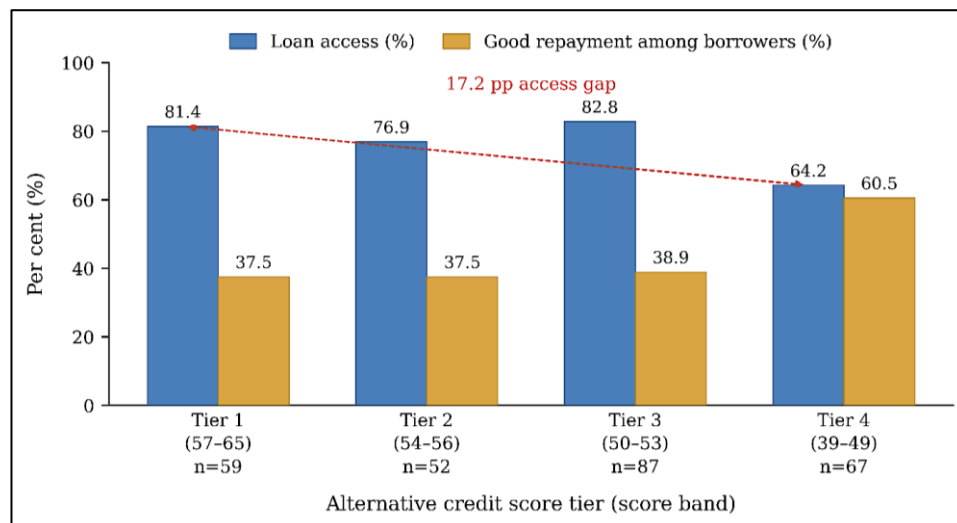


Figure 5: Four-tier Score against Observed Access and Repayment

Note: Loan access is the percentage of respondents in each band who obtained a crop loan; good repayment is the percentage of borrowers in each band in repayment categories 3 and 4 and is computed on the 203 borrowers rather than on all 265 respondents. The dotted line marks the 17.2 percentage-point access gap between Tier 1 and Tier 4. Score bands are shown in parentheses and band sizes below each tier. pp = percentage points, n = number of respondents in the band.

Figure 5 depicts the two dependent variables relative to the four scores bands. The differences observed in loan availability rates across Tier 1-3 were minimal, ranging from 76.9% to 82.8% and declined sharply to 64.2% within Tier 4 – implying a discrepancy of 17.2% below Tier 1. Conversely, borrower repayment quality displayed an inverse relationship, remaining constant at around 38% for Tier 1-3 before rising to 60.5% within Tier 4. Such divergence in trends amongst the two series in the lower spectrum was attributed to survivor bias.

Two observations can be made. The first one is that the access gradient exists but only applies to the lowest end of the score distribution. The Tiers 1 to 3 fit in a fairly close range between 76.9% and 82.8%, while Tier 4 drops to 64.2%, which is 17.2 percentage points below Tier 1. In effect, the instrument therefore works as a screen, which is

the function an alternative score is intended to serve for thin-file borrowers.

There is another interesting anomaly which requires explanation. While Tier 4 is characterized by the highest repayment ratio among borrowers at 60.5%, as compared to 37.5% of Tier 1, the main reason behind it lies not in the flaws of the scoring system but in the selection bias of survivors. The applicants in Tier 4 face the toughest screening process and thus are the only strongest applicants in Tier 4 that obtain the loans, creating a sample which is selected among those people who repay the loan. In fact, this example reinforces the need for a new scoring system, because it identifies a group of borrowers who score poorly under conventional screening yet repay reliably.

Discussion

The revised estimates put alternative data scoring in proper perspective. The Matthews correlation coefficient at Stage 1 being 0.363 shows that there is indeed some discrimination ability, although not much. This does not mean that the alternative data approach should be jettisoned. It means that the alternative data approach cannot operate under its own steam. Its immediate value would lie in functioning as a screening tool to spot bottom quartile cases for action and cases where the borrower's conduct is unexpected from the point of view of collateral-based evaluation. Human judgement must intervene and the results of the models used for important decisions are not infallible.

It is worth situating these figures against the wider literature. Benchmarking studies of consumer credit scoring routinely report ROC-AUC values in the range of 0.75 to 0.85 on large bureau datasets with rich repayment histories and reviews of machine learning in rural finance document a similar spread where the underlying data are behavioural rather than historical. Digital-footprint studies reporting higher figures typically draw on platform-generated transaction traces captured passively at scale, which differ in both volume and granularity from a 22-item survey instrument administered once in the field (6, 8). Our values should therefore be read as consistent with, rather than inferior to, the credible portion of this literature. The values that our results sit awkwardly beside are those exceeding 0.90 on small field samples and the protocol experiment reported here offers a parsimonious explanation for a substantial share of that gap.

That experiment is the paper's central methodological contribution. Holding data, features, hyperparameters and random seed constant, relocating a single pre-processing step moved the reported ROC-AUC from 0.956 to 0.783. This is consistent with the general formulation of leakage as the inadvertent introduction of information about the target into the training process, with the demonstrated bias of cross-validation estimates when model selection is performed on the full sample and with the specific warning that resampling applied before partitioning yields overoptimistic estimates on imbalanced data. What the present result adds is a magnitude from field data in a substantive

application domain. An inflation of 0.173 AUC units is large enough to convert a model that should not be deployed into one that appears close to deployable and it is invisible to any reader unless the protocol is stated. We would therefore argue that the placement of the resampling step relative to partitioning should be treated as a reporting requirement in this literature rather than an optional methodological detail.

The explanatory results carry more direct policy content. Crop insurance enrolment emerges as the dominant determinant of credit access, with no other variable approaching it and with a rank that does not vary across resamples. The relationship is unlikely to be causal in any direct sense. Experimental evidence establishing a causal link between insurance and agricultural investment has required the randomised provision of cover (30); the present cross-sectional association admits no such inference. Enrolment in an insurance scheme does not itself confer creditworthiness; it functions as a proxy for engagement with the formal sector. It signals that the farmer has completed a formal transaction, possesses the documentation and an institutional trace of the kind that a thin-file borrower normally lacks. The item is dichotomous and 57% of respondents were not enrolled, so what dominates prediction is not a graded measure of formal engagement but a single binary institutional trace. If this reading is correct, it has a constructive implication: schemes that generate formal transaction records, including crop insurance and income-support transfers, produce a credit-relevant signal as a by-product and that signal could be made available to lenders through consented data-sharing rather than left dormant. That the signal is binary makes this implication more tractable rather than less, since it requires only a record of participation, not a graded assessment. Distance to a bank branch is the second stable determinant and the only variable ranking highly at both stages. Its persistence across access and repayment suggests that proximity to a formal institution remains materially consequential for smallholders. Distance imposes transaction costs not only at the point of application but throughout the repayment cycle, since each instalment requires travel, foregone labour time and, in the agency areas of the study zone, a journey that may consume a working day.

The finding that farm location ranks second for repayment is consistent with the same mechanism. Taken together, these results suggest that the binding constraints on smallholder credit in this setting are as much logistical as informational and that a scoring model which ignores geography will misattribute to borrower conduct what is in fact a function of infrastructure.

The four-tier results have a practical reading. The concentration of the access gradient at the bottom of the distribution means the instrument is best deployed as a screen rather than as a ranking device and the counter-intuitive repayment pattern in Tier 4 identifies precisely the population an alternative score is meant to serve: borrowers whom conventional screening rejects but who repay when lent to. Deployment must have its safeguards. A fairness test must be conducted with respect to gender, social category and land-holding class before deployment can be considered, since credit scoring algorithms can be discriminatory even if protected characteristics are removed from them (31). Further, a structured audit must accompany deployment, not come after it (32). The evidence that digital financial inclusion has produced uneven gains across gender and socio-economic groups in India makes this requirement concrete rather than precautionary.

Conclusion

Thin-file credit scoring built on alternative data carries meaningfully more predictive power than random assignment and its explanations are stable enough to inform policy, provided that fairness assessment and audit accompany any deployment. Predictive performance is nonetheless lower than parts of the literature suggest and this becomes evident once resampling is performed correctly. For the 265 farmers surveyed, leakage-free estimation yields a ROC-AUC of 0.783 for credit access and 0.819 for repayment, whereas the conventional protocol yields 0.956 and 0.870 on the same sample. This study seeks to derive a credible benchmark for alternative data credit scoring of thin-file smallholder farmers and evaluate the strength of its justification for policy. Based on a sample of 265 farmers, leakage-free estimation results in an ROC-AUC value of 0.783 and 0.819, respectively, for access to credit and likelihood of repayment using this approach compared to a value of 0.956 and 0.870 using

traditional approaches on identical data. Alternative data possess significant predictive capacity relative to pure chance allocation but not as great as suggested in some literature. The three contributions herein arise. First, the empirical evidence represents the inaugural quantification of resampling bias – 0.173 ROC-AUC inflated per field dataset – caused by resampling in India, with all else being held constant save one preprocessing decision within the machine learning pipeline. Second, this paper marks the first use of SHAP rank stability tests in agricultural credit scoring. A variable may be found to have a large mean absolute SHAP value and simultaneously have a rank standard deviation of 4.84 between resamples. In this way, the predictor should not be used as policy advice. Third, the 4C instrument comprises 22 questions designed for field implementation.

Methodological and practical implications emerge. Researchers need to ensure adequate documentation of whether resampling precedes or follows partitioning. It becomes necessary to provide empirical support for the reliability of SHAP rankings. In practical settings, the enrolment in agricultural insurance appears to be the most robust predictor of eligibility alongside distance to banks, whereas distance to banks, along with farm location, predicts default behaviour. Consequently, the binding constraint lies more within logistical and administrative issues than behavioural concerns. The proposed four-level tool acts like a filter rather than a ranker, helping identify the potential clients served by the non-conventional scoring system, specifically those who get rejected based on traditional filters yet make good payers upon borrowing opportunities.

Limitations

- (a) **Sample size:** Stage 1 uses 265 farmers and Stage 2 uses 203. Confidence intervals are correspondingly wide and the estimates are best read as a benchmark range rather than as point values.
- (b) **Geographic concentration:** Srikakulam contributes 74.7% of observations. This mirrors the structure of the study region but limits generalisation to districts with different infrastructure.
- (c) **Cross-sectional design:** Repayment is observed once, so the analysis cannot separate transient from persistent difficulty, nor

establish causal direction for the insurance result.

- (d) **Self-reported outcomes:** Both dependent variables are farmer-reported and were not validated against lender records, so social-desirability bias on repayment cannot be excluded.
- (e) **Internal validation only:** Stability is estimated by repeated partitioning of a single sample; external validation in a different agro-climatic or lending environment remains necessary.
- (f) **Item variance:** Four items out of twenty-two items fall within one response option: disturbances frequency (94.0% average score = 3.91), agriculture related training (87.2%), banking transaction awareness (82.3%) and consumer durables purchased on credit (80.4%). Construct C4 contains three of these four items, corresponding to its smallest standard deviation of 1.11 compared to others and contributing to its emergence as the constraining construct. Such low item variance limits the ability of an item to discriminate respondents.

Therefore, the reported performance should therefore be read as conditional on this response distribution. Revision of the items to improve their variance would be a viable improvement to the scale.

Future Scope

Five directions follow from these limitations. External validation should be pursued in a second agro-climatic zone and, ideally, against a lender's own portfolio, which would address the reliance on internal validation and test whether the 4C structure transfers. Longitudinal follow-up would allow transient repayment difficulty to be separated from persistent difficulty and would permit the crop-insurance association to be examined causally rather than based on correlational evidence only, for instance by exploiting staggered scheme enrolment. Linkage to administrative records, including crop-insurance enrolment data, income-support payment records and bank core-banking systems, would replace farmer-reported outcomes and remove the social-desirability concern. Extension to consented digital-footprint variables, such as transaction frequency and mobile-account tenure, would test whether passively captured behavioural data add discriminatory power over the survey instrument

in this population and would require data-protection safeguards commensurate with the sensitivity of the data. Finally, fairness auditing across gender, social category and landholding class should precede any operational deployment. The learning curves in Figure 4 give this programme a concrete priority: neither curve has plateaued at the maximum available sample size, so extending the survey is likely to yield further gains in discrimination and establishing the size at which performance stabilises is a necessary precondition for any claim about the ceiling of the approach.

Abbreviations

4C: Connectivity and Digital Capital, Contextual Socio-demographic Capital, Cultivation and Farm Capacity, Credit and Financial Behaviour, AUC: Area Under the Curve, MCC: Matthews Correlation Coefficient, ROC: Receiver Operating Characteristic, SD: Standard Deviation, SHAP: SHapley Additive exPlanations, SMOTE: Synthetic Minority Over-sampling Technique.

Acknowledgement

The authors gratefully acknowledge the support and assistance received from the School of Management, Centurion University of Technology and Management, Andhra Pradesh and the Department of MBA, Koneru Lakshmaiah Education Foundation, Andhra Pradesh, in facilitating data collection for this study.

Author Contributions

Anil Kumar Jonnalagadda: Percform Experiments, Data Analysis, Writing - Original Draft, Narasimha Rao MVAL: Study Design, Supervision, Ramesh Babu S: Conceptualization, Data Analysis. All authors reviewed and approved the final version.

Conflict of Interest

The authors declare that there is no conflict of interest associated with this publication.

Data Availability

The datasets generated and analysed during the present study shall be provided by the corresponding author on reasonable request. Personal identifiers were removed and responses encoded and secure storage and aggregate reporting ensured the privacy and anonymity of participants.

Declaration of Artificial Intelligence

(AI) Assistance

Artificial intelligence assistance was utilized during the manuscript drafting process, including language revision, analysis code review and internal validation of reported figures. AI technology aided in coding for Figure 1 to 5 redrafting based on the authors' own data set, ensuring enhanced resolution and typography per journal guidelines. Neither of the figures nor any components thereof have been generated using generative artificial intelligence technologies; moreover, AI did not create, edit or infer any figure contents. All plotted data points originate from the study data set, while design aspects were defined by the authors. Methods, data collection, model specification, statistical analysis, as well as interpretation, remain exclusively the authors' intellectual property. Authors attest to having reviewed and approved all reported content. The authors take full responsibility for the content's originality, interpretation and accuracy.

Ethics Approval

Ethical clearance for this study was granted by the Institutional Ethics Committee of Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, before data collection commenced. The study was classified as minimal-risk social science investigation without the issuance of any numbered certificate by the committee; thus, there is no office reference number available. The research processes complied fully with the institutional code of ethics and ethical standards stipulated in the World Medical Association Declaration of Helsinki (2013). Voluntary participation was ensured and verbal consent was obtained from all respondents prior to the interview. Each individual received a detailed explanation of the purpose of the study, anonymised use of the data for academic purposes and participant withdrawal rights. Consent was sought verbally in Telugu since participants belonged to the low-literacy rural population where a signature requirement would have acted as a hurdle to their participation.

Funding

The authors received no grant from any public, private or non-profit organisation for undertaking the present study.

References

1. Stiglitz JE, Weiss A. Credit rationing in markets with imperfect information. *Am Econ Rev.* 1981;71(3):393-410.
<https://www.jstor.org/stable/1802787>
2. Burgess R, Pande R. Do rural banks matter? Evidence from the Indian social banking experiment. *Am Econ Rev.* 2005;95(3):780-795.
<https://doi.org/10.1257/0002828054201242>
3. Kumar A, Das R, Aditya KS, *et al.* Examining institutional credit access among agricultural households in eastern India: trends, patterns and determinants. *Agric Finance Rev.* 2021;81(2):250-64.
<https://doi.org/10.1108/AFR-04-2020-0054>
4. Chakrabarty KC. Financial inclusion: a road India needs to travel. *RBI Bull.* 2011;1851-1856.
https://www.rbi.org.in/scripts/BS_ViewBulletin.aspx?Id=12680
5. Demirgüç-Kunt A, Klapper L, Singer D, *et al.* The Global Findex Database 2017: measuring financial inclusion and opportunities to expand access to and use of financial services. *World Bank Econ Rev.* 2020;34(Suppl_1):S2-S8.
<https://doi.org/10.1093/wber/lhz013>
6. Berg T, Burg V, Gombović A, Puri M. On the rise of FinTechs: credit scoring using digital footprints. *Rev Financ Stud.* 2020;33(7):2845-2897.
<https://doi.org/10.1093/rfs/hhz099>
7. Óskarsdóttir M, Bravo C, Sarraute C, *et al.* The value of big data for credit scoring: enhancing financial inclusion using mobile phone data and social network analytics. *Appl Soft Comput.* 2019;74:26-39.
<https://doi.org/10.1016/j.asoc.2018.10.004>
8. Björkegren D, Grissen D. Behavior revealed in mobile phone usage predicts credit repayment. *World Bank Econ Rev.* 2020;34(3):618-634.
<https://doi.org/10.1093/wber/lhz006>
9. Jagtiani J, Lemieux C. The roles of alternative data and machine learning in fintech lending: evidence from the LendingClub consumer platform. *Financ Manage.* 2019;48(4):1009-1029.
<https://doi.org/10.1111/fima.12295>
10. Kumar A, Sharma S, Mahdavi M. Machine learning technologies for digital credit scoring in rural finance: a literature review. *Risks.* 2021;9(11):192.
<https://doi.org/10.3390/risks9110192>
11. Frost J, Gambacorta L, Huang Y, *et al.* BigTech and the changing structure of financial intermediation. *Econ Policy.* 2019;34(100):761-799.
<https://doi.org/10.1093/epolic/eiaa003>
12. Djeundje VB, Crook J, Calabrese R, *et al.* Enhancing credit scoring with alternative data. *Expert Syst Appl.* 2021;163:113766.
<https://doi.org/10.1016/j.eswa.2020.113766>
13. Duvendack M, Sonne L, Garikipati S. Gender inclusivity of India's digital financial revolution for attainment of SDGs: macro achievements and the micro experiences of targeted initiatives. *Eur J Dev Res.* 2023;35(6):1369-91.
<https://doi.org/10.1057/s41287-023-00585-x>
14. Giné X, Klöner S. Credit constraints as a barrier to technology adoption by the poor: Lessons from South Indian small-scale fisheries. In: Nissanke M, Thorbecke E, eds. *Globalization and the Poor in Asia.*

- Studies in Development Economics and Policy. London: Palgrave Macmillan; 2008:221-49.
https://doi.org/10.1057/9780230594005_9
15. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res.* 2002;16:321-357.
<https://doi.org/10.1613/jair.953>
 16. Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in data mining: formulation, detection and avoidance. *ACM Trans Knowl Discov Data.* 2012;6(4):15.
<https://doi.org/10.1145/2020408.2020496>
 17. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics.* 2006;7(91):1-8.
<https://doi.org/10.1186/1471-2105-7-91>
 18. Santos MS, Soares JP, Abreu PH, Araújo H, Santos J. Cross-validation for imbalanced datasets: avoiding overoptimistic and overfitting approaches. *IEEE Comput Intell Mag.* 2018;13(4):59-76.
<https://doi.org/10.1109/MCI.2018.2866730>
 19. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell.* 2019;1:206-215.
<https://doi.org/10.1038/s42256-019-0048-x>
 20. Knowles B, D'Cruz J, Richards JT, Varshney KR, Humble AI. *Commun ACM.* 2023;66(9):73-79.
<https://doi.org/10.1145/3587035>
 21. Breiman L. Random forests. *Mach Learn.* 2001;45:5-32.
<https://doi.org/10.1023/A:1010933404324>
 22. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. *PLoS One.* 2019;14(11):e0224365.
<https://doi.org/10.1371/journal.pone.0224365>
 23. Lessmann S, Baesens B, Seow HV, *et al.* Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research. *Eur J Oper Res.* 2015;247(1):124-136.
<https://doi.org/10.1016/j.ejor.2015.05.030>
 24. Reddy AA, Bantilan MCS. Regional disparities in Andhra Pradesh, India. *Local Econ.* 2013;28(1):123-35.
<https://doi.org/10.1177/0269094212463791>
 25. Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. *J Big Data.* 2020;7(94):1-45.
<https://doi.org/10.1186/s40537-020-00369-8>
 26. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient over F1 score and accuracy in binary classification evaluation. *BMC Genomics.* 2020;21(6):2-13.
<https://doi.org/10.1186/s12864-019-6413-7>
 27. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One.* 2015;10(3):e0118432.
<https://doi.org/10.1371/journal.pone.0118432>
 28. Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. *Knowl Inf Syst.* 2014;41(3):647-665.
<https://doi.org/10.1007/s10115-013-0679-x>
 29. Lundberg SM, Erion G, Chen H, *et al.* From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell.* 2020;2(1):56-67.
<https://doi.org/10.1038/s42256-019-0138-9>
 30. Karlan D, Osei R, Osei-Akoto I, Udry C. Agricultural decisions after relaxing credit and risk constraints. *Q J Econ.* 2014;129(2):597-652.
<https://doi.org/10.1093/qje/qju002>
 31. Kozodoi N, Jacob J, Lessmann S. Fairness in credit scoring: assessment, implementation and profit implications. *Eur J Oper Res.* 2022;297(3):1083-1094.
<https://doi.org/10.1016/j.ejor.2021.06.023>
 32. Laine J, Minkkinen M, Mäntymäki M. Ethics-based AI auditing: a systematic literature review on conceptualizations of ethical principles and knowledge contributions to stakeholders. *Inf Manage.* 2024;61(5):103969.
<https://doi.org/10.1016/j.im.2024.103969>

How to Cite: Jonnalagadda AK, Rao MVALN, Babu SR. Thin-file, Not Signal-free: An Alternative-data Credit Score for Smallholder Farmers Built with Explainable Machine Learning. *Int Res J Multidiscip Scope.* 2026;7(4):1505-1520. DOI: 10.47857/irjms.2026.v07i04.013934