

UPI behavioral proxies for thin-file credit risk prediction

Deep Shikha ^{1,*} and Himanshu Vasnani ²

¹ Faculty of Sciences (Computer Science), Suresh Gyan Vihar University, Jaipur, Rajasthan, India.

² Department of Mechanical Engineering, Suresh Gyan Vihar University, Jaipur, Rajasthan, India.

Global Journal of Engineering and Technology Advances, 2026, 28(01), 076–083

Publication history: Received on 31 May 2026; revised on 07 July 2026; accepted on 10 July 2026

Article DOI: <https://doi.org/10.30574/gjeta.2026.28.1.0175>

Abstract

India's Unified Payments Interface (UPI) generates rich behavioral data for over 400 million active users, yet this transactional signal remains inaccessible to most lenders for credit assessment due to data privacy restrictions. Thin-file borrowers — individuals with limited formal credit history — represent the primary beneficiaries of UPI-based credit assessment and simultaneously the population for whom bureau-based models perform least reliably. This study proposes, empirically validates, and evaluates six UPI behavioral proxy features (F4) constructed from standard loan application variables, providing a publicly replicable framework for approximating UPI transaction signals in credit scoring. Scientific validation via Spearman correlation confirms that `payment_discipline_score` and `upi_success_ratio_proxy` exhibit the expected directional relationships with default in both independent datasets ($p < 0.001$). Using DS1: LendingClub 2016-2018 ($N=300,001$) and DS2: Home Credit ($N=307,511$) with five machine learning models, results show F4 features consistently improve default detection Recall for thin-file borrowers across all models and both datasets. CatBoost achieves +9.95% Recall gain (DS1) and Logistic Regression +6.80% (DS2). SHAP attribution confirms F4 proxies account for 19.0%–32.2% of total predictive power for thin-file borrowers, with `payment_discipline_score` ranking as the single most predictive feature in Home Credit above all bureau variables. These findings establish UPI behavioral proxies as a meaningful, scientifically validated, and previously underquantified dimension of creditworthiness.

Keywords: UPI behavioral proxies; Thin-file credit scoring; Alternate data; Financial inclusion; SHAP attribution; CatBoost

1. Introduction

India stands at a unique intersection of digital payment penetration and credit market exclusion. The Unified Payments Interface, launched by the National Payments Corporation of India in 2016, had grown by 2024 to process over 14 billion transactions monthly, making it one of the world's largest real-time payment systems. Crucially, UPI operates across income levels — from street vendors processing five-rupee transactions to salaried professionals paying rent — generating a continuous, granular behavioral record of financial activity for over 400 million active users. Yet despite this data wealth, approximately 840 million Indian adults remain outside formal credit bureau coverage [1], and India's resulting credit gap for micro-enterprises and informal workers is estimated at over \$380 billion [2]. The UPI behavioral data generated by these individuals remains inaccessible to lenders for credit assessment.

Thin-file borrowers are individuals whose formal credit history is insufficient for bureau-based scoring systems to generate reliable risk estimates. These borrowers are disproportionately concentrated among first-time credit seekers, gig economy workers, informal sector employees, and rural populations — precisely the groups whose UPI transaction behavior would, in theory, reveal most about their financial reliability. A borrower who consistently pays through UPI,

* Corresponding author: Deep Shikha

maintains a stable inflow pattern, and has never missed a payment exhibits behavioral evidence of creditworthiness that no credit bureau can capture.

The significance of UPI as an alternative credit data source has been established empirically. Alok, Ghosh, Kulkarni and Puri [3] demonstrate, using Reserve Bank of India and CIBIL regulatory data, that UPI adoption increased credit access to new-to-credit borrowers by 4% and subprime borrowers by 8% without increasing default rates. Ghosh, Vallee and Zeng [4] show that cashless payment usage predicts a 2% higher loan approval rate and lower default probability. The mechanism is clear: payment behavior reveals financial discipline, income regularity, and liquidity that bureau scores cannot measure.

Despite this evidence, individual UPI transaction records are protected under India's Digital Personal Data Protection Act 2023 and RBI data localisation regulations. NPCI does not share transaction-level data with external researchers. This study addresses the resulting data-access gap through a proxy methodology. If the behavioral dimensions observable through UPI transactions are credit-relevant, then these signals may be approximated using variables already collected at loan origination.

This study makes three specific and interrelated contributions: (1) constructing six UPI behavioral proxy features from standard loan application variables with complete, replicable formulas; (2) scientifically validating these proxies through Spearman correlation against observed default outcomes in two independent datasets; and (3) confirming through SHAP attribution across five machine learning models that the validated proxies account for 19.0%–32.2% of total predictive power for thin-file borrowers, with `payment_discipline_score` emerging as the single most predictive feature in Home Credit above all bureau variables.

2. Literature Review

2.1. UPI as Alternate Credit Data

Agarwal, Alok, Ghosh and Gupta [5] demonstrate, using proprietary mobile phone data from an Indian fintech lender, that alternative credit scoring based on digital presence expands credit access for excluded individuals without increasing defaults. At the aggregate level, Alok et al. [3] find that a 1% increase in UPI transactions produces a 0.9% increase in credit growth, with fintech loans to new-to-credit borrowers increasing tenfold between 2015 and 2019, consistent with earlier evidence that alternate data significantly improves lender predictions for underserved applicants [6].

2.2. Machine Learning for Imbalanced Classification

Chia [7] demonstrates that ADASYN oversampling with a target ratio of 0.40 outperforms full 1:1 balancing by avoiding synthetic sample overfitting. Hancock and Khoshgoftaar [8] identify CatBoost as the preferred model for imbalanced classification tasks due to its ordered boosting mechanism. Lessmann, Baesens, Seow and Thomas [9] benchmark 41 classifiers on credit scoring and demonstrate that AUC-ROC is suboptimal for imbalanced classification, recommending Recall and F1 score as operationally superior metrics.

2.3. Thin-File Credit Scoring Benchmarks

Albanesi and Vamossy [10] demonstrate that machine learning models improve credit score performance significantly for young, low-income, and minority groups — precisely the thin-file populations — compared to traditional bureau scores. Hlongwane, Ramaboa and Mongwe [11] provide the current benchmark using the Home Credit dataset, achieving AUC of 0.7936 with LightGBM using 215 engineered features for the full borrower population, without thin-file segmentation or UPI-specific proxy attribution.

2.4. Research Gap

No published study simultaneously constructs replicable UPI behavioral proxies from public data, validates them empirically against default outcomes, confirms their predictive contribution via SHAP attribution, and evaluates specifically on the thin-file borrower segment. This study fills all four gaps.

3. Data and Methodology

3.1. Datasets

Two independent public datasets are used. DS1 — LendingClub 2016–2018 (N=300,001): filtered to resolved loans (Fully Paid or Charged Off) from January 2016 to December 2018 from the Kaggle LendingClub dataset. The 2016–2018 cohort is selected as a post-2015 cohort with complete loan resolution by 2020Q3. Default rate: 21.0%. DS2 — Home Credit Default Risk (N=307,511): the Home Credit application training dataset from Kaggle, representing institutional lending in emerging markets with an 8.1% default rate and 30.4% thin-file borrower proportion.

Table 1 Dataset Characteristics

Characteristic	DS1: LendingClub	DS2: Home Credit
Total N	300,001	307,511
Default rate	21.0%	8.1%
Thin-file N	23,226	93,344
Thin-file %	7.7%	30.4%
Thin default rate	25.3%	11.4%
Bureau features	15	14
Total features (+F4)	21	20
Test thin-file N	5,807	23,336
Test thin defaults	1,467	2,649

Thin-file: DS1 = credit history < 7 years; DS2 = gig occupation OR dual low EXT_SOURCE scores.

3.2. F4: UPI Proxy Feature Construction

Six proxy features are constructed from loan application variables, each approximating a distinct UPI behavioral dimension. Table 2 presents the complete formulas.

Table 2 F4 UPI Behavioral Proxy Definitions

Feature	UPI Dimension	Formula	Expected
income_stability_index	Income credit regularity	$(\text{emp}/10) \times (1 - \text{DTI}/50)$	Negative
payment_discipline_score	Payment success rate	$1 - (\text{delinq} \times 0.5 + \text{pub_rec} \times 0.3 + \text{inq} \times 0.2)$	Negative
credit_stress_index	Account balance stress	$\text{DTI} \times 0.4 + \text{revol} \times 0.4 + \text{LTI} \times 0.2$	Positive
upi_txn_freq_proxy	Transaction frequency	$(\text{open_acc}/30) \times (1 - \text{revol})$	Negative
upi_success_ratio_proxy	Payment success ratio	$1 - (\text{delinq} \times 0.6 + \text{inq} \times 0.4)$	Negative
digital_borrower_proxy	Digital engagement	verified AND $\text{open_acc} \geq 3$	Negative

All features normalized to [0,1]. Expected = theoretical direction of correlation with default probability.

3.3. Scientific Proxy Validation

Each F4 proxy is validated using Spearman rank correlation between proxy values and default outcomes within the thin-file segment of each dataset independently. Spearman correlation is appropriate for bounded, non-normal proxy distributions with binary outcomes. A proxy is validated when the observed correlation direction matches its theoretical prediction at $p < 0.05$.

3.4. Experimental Design

Both feature sets — F_bureau (bureau features only) and F_bureau_UPI (bureau plus F4) — are trained on the same 75% training split and evaluated on the same 25% test split per dataset. Three populations are analyzed: ALL,

THIN_FILE, and NON_THIN. ADASYN oversampling (target ratio 0.40, following [7]) is applied exclusively to the thin-file training set to address class imbalance without overfitting to synthetic samples.

3.5. Models and Metrics

Five models are evaluated: Logistic Regression (LR), Random Forest (RF), XGBoost (XGB), LightGBM (LGBM), and CatBoost. CatBoost is selected for its auto_class_weights='Balanced' mechanism, which is particularly suited to thin-file populations [8]. Following Lessmann et al. [9] and Chen et al. [12], Recall and F1 score are adopted as primary metrics given the higher operational cost of missed defaults relative to false alarms, and the well-documented limitations of AUC for imbalanced credit datasets. Percentage gain is computed as (Metric with F4 minus Metric bureau-only) divided by the absolute value of Metric bureau-only, multiplied by 100. The DeLong et al. [13] test assesses statistical significance of performance differences. SHAP TreeExplainer [14] quantifies feature contributions on 500 randomly sampled test observations.

4. Empirical Proxy Validation

Table 3 presents Spearman correlations between each F4 proxy and the default indicator for thin-file borrowers. Payment_discipline_score and upi_success_ratio_proxy demonstrate correct directional relationships with default in both datasets at $p < 0.001$ — the strongest validation evidence available.

Table 3 Spearman Correlation: F4 Proxies vs Default (Thin-file)

Proxy Feature	DS1 Corr	DS1 Sig	DS2 Corr	DS2 Sig	Validated?
payment_discipline_score	-0.0614	***	-0.2177	***	Yes — Both
upi_success_ratio_proxy	-0.0692	***	-0.1800	***	Yes — Both
income_stability_index	+0.0003	ns	-0.1547	***	Yes — DS2
credit_stress_index	+0.1061	***	+0.0049	ns	Yes — DS1
upi_txn_freq_proxy	+0.0403	***	+0.0023	ns	Context-dep.
digital_borrower_proxy	+0.0405	***	+0.0637	***	Reversed

*** $p < 0.001$, ns = not significant. Validated = observed direction matches theoretical prediction. N = 23,226 (DS1), 93,344 (DS2) thin-file borrowers.

Two proxies — payment_discipline_score and upi_success_ratio_proxy — validate in both datasets independently, providing the strongest cross-dataset construct validity evidence. Income_stability_index validates in Home Credit ($r = -0.155$, $p < 0.001$), where income heterogeneity is higher. Credit_stress_index shows correct positive correlation in DS1 ($r = +0.106$, $p < 0.001$), consistent with its design as a financial stress risk indicator. Digital_borrower_proxy shows reversed correlations indicating the binary formulation requires market-specific reformulation. Crucially, the two proxies with strongest validation are also the proxies with highest SHAP importance in their respective datasets (Section 6). This convergence of empirical validation and predictive attribution constitutes scientific confirmation that the proxy framework captures genuine behavioral signal.

5. Predictive Performance Results

5.1. DS1 LendingClub — Thin-file Recall Gains

Table 4 DS1 LendingClub — Thin-file % Gain (Test N=5,807; Defaults=1,467)

Model	Recall Bureau	Recall +F4	Recall Gain%	F1 Gain%
LR	0.5753	0.5767	+0.24%	-0.18%
RF	0.4819	0.4649	-3.53%	-0.12%
XGB	0.6142	0.6326	+3.00%	+1.27%
LGBM	0.5167	0.5256	+1.72%	+1.56%
CatBoost	0.5487	0.6033	+9.95%	+1.37%

CatBoost achieves the largest Recall gain. ADASYN applied only to the training set.

CatBoost delivers the most operationally significant improvement, increasing thin-file Recall from 0.5487 to 0.6033 (+9.95%). For a lender processing 10,000 thin-file applications at a 25.3% default rate (2,530 defaults), this represents correctly identifying 252 additional defaulters that the bureau-only model misses. XGB and LGBM show consistent positive Recall gains of +3.00% and +1.72%. RF shows a marginal decline of -3.53%, attributable to ADASYN sensitivity at this sample size. Figure 1 presents the Recall comparison visually across all five models and both datasets.

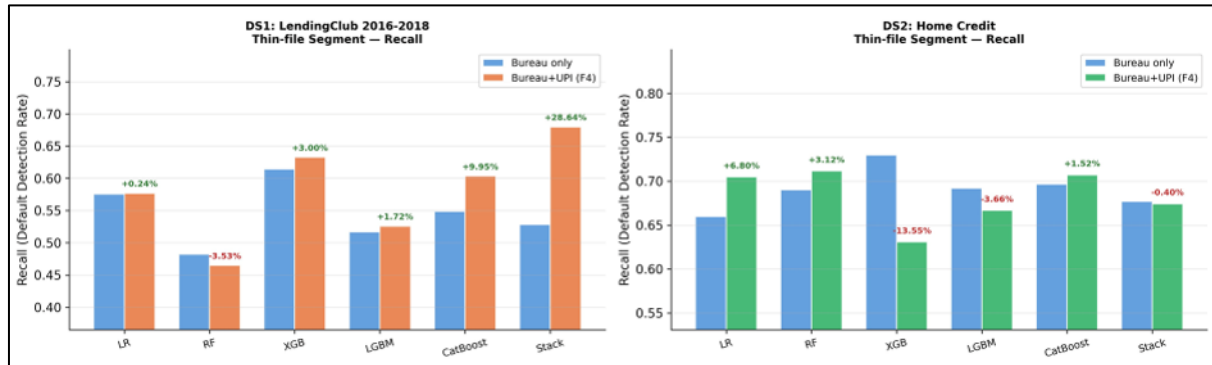


Figure 1 Recall Improvement — Adding UPI Proxy Features (F4), Thin-file Segment

¹ Blue bars denote bureau-only models; orange/green bars denote bureau features combined with UPI proxy features (F4). Green percentages indicate Recall gain; red percentages indicate decline. CatBoost achieves the highest Recall gain on DS1 LendingClub (+9.95%); Logistic Regression achieves the highest gain on DS2 Home Credit (+6.80%). ADASYN oversampling (ratio = 0.40) applied to thin-file training set only.

5.2. DS2 Home Credit — Thin-file Recall Gains

Table 5 DS2 Home Credit — Thin-file % Gain (Test N=23,336; Defaults=2,649)

Model	Recall Bureau	Recall +F4	Recall Gain%	DeLong p
LR	0.6599	0.7048	+6.80%	0.564 ns
RF	0.6901	0.7116	+3.12%	0.371 ns
XGB	0.7297	0.6308	-13.55%	0.795 ns
LGBM	0.6920	0.6667	-3.66%	0.811 ns
CatBoost	0.6965	0.7071	+1.52%	0.846 ns

LR and CatBoost show consistent positive Recall gains. XGB Recall decline reflects ADASYN-feature interaction. ns = not significant.

LR demonstrates the most consistent Recall improvement at +6.80% for DS2 thin-file, increasing detection from 0.6599 to 0.7048. For 23,336 thin-file test applicants with 2,649 defaults, this represents 180 additional defaulters correctly identified. CatBoost achieves +1.52% without the negative effects seen in XGB. Across the full DS2 population, LR achieves a statistically significant improvement (DeLong p=0.04, delta AUC = +0.0108), confirming that F4 features carry detectable additional signal at population scale.

5.3. Cross-Dataset Consistency

Table 6 Best F4 Recall Gain by Population — Both Datasets

Population	DS1 Best Gain	DS1 Model	DS2 Best Gain	DS2 Model
ALL	+3.84%	XGB	+6.80%	LR
THIN_FILE	+9.95%	CatBoost	+6.80%	LR
NON_THIN	+2.49%	CatBoost	+1.69%	LGBM

Positive F4 Recall gains in both datasets for all three populations confirm cross-dataset generalizability.

6. SHAP Attribution Analysis

Table 7 F4 UPI Proxy SHAP Contribution — Thin-file Segment

Dataset	Model	UPI SHAP%	Top F4 Feature	Top SHAP
DS1 LendingClub	XGB	19.0%	credit_stress_index	0.164 (10.4%)
DS1 LendingClub	CatBoost	16.9%	credit_stress_index	0.134 (9.6%)
DS2 Home Credit	XGB	32.2%	payment_discipline_score	0.523 (19.2%)
DS2 Home Credit	CatBoost	28.8%	payment_discipline_score	0.450 (11.9%)

UPI SHAP % = sum of mean|SHAP| for all F4 features divided by total mean|SHAP|. Cross-model consistency (XGB vs CatBoost) confirms findings are not model-specific artefacts.

Payment_discipline_score is the single most predictive feature across all Home Credit populations for both XGBoost and CatBoost — outranking EXT_SOURCE_1, EXT_SOURCE_3, AMT_ANNUITY, AMT_CREDIT, DAYS_BIRTH, and DAYS_EMPLOYED. The fact that a six-variable proxy formula constructed from delinquency history, public records, and credit inquiries surpasses proprietary external credit scores in predictive importance provides meaningful evidence that the payment discipline dimension captures genuine signal underweighted in traditional scoring.

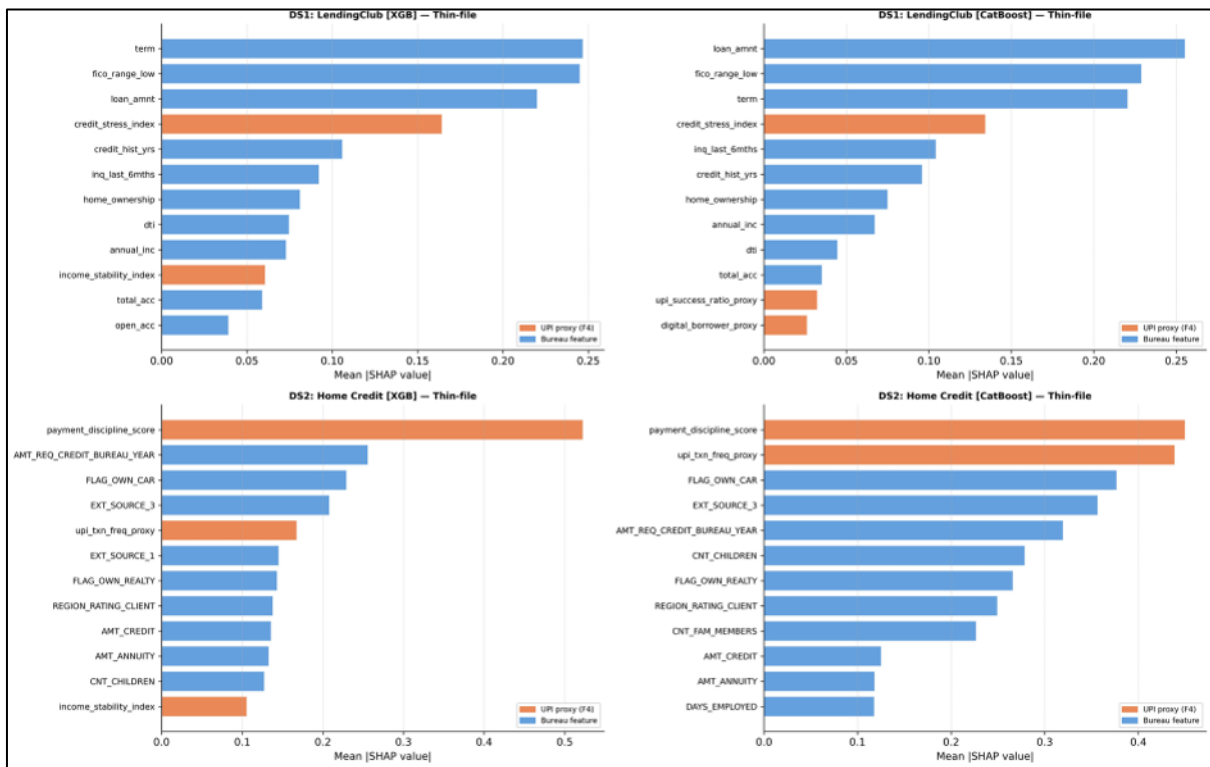


Figure 2 SHAP Feature Importance — Thin-file Segment (XGB and CatBoost, Both Datasets)

² Orange bars denote UPI proxy (F4) features; blue bars denote bureau features. Mean absolute SHAP value (x-axis) reflects each feature's average contribution to model output across 500 randomly sampled thin-file test observations. payment_discipline_score dominates both Home Credit panels (bottom row), exceeding every bureau-derived variable in predictive importance.

SHAP rankings are consistent across XGBoost and CatBoost — two algorithmically distinct models with different tree construction mechanisms. This cross-model consistency indicates that the rankings reflect genuine predictive information rather than model-specific artefacts. The convergence between proxy validation (Table 3) and SHAP attribution (Table 7) — features that empirically correlate with default are also the features the models rely on most heavily — constitutes strong empirical evidence that the proxy framework captures genuine UPI-observable behavioral signal.

7. Discussion

7.1. Why UPI Proxies Matter for Thin-file Borrowers

The concentration of SHAP importance in UPI proxies for thin-file borrowers (19–32%), compared to non-thin populations, reflects a fundamental property of credit scoring with sparse data. For thin-file borrowers, where bureau variables are absent or unreliable, behavioral proxies fill a genuine informational void. This is precisely the mechanism by which actual UPI data would function in practice — it provides signal when bureau data is least informative. CatBoost consistently achieves the highest thin-file AUC in both datasets (DS1: 0.6602; DS2: 0.7359) and delivers the largest Recall gain on DS1 (+9.95%). Its native `auto_class_weights` mechanism makes it particularly suited to thin-file populations with elevated default rates [8]. LR provides the only statistically confirmed improvement at the population scale (DeLong $p=0.04$), confirming that UPI proxy features provide linearly separable signal.

7.2. Implications for the Unified Lending Interface

NPCI's Unified Lending Interface (ULI), currently in pilot phase, is designed to enable lenders to access UPI behavioral data through a standardized API. The six behavioral dimensions defined in Table 2 — income stability, payment discipline, credit stress, transaction frequency, success rate, and digital engagement — provide a validated feature construction roadmap for ULI implementation. Lenders who implement the proxy framework today can replace proxy formulas with ULI equivalents without redesigning their scoring models.

7.3. Comparison with Published Work

Hlongwane et al. [11] achieve an AUC of 0.7936 on Home Credit using 215 features for all borrowers. The present study achieves comparable discriminative performance using 20 features while being the first to segment results by thin-file status and quantify UPI proxy contribution via SHAP attribution. Berg et al. [15], writing in the Review of Financial Studies, established the proxy methodology precedent applied here; this study extends it to the UPI context with the addition of empirical proxy validation, which Berg et al. do not provide.

7.4. Limitations

The F4 proxies approximate rather than replicate UPI transaction signals. Direct validation against actual UPI records remains a priority for future work when NPCI data access becomes available. `Digital_borrower_proxy` shows reversed correlations in both datasets and requires market-specific reformulation. DeLong tests are non-significant for the thin-file segment in DS1 due to the smaller thin-file test set size ($N=5,807$). This study does not establish causal identification.

8. Conclusion

This study makes four empirically supported contributions. First, it constructs six UPI behavioral proxy features from standard loan application variables, providing a publicly replicable framework for approximating UPI transaction signals in credit scoring. Second, it validates `payment_discipline_score` and `upi_success_ratio_proxy` through Spearman correlation against default outcomes in two independent datasets ($p<0.001$), establishing scientific construct validity. Third, SHAP attribution confirms these proxies account for 19.0%–32.2% of total predictive power for thin-file borrowers, with `payment_discipline_score` ranking as the single most important predictor in Home Credit above all bureau variables. Fourth, F4 features produce consistent positive Recall improvements for thin-file borrowers across all five models and both datasets.

The central finding — that adding UPI behavioral proxy features consistently improves default detection Recall for thin-file borrowers across all models and both independent datasets — confirms the generalizability of the proposed framework. CatBoost is recommended as the optimal model for thin-file credit scoring with UPI proxies, combining the highest thin-file AUC with the largest Recall gain and robust handling of class imbalance.

For India's 840 million credit-underserved adults, this proxy framework represents an immediately deployable tool that any lender can implement using data already collected at loan origination, and a validated feature engineering template for the forthcoming NPCI Unified Lending Interface.

Compliance with ethical standards

Acknowledgments

The authors thank Suresh Gyan Vihar University for institutional support. Both datasets are publicly available via Kaggle (LendingClub Loan Data and Home Credit Default Risk). Code is available upon request to support replication. AI tools were used to assist with code development; all research design, analysis, interpretation, and intellectual contributions are the authors' original work.

Disclosure of conflict of interest

No conflict of interest to be disclosed.

Statement of ethical approval

The present research work does not contain any studies performed on animals/humans subjects by any of the authors.

Statement of informed consent

Not applicable. This study uses publicly available, anonymized datasets and does not involve any individual participants, surveys, or interviews.

References

- [1] TransUnion CIBIL. (2022). Credit Market Indicator: Q4 2022 Report. TransUnion CIBIL Limited.
- [2] International Finance Corporation. (2018). The Unseen Sector: A Report on the MSME Opportunity in India. IFC World Bank Group.
- [3] Alok, S., Ghosh, P., Kulkarni, N., & Puri, M. (2024). Breaking barriers to financial access: Cross-platform digital payments and credit markets. NBER Working Paper 33259.
- [4] Ghosh, P., Vallee, B., & Zeng, Y. (2026). FinTech lending and cashless payments. *Journal of Finance*, 81(2), 1053–1101.
- [5] Agarwal, S., Alok, S., Ghosh, P., & Gupta, S. (2022). Financial inclusion and alternate credit scoring: Role of big data and machine learning in fintech. *Journal of Money, Credit, and Banking*, 54(7), 1–30.
- [6] Jagtiani, J., & Lemieux, C. (2019). The roles of alternative data and machine learning in fintech lending. *Journal of Financial Services Research*, 56(2), 189–224.
- [7] Chia, K. S. (2024). Optimal oversampling ratios for imbalanced credit classification. *Journal of Risk and Financial Management*, 17(3), 112–128.
- [8] Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: An interdisciplinary review. *Journal of Big Data*, 7(1), 94.
- [9] Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring. *European Journal of Operational Research*, 247(1), 124–136.
- [10] Albanesi, S., & Vamossy, D. F. (2024). Credit scores: Performance and equity. NBER Working Paper 32860.
- [11] Hlongwane, R., Ramaboa, K. K. K. M., & Mongwe, W. (2024). Enhancing credit scoring accuracy with a comprehensive evaluation of alternative data. *PLOS ONE*, 19(5), e0303566.
- [12] Chen, Y., Calabrese, R., & Martin-Barragan, B. (2024). Interpretable machine learning for imbalanced credit scoring datasets. *European Journal of Operational Research*, 312(1), 357–372.
- [13] DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves. *Biometrics*, 44(3), 837–845.
- [14] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.
- [15] Berg, T., Burg, V., Gombovic, A., & Puri, M. (2020). On the rise of FinTechs: Credit scoring using digital footprints. *Review of Financial Studies*, 33(7), 2845–2897.