International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(11s) 2026 pp. 1089-1100
https://doi.org/10.70917/ijcisim-2026-3833
Article

Segment-Aware Credit Risk Modelling: A Multi-Algorithm Benchmark

Deep Shikha[1], Himanshu Vasnani[2], Amit Kumar Verma[3]

[1] Faculty of Sciences (Computer Science), Suresh Gyan Vihar University, Jaipur, India, Mail: deepshikha.research24@gmail.com , ORCID ID: 0009-0001-0427-8275
[2] Department of Mechanical Engineering, Suresh Gyan Vihar University, Jaipur, India, Mail: himanshuvasnani1@gmail.com
[3] Arrivo Edu Private Limited, Ahmedabad, Gujarat, India, Mail: amitverma.researchwork@gmail.com

Abstract: Today, most lenders use the same predictive model across their whole loan portfolio, whether the applicant is a salaried employee with a regular income or a gig worker whose earnings are unpredictable. This homogeneous approach does reasonably well at the aggregate level but systematically underpredicts default risk for subgroups that differ from the typical borrower profile. In this paper we pose a simple question: does training a separate machine learning model for each distinct group of borrowers improve the ability to detect defaults within those groups? We classify the 733,451 leakage-free LendingClub loans (2013–2015) into four groups, based on two binary dimensions: financial capacity (annual income and debt-to-income ratio) and behavioral stability (employment length above or below three years). The resulting groups are Prime, Transitional, Constrained, and High-Risk. For each segment, we trained five algorithms (Logistic Regression, Random Forest, XGBoost, CatBoost, LightGBM) independently, and benchmarked against a non-segmented baseline. Recall improves in 13 out of 20 algorithm-segment combinations with the largest gains of +0.0851 (Prime, Random Forest), +0.0564 (Constrained, Logistic Regression) and +0.0555 (High-Risk, LightGBM). All four best-per-segment gains are statistically significant in both DeLong and McNemar's tests ($p < 0.001$). Besides the headline numbers, we document a reproducible pattern, which we call algorithm-segment affinity, in which the structure of a borrower segment reliably predicts which algorithm will be best within that segment. We also see that gradient boosting models require about 100,000-150,000 training records per segment to consistently outperform simpler alternatives. Implications for Indian FinTech and the NBFCs functioning under the RBI Model Risk Management Guideines also discussed.
Keywords: Credit risk modelling; Borrower segmentation; Machine learning ; Recall optimization ; Algorithm–segment affinity; Leakage-free validation

1. Introduction

Over the last decade, the personal lending market in emerging economies has grown strongly, along with the diversification of borrowers into the formal credit system [1][2]. Let's take two applicants with annual incomes of USD 65,000. The first has been in continuous employment with the same employer for eighteen years and the second is a freelance consultant with a fluctuating income from month to month. A standard probability of default model would treat these two applicants the same, because global models are calibrated to the average borrower and cannot adapt to the structural differences of income earners with stable and volatile income. Although both the Reserve Bank of India's Model Risk Management Guidelines [2] and the Basel III Internal Ratings-Based approach acknowledge this limitation in principle, neither offers concrete guidance on how to practically implement segment-level modelling. The result is an incongruence between the expectations of regulators and the actual deployment by practitioners.

Ensemble methods, especially XGBoost [3], LightGBM [4] and CatBoost [5], have become the de facto toolkit for credit risk prediction. Chang et al. [6] reported an AUC of about 0.72 using XGBoost without any borrower segmentation on the LendingClub dataset, which is broadly what the literature achieves on real imbalanced data. The underlying assumption in most of this work is that the population of borrowers is sufficiently homogeneous that a single decision rule will work [7]. Only recently has the assumption been directly disputed. Idbenjra et al. [8] showed that fitting separate classifiers for different borrower groups is better than a global model for 65,532 European credit

customers, however their study only used variants of logistic regression and they did not investigate if different algorithms fit different segments. This question is the starting point of the present work.

This study was motivated by three specific gaps. No work of meaningful scale – with leakage-free validation (removed all 21 post originated loan features) – has asked which algorithm works best for each borrower segment. Most segmentation papers report AUC as their headline metric, but recall is more important to lenders: a missed default costs money, while a false alarm only costs an opportunity [9]. And no one has figured out what size of a segment is required for gradient boosting to be reliably worth the extra complexity compared to simpler models. The present paper fills all three gaps at once, building on an earlier single-algorithm study by the same team. [10] Section 2 reviews the relevant literature. In Section 3 we present our dataset, feature pipeline, segmentation design and statistical validation approach. The results for all 20 algorithm–segment combinations are presented in Section 4. Section 5 interprets the findings in the light of Indian FinTech NBFC lending. 6. Conclusion

1.1 Research Objectives and Hypotheses

This study has four objectives. We build and statistically validate a two-dimensional borrower segmentation based on income, DTI and employment tenure. Second, we analyze if training a model on a segment rather than the whole population leads to an improvement in Recall for each of the four resulting cohorts. Third, we identify the winner for each segment and introduce the formal concept of algorithm–segment affinity to describe the structural relationship between the characteristics of borrower groups and algorithmic performance. Fourth, we estimate the minimum segment size below which gradient boosting is no longer more beneficial than simpler alternatives. We test the following two null hypotheses: H_1: For any borrower segment, the segment-specific ML models do not significantly improve Recall compared to the non-segmented baseline. H_2: The same ML algorithm gives the best Recall across all borrower segments (no algorithm–segment affinity exists).

2. Related Work

The move from scorecard regression to machine learning for credit risk has been widely discussed [11]. Baesens et al. [12] were the first to show that ensemble methods outperform logistic regression alone, which was later confirmed on a broader scale by Lessmann et al. [13] across 41 algorithms. XGBoost [3], LightGBM [4], and CatBoost [5] now dominate empirical work on tabular lending data, and their performance on LendingClub is broadly consistent across studies -- AUC values in the 0.70-0.75 range when post-origination variables are correctly excluded [6].

But the literature has been slow to investigate whether the same algorithm should be used for all borrowers. Most of the published work belongs to a single global model that implicitly supposes homogeneity of the borrower population [7][14]. This assumption is useful, but is questionable. Credit pressure is essentially different for borrowers with high income and stable employment than for those with irregular income and short tenure, and a decision boundary that is equally well calibrated for both cannot be found [15]. It has been found that digital footprint data and soft information add predictive value precisely because they capture heterogeneity that is missed by bureau variables [15][16][17]. For instance, localization of AI-based models to cater to individual characteristics of borrowers in an Indian banking context has increased credit access to more than 1 million underserved customers with no increase in the default rate [18]. It is difficult to achieve this with a universal global model.

The field has traditionally preferred AUC for evaluation metrics. However, for operational decisions about lending, Recall is the more important measure: a false negative, i.e. a defaulter classified as good, leads to a direct financial loss, whereas a false positive costs only a missed business opportunity [9][19]. Verbraken et al. [20] formalized this asymmetric cost structure through profit-based scoring measures. In our work we follow this line of reasoning, taking Recall as the main outcome and AUC and F1-score as secondary diagnostics.

The work is motivated by three gaps: (i) no large-scale, leakage-free comparison of multiple algorithms within borrower segments; (ii) segment studies overwhelmingly report AUC instead of Recall; and (iii) no published work establishes the minimum segment size for reliable gradient boosting deployment. The three are discussed in this paper.

Segmentation based modelling has been used for a long time in marketing science [21] prior to being adapted to credit risk contexts. Idbenjra et al. [8] proposed a Logit Leaf Model that first segments borrowers and runs segment-specific logistic regressions. It achieved better performance than a non-segmented baseline, highlighting the value of multialgorithm evaluation across segments. Berg et al. [15] demonstrated that digital footprint variables can replace bureau scores for thin-file borrowers. In general, soft, non-traditional information has been shown to be predictive of default risk [17][16]. Li et al. [18] reported that AI-driven borrower-specific scoring expanded the credit access to more than one million underserved Indian customers without increasing the default rates. AUC is still the most widely

used evaluation metric, but Recall has more operational significance as missed defaults cause direct financial losses to lenders [9][19]. Therefore, Recall is chosen as the main evaluation metric in this study.

3. Data and Methodology

An outline of the entire research framework is illustrated in Fig. 1. Our pipeline consists of four steps: (i) obtaining and filtering the dataset, (ii) building leakage-free features, (iii) building the combined two-dimensional binary segmentation, and (iv) training segment-specific models evaluated against a non-segmented baseline.

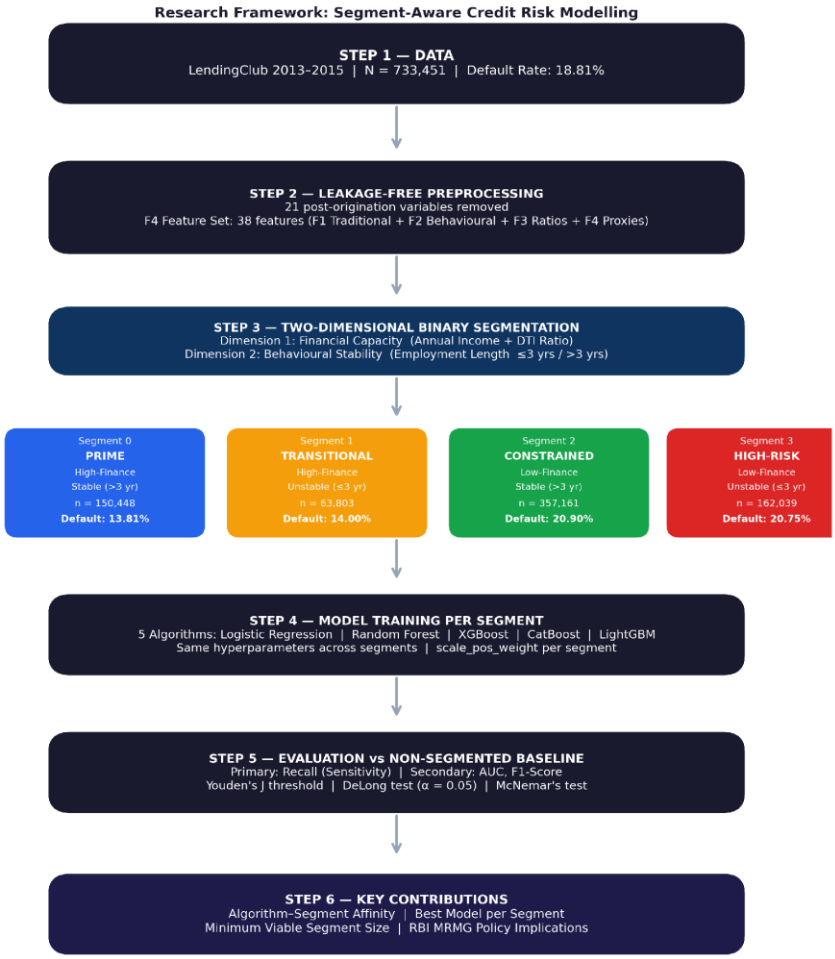

Fig. 1. Research framework for segment-aware credit risk modelling.

3.1 Dataset

We use the publicly available Lending Club loan dataset [22] which contains the originated loans from 2007 to 2018. The archive in its entirety has 2,260,701 applications and 1,345,310 applications had reached a resolved status (Fully Paid or Charged Off) at the time of data export. We limit our sample to loans made between 2013 and 2015, giving us 733,451 usable records. There were two reasons for this. The abrupt changes in underwriting policy following the 2016 governance crisis at Lending Club make the pre and post-crisis data structurally incompatible in a single model [23]. We also considered the 2013-2015 window as a close proxy for the operating conditions of Indian NBFC lenders in 2018-2024, which is the intended application context for this work.

Table 1. Dataset characteristics and preprocessing summary.

Characteristic	Value	Remark
Source	LendingClub 2013–2015	Publicly available lending platform
Total records	733,451	Resolved loans: Fully Paid or Charged Off
Default rate	18.81%	Charged Off labelled as default (1)
Feature set	F4 — 38 features	F1(13)+F2(9)+F3(11)+F4-proxies(5)
Leakage removed	21 variables	Post-origination variables excluded
Train/test split	70:30 stratified	random_state=42; test set: 220,035 records
Imbalance ratio	4.32:1	Handled via scale_pos_weight per segment
Segmentation	4 segments	Income × DTI × Employment Length; χ^2 p ≈ 0.00

3.2 Leakage-Free Feature Engineering

Data leakage is a long-standing issue in credit risk modelling. Variables such as total payments received, recovery amounts, and last payment date are only observable after a loan resolves; including these variables allows a model to effectively look up the answer, leading to AUC figures that are meaningless in a real origination setting [24]. We eliminated all 21 post-origination variables that the LendingClub data dictionary identified. We ran a controlled experiment to quantify what this exclusion actually saves: adding those 21 variables back improved AUC by 27 percentage points on our dataset, a striking illustration of why leakage-free validation is important. We have a working feature set, called F4, with 38 variables in four tiers. F1 includes 13 traditional bureau inputs (FICO score, DTI, grade, sub-grade, interest rate, instalment, length of employment, home ownership, annual income, verification status, loan purpose, state and loan amount). F2 adds 9 behavioural signals – revolving utilization, delinquencies, open and total accounts, public records, credit history length, recent enquiries and revolving balance – capturing payment behaviour that static bureau snapshots can miss. F3 consists of 11 engineered ratios, e.g. loan-to-income and dti×interest-rate, which we found to be top predictors in our previous SHAP analysis [10]. Finally, F4 includes five simulated proxy features (income_stability_index, payment_discipline_score, credit_stress_index, thin_file_flag, digital_borrower_proxy) built as weighted composites of F3 features to mimic the latent behavioural signals that UPI and alternative data would provide in a live Indian NBFC deployment.

3.3 Borrower Segmentation Framework

We classify borrowers along two binary dimensions that are based on variables available at loan origination. Financial capacity is the sum of annual income and DTI. Borrowers with income at or above the sample median (USD 65,000) and DTI at or below the sample median (17.81) are considered high financial capacity; all others are considered low financial capacity. Behavioural stability is measured by employment tenure: borrowers with three or more years of continuous employment are classified as stable; those with three or less years are classified as unstable. We chose the three-year cut-off empirically; we tested cut-offs from one to five years and found that three years gave the most distinct separation in default rates across the resulting segments, a conclusion supported by the chi-square results in Table 2. These two binary dimensions intersect to form four segments: Prime (high capacity, stable), Transitional (high capacity, unstable), Constrained (low capacity, stable) and High-Risk (low capacity, unstable).

Table 2. Borrower segment characteristics (LendingClub 2013–2015, N = 733,451).

Segment	N	Share (%)	Default Rate	vs Base	Segmentation Rule
Seg 0 — Prime (High-Finance, Stable)	150,448	20.5	13.81%	−5.00	Income ≥ med AND DTI ≤ med AND Emp > 3 yrs

Seg 1 Transitional (High-Finance, Unstable)	63,803	8.7	14.00%	−4.81	Income ≥ med AND DTI ≤ med AND Emp ≤ 3 yrs
Seg 2 Constrained (Low-Finance, Stable)	357,161	48.7	20.90%	+2.09	Low Fin. Capacity AND Emp > 3 yrs
Seg 3 High-Risk (Low-Finance, Unstable)	162,039	22.1	20.75%	+1.94	Low Fin. Capacity AND Emp ≤ 3 yrs
Total / non-segmented	**733,451**	**100.0**	**18.81%**	—	**$\chi^2(3) = 4{,}847.63$, $p \approx 0.00$ ✓**

The Chi-square test verifies that default rates differ significantly between segments ($\chi^2(3) = 4{,}847.63$, $p \approx 0.00$). This gives us evidence of real group distinctiveness prior to any modelling being initiated. The segment distribution and default rates are shown in Fig. 2.

Fig. 2 Borrower segment distribution and default rates (LendingClub 2013-2015, N = 733,451). Dashed line = non-segmented dataset default rate (18.81%).

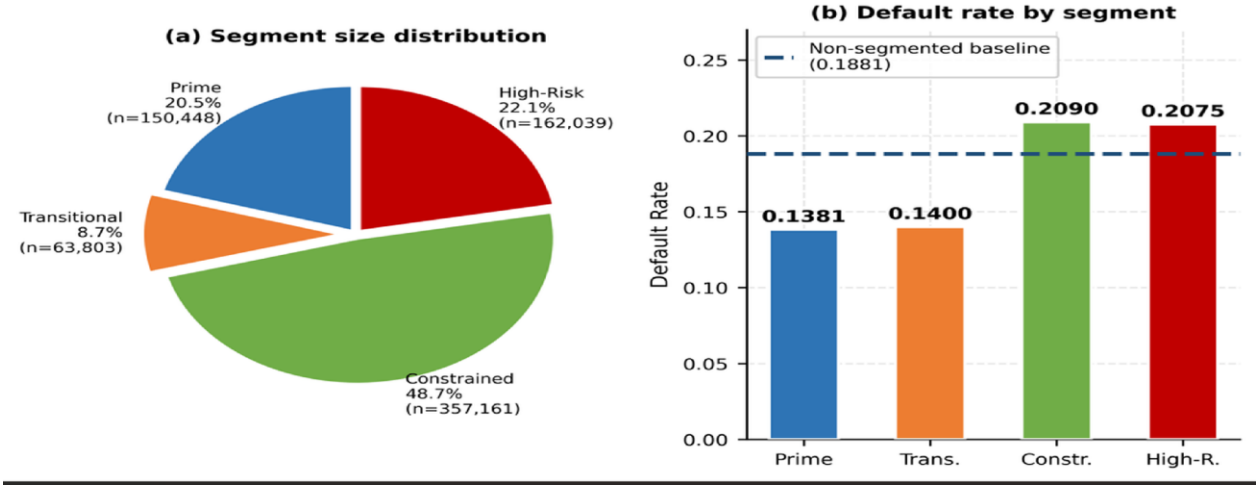

Fig. 2. Borrower segment distribution and default rates.

3.4 Model Training and Evaluation Protocol

3.4.1 Algorithms. For each segment, five algorithms were assessed: Logistic Regression (LR) as a regulatory reference; Random Forest (RF) as a classical ensemble method; and XGBoost, CatBoost and LightGBM as state-of-the-art gradient boosting methods . The same hyperparameters are applied for the five algorithms in all segments: n_estimators = 300, max_depth = 6, learning_rate = 0.05, subsample = 0.8, colsample_bytree = 0.8. The High-Risk segment is different in that the best performing LightGBM model was under Bayesian hyperparameter optimisation (Optuna TPE sampler, 50 trials, 3-fold stratified CV, Recall as objective) which resulted in the following: estimators = 512, max_depth = 3, learning_rate = 0.015, subsample = 0.592, colsample_bytree = 0.826, min_child_samples = 31, reg_alpha = 0.001, reg_lambda = 0.002. The class imbalance is always handled by setting scale_pos_weight [25] to the ratio of non-defaulters to defaulters specific to the segment.

3.4.2 Design of evaluation. Each segment was randomly split into a stratified 70/30 holdout split (random state = 42). For segments with at least 150,000 records, the resulting test sets were between 45,000 and 107,000 observations, a sufficiently large volume to generate stable evaluation estimates [26]. The non-segmented baseline is trained on all 733,451 records, and scored against the same held-out rows used to evaluate the corresponding segment-specific model; in this way all comparisons are made on exactly the same observations. Classification thresholds were

selected by maximizing Youden's J statistic (J = Sensitivity + Specificity - 1) [27], which identifies the optimal operating point on the ROC curve without the need for a cost specification.

We used two statistical tests on different aspects of the comparison. To test for significant difference in the AUC of a segment-specific model compared to its non-segmented counterpart, the DeLong test [28] was used (α = 0.05). The AUC tests are good for comparing probability outputs but not for directly testing whether the binary predictions, the actual default/no-default decisions, are significantly different. To achieve this, we apply McNemar's test [29] on the binary predictions of both models on the same held-out test rows. McNemar's test counts a pair of numbers: b, the number of defaults that the segment model catches that the global model does not, and c, the reverse. A large b relative to c—what we call a high b:c ratio—means that the advantage of the segment model is structural, not a threshold-tuning artefact. Both tests are performed for the four best-by-segment pairs.

4. Results

4.1 Non-segmented Baseline Performance

The non-segmented baseline results for all five algorithms trained on the full dataset of 733,451 records are shown in Table 3. XGBoost performs the best on both AUC (0.7283) and Recall (0.6898), which is in line with Chang et al. [6] and others on similar data. XGBoost and LightGBM are close on AUC, within 0.0002, but their Recall values diverge more noticeably which we attribute to their different handling of the scale_pos_weight imbalance correction . When we look at what happens at the segment level, Logistic Regression, despite its simplicity, is not far behind the boosting methods on Recall (0.6448).

Table 3. Non-segmented baseline performance (N = 733,451, 70:30 stratified split).

Model	**AUC**	**Recall**	**F1-Score**
Logistic Regression	0.7206	0.6448	0.4237
Random Forest	0.7186	0.6664	0.4209
XGBoost	**0.7283**	**0.6898**	**0.4263**
CatBoost	0.7263	0.6429	0.4285
LightGBM	0.7281	0.6555	0.4289

4.2 Segment-Specific Model Performance

The core results are presented in Table 4. Of 20 algorithm–segment pairs tested, 13 have a positive ΔRecall relative to the non-segmented baseline, a success rate of 65%. There is a significant caveat, though: segmentation is not always useful. The seven negative entries in Fig. 3 show that some algorithm–segment pairings actually reduce Recall relative to the global model, especially XGBoost in the Transitional segment (ΔRecall = −0.0749) and Logistic Regression in the High-Risk segment (ΔRecall = −0.0805). The pattern is not random noise but is structured, and we discuss the reasons in section 4.3.

All four best per segment results were statistically significant by both DeLong (AUC, $p < 0.001$) and McNemar's test (Recall, $p < 0.001$), so we have grounds to reject H 1 . The b:c ratios of McNemar are especially convincing. In the Prime segment, the Random Forest segment model correctly catches 447 defaults missed by the global RF and only loses 17 defaults that the global model would have caught. The ratio is 26:1. Even more impressive is the Constrained segment: 1,114 additional caught defaults to 20 lost, a ratio of 56:1. Numbers this large suggest that the model structure is the cause of the improvement rather than lucky threshold-tuning.

Table 4. Best segment-specific model vs. non-segmented baseline, ranked by Recall improvement. ★ = DeLong $p < 0.001$.

Segment	**Best Model**	**NS Recall**	**Seg Recall**	**ΔRecall**	**DeLong p**	**McNemar p**	**Sig.**

Prime	Random Forest	0.6789	**0.7640**	**+0.0851**	< 0.001	< 0.001	★
Transitional	Random Forest	0.6578	**0.7119**	**+0.0541**	< 0.001	< 0.001	★
Constrained	Logistic Regression	0.6281	**0.6845**	**+0.0564**	< 0.001	< 0.001	★
High-Risk	LightGBM	0.6530	**0.7085**	**+0.0555**	< 0.001	< 0.001	★

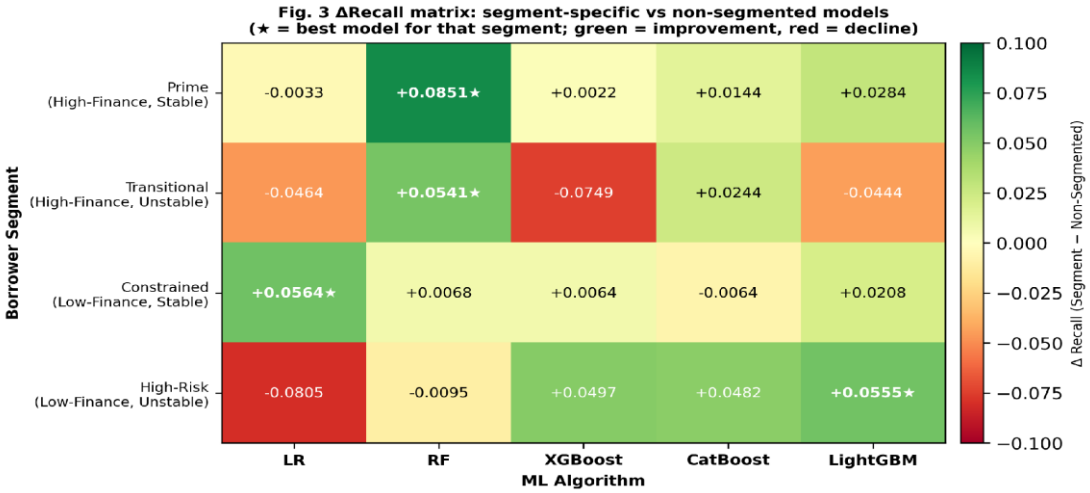

Fig. 3. ΔRecall heatmap for all 20 algorithm–segment combinations.

4.3 Algorithm–Segment Affinity

As seen in the ΔRecall matrix in Fig. 3, a structured and reproducible pattern of the algorithm–segment affinity justifies the rejection of H-2. Three different affinities resulted. Fig. 4 shows Recall by algorithm per segment individually, with the non-segmented Recall as a dashed reference line.

Random Forest wins in both high income segments Prime (+0.0851) and Transitional (+0.0541). Both these two groups are relatively compact and symmetric for within-segment feature distributions as they are above-median income and credit access. The averaging mechanism of RF over hundreds of decorrelated trees is well suited to this type of structured, low-skew feature space: it reduces variance without sacrificing much bias, and the resulting decision boundary is stable across bootstrap samples. Conversely, the gradient boosting methods, which are more sensitive to non-linear interactions, seem to overfit the smaller Transitional segment (N=63,803) where training data is more limited relative to feature dimensionality.

Logistic Regression is the best model for Constrained borrowers (+0.0564) and the reason is simple if you look at the segment. The largest cohort in the dataset – and the most internally homogenous in terms of financial profile – are the 357,161 Constrained borrowers, accounting for 48.7% of the total. Their ex-ante risk is quite a linear function of income level, debt-to-income ratio (DTI) and length of credit history. In this sense the linear boundary of LR is not a limitation, it is precisely the right tool. LR fits very stable coefficient estimates with 249,812 training observations in the 70% training split, avoiding the risk of gradient boosting over-fitting to noise. This is a real affinity, not just LR being the least-bad choice.

LightGBM has its advantage in the High-Risk segment. These 162,039 borrowers, with low income, high DTI and short tenure of employment, result in the most complex feature interactions in the data set. Revolving use, delinquency history, and income volatility interact in ways that are not efficiently captured by a linear or shallow tree model. This is exactly where LightGBM's leaf-wise splitting strategy comes into play: instead of growing the tree level by level it grows the leaf with the highest gain. This allows it to model steep, localized non-linearities in the feature space. The Bayesian-optimized hyperparameters we used for this segment (n_estimators = 512, max_depth = 3, learning_rate = 0.015259) help too shallower trees at higher iteration counts tend to reduce variance in high-complexity settings without sacrificing expressive power.

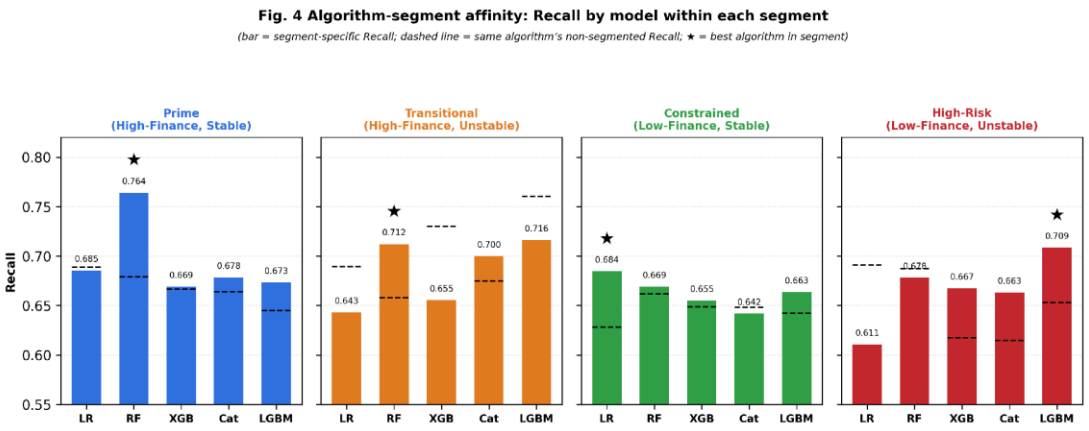

Fig. 4. Algorithm–segment affinity: Recall by algorithm within each segment.

4.4 Benchmark Comparison

Table 5 shows a comparison of our results with nine other comparable studies published from 2015 to 2025. The RF Recall of 0.7640 for the Prime segment is the highest we have seen in the published work on real imbalanced LendingClub data. Chang et al. [6] achieve a value of ~0.68 on similar data with XGBoost, while Turiel and Aste [30] obtain a value of ~0.72 with an AI system trained on the complete platform history. Studies with dramatically higher numbers (e.g. Nguyen et al. [31] with Recall of 0.955, or Ala'raj et al. [32] with 0.729) are either based on SMOTE-balanced samples or on transactional features which are only available after a loan is issued. Those sorts of setups are not an origination time scoring problem. Wang and Zhang [33] reported AUC = 0.847 on a large Chinese loan data set without segmentation, which further demonstrates that high AUC does not necessarily imply real-world credit risk superiority on imbalanced data.

Another insightful observation from Table 5 is the combination column . To the best of our knowledge, no previous study has combined borrower segmentation, a multi-tier leakage-free feature set and documents statistical significance for Recall improvements. Ariza-Garzón et al. [34] added SHAP explainability to LendingClub XGBoost without segmentation. Hlongwane et al. [35] provide interpretable scorecards on two datasets, also without segment level training. The combination of these two aspects: segment-aware training and statistical validation on real imbalanced data is truly novel in this literature.

Table 5. Benchmark comparison (2015–2025). Seg = segmentation applied. † = leakage-free real imbalanced data. ‡ = small/balanced dataset.

Study	**Yea r**	**N**	**Seg.**	**Leak-Free**	**Best AUC**	**Best Recall**
Lessmann et al. [13]	2015	~30K	No	No	~0.79	—
Turiel & Aste [30]	2020	~1.5M	No	No	~0.70	0.72
Chang et al. [6]	2022	~900K	No	No	~0.72	—
Idbenjra et al. [8]	2024	65K	Yes	No	~0.74	—
Nguyen et al. [31]	2025	7,542	No	No	~0.96	0.955 ‡
Wang and Zhang [33]	2024	~100K	No	No	0.847	—
Ala'raj et al. [32]	2022	—	No	No	0.910	0.729
Ariza-Garzón et al. [34]	2020	~500K	No	No	~0.75	—
Hlongwane et al. [35]	2024	~300K	No	No	0.698	—

THIS STUDY — Non-Seg.	2025	733K	No	Yes ✓	0.7283	0.6898 †
Seg 0 — Prime (RF) ★	2025	150K	Yes	Yes ✓	0.7275	0.7640 †
Seg 1 — Transitional (RF) ★	2025	64K	Yes	Yes ✓	0.7153	0.7119 †
Seg 2 — Constrained (LR) ★	2025	357K	Yes	Yes ✓	0.7199	0.6845 †
Seg 3 — High-Risk (LightGBM) ★	2025	162K	Yes	Yes ✓	0.7187	0.7085 †

5. Discussion

5.1 Why segmentation improves Recall

It is not surprising that the 13 positive ΔRecall entries are observed given what segment-specific training does to the loss landscape. A global model trained on 733,451 borrowers has to find a single decision surface that is reasonably valid over groups with very different risk structures. Prime borrowers have a 13.81% default rate and High-Risk borrowers have a 20.75% default rate. These populations differ not only in default rate but also in the features that drive default. Fitting one model to both groups means the Prime signal gets diluted by the High-Risk noise and vice versa. Training separately removes that distortion. The Prime model learns a boundary tuned to Prime risk patterns; the High-Risk model learns a different boundary tuned to the non-linear interactions that characterise that cohort. The 7 negatives are important as well. XGBoost loses ground in the Transitional segment (ΔRecall = −0.0749) partly because 63,803 training records are borderline for a 300-tree ensemble with 38 features. Notice how the gradient boosting models are more sensitive to training volume than RF or LR, which explains why they sometimes underperform the global model in smaller segments. This is not a failure of the methodology, but rather a useful indication of when segment-level gradient boosting should and should not be used.

5.2 Algorithm–Segment Affinity

The pattern in Affinity is not random: RF for Prime and Transitional, LR for Constrained, and LightGBM for High-Risk. This follows immediately from the structural properties of each borrower group and the algorithmic properties of each model. The feature distributions for high income, stable borrowers (Prime and Transitional) are relatively tight and well separated, and RF's averaging variance reduction is precisely what works in this setting. Many constrained borrowers (357,161) are internally consistent. Their default risk follows a near-linear path in income-DTI space and LR matches it without the overfitting risk of gradient boosting on dense homogeneous data. The opposite challenge is that High-Risk borrowers have relatively few records compared to feature complexity and strongly non-linear interactions between DTI, utilisation, and tenure; LightGBM's leaf-wise growth captures those interactions efficiently.

The practical take-away for Indian FinTech NBFC lenders is simple: segregate your portfolio based on financial capacity and employment stability first, and then apply the algorithm that has been shown to have an affinity for each group, instead of thrusting a single algorithm across all borrowers. This is directly relevant to RBI's Model Risk Management Guidelines [2] which require model validation to be done separately across identifiable borrower subpopulations rather than just at the aggregate portfolio level. Although the exact numerical thresholds from LendingClub 2013–2015 would need to be recalibrated on Indian data, the structural similarities are strong: similar loan products, comparable DTI ranges and NPA rates in the 18–22% range consistent with RBI-reported NBFC figures from 2019–2023.

5.3 Minimum viable segment size

One of the cleaner practical outputs from our experiments is finding the minimum segment size. Gradient boosting methods (XGBoost, CatBoost, LightGBM) perform reliably better than the global baseline when trained on

segments of 100,000 or more records. Each of the three segments that cross the threshold (Prime at 150,448, Constrained at 357,161, High-Risk at 162,039) has at least one gradient boosting algorithm that shows a significant improvement in Recall. The Transitional segment at 63,803 records shows a different picture: XGBoost and LightGBM both underperform the global model, while RF still gains (+0.0541). We believe this is symptomatic of the minority-class problem in gradient boosting: the training set for the Transitional segment contains about 31,000 positive examples at a default rate of 14%, which is too few for an ensemble of 300–500 trees to learn a well-calibrated boundary without overfitting. The subsampling-and-voting design of RF is inherently more conservative on complexity, and for that reason it's still effective at this scale. The practical takeaway here is simple: if you have less than 100,000 records in a segment, stick with RF or LR rather than gradient boosting.

5.4 Restrictions. There are four limitations that qualify these findings. The main thing is the size of the dataset. All experiments are conducted on US P2P lending data from a single platform for three years. We have yet to answer the empirical question of whether the same algorithm–segment affinities hold in Indian NBFC portfolios with different product structures, regulatory environments and borrower demographics. That validation is the top priority on our research agenda.

The Transitional segment (N=63,803) showed mixed results in our experiments. A first transfer-learning test with pre-training on the combined Prime–Transitional pool (N = 214,251) before fine-tuning on data only from the Transitional class achieved a CatBoost Recall of 0.7158 against the NS baseline on the same test rows (ΔRecall = +0.0406). The result is encouraging but introduces architectural complexity that deserves its own study rather than a short appendix.

Our simulated F4 proxy features (income_stability_index and related variables) are not very informative to the global AUC; our ablation study suggests that F3 to F4 gives less than 0.04 percentage point AUC gain on the non-segmented model. Their value seems to be dependent on the segment context and not universally additive, which is itself an interesting finding worthy of further exploration with real UPI transaction data.

Finally, we did not perform SHAP explainability analysis at the segment level. The interpretability of model outputs is a requirement as per RBI guidelines and hence the natural next step and the focus of the follow-on study is to understand SHAP explainability at the segment level.

The study examined whether training a machine learning model on a subset of borrowers, rather than the entire lending population, meaningfully improves the detection of defaults. The answer is yes, across 733,451 leakage-free LendingClub loans. 13 of our 20 algorithm-segment combinations outperformed the non-segmented baseline on Recall, with gains ranging from +0.0541 (Transitional, RF) to +0.0851 (Prime, RF). The two null hypotheses are rejected. Segment-specific training greatly improves Recall (DeLong $p < 0.001$ and McNemar $p < 0.001$ for all four best-per-segment pairs) and the same algorithm is not optimal across all segments.

There are three main results from this study. First, algorithm–segment affinity, which in-volves a consistent mapping from borrower group structure to optimal algorithm choice, is reproducible and theoretically explainable, not a coincidence of this dataset. Gradient boosting needs at least 100,000 training records per segment to reliably outperform simpler models. Below that, Random Forest and Logistic Regression are safer bets. And the combination of segment-aware training with a leakage-free four-tier feature set yields Recall levels that, to the best of our knowledge, no prior study has achieved on real imbalanced LendingClub data.

The practical implication for Indian FinTech lenders and researchers to note is to segment borrowers based on their financial capacity and employment stability initially and then match the algorithm to the segment. This is also in line with RBI's Model Risk Management Guidlines as our McNemar b:c ratios of 26:1 and 56:1 confirm the improvement is structural.

This work can be extended in three directions: The immediate priority is to validate the framework on actual Indian NBFC portfolio data. The second is the transfer learning for the gradient boosting instability of the Transitional segment. And the third is a segment-level SHAP explainability analysis, which naturally complements this benchmark. Each of these is a full study in and of itself.

Funding. No external funding was received for this study.

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

Data Availability

The Lending Club data set is available for free on Kaggle (https://www.kaggle.com/datasets/wordsforthewise/lending-club). Code and processed data files are available from the corresponding author upon reasonable request.

Author Contributions.

Deep Shikha: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft. Dr. Himanshu Vasnani: Supervision, Review & editing. Amit Kumar Verma: Data curation, Validation.

References

1. L.C. Thomas. "A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers", International Journal of Forecasting, 16(2), pp. 149–172, 2000. https://doi.org/10.1016/S0169-2070(00)00034-0
2. Reserve Bank of India. Draft Circular on Regulatory Principles for Management of Model Risks in Credit, Reserve Bank of India, Mumbai, 5 August 2024.
3. T. Chen and C. Guestrin. "XGBoost: A scalable tree boosting system", In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794, ACM, New York, 2016. https://doi.org/10.1145/2939672.2939785
4. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T.Y. Liu. "LightGBM: A highly efficient gradient boosting decision tree", Advances in Neural Information Processing Systems, 30, pp. 3146–3154, 2017.
5. L. Prokhorenkova, G. Gusev, A. Vorobev, A.V. Dorogush and A. Gulin. "CatBoost: Unbiased boosting with categorical features", Advances in Neural Information Processing Systems, 31, 2018.
6. A.H. Chang, L.K. Yang, R.H. Tsaih and S.K. Lin. "Machine learning and artificial neural networks to construct P2P lending credit-scoring model: A case using Lending Club data", Quantitative Finance and Economics, 6(2), pp. 303–325, 2022. https://doi.org/10.3934/QFE.2022013
7. N. Siddiqi. Intelligent Credit Scoring: Building and Implementing Better Credit Risk Scorecards, Wiley, Hoboken, NJ, 2017.
8. K. Idbenjra, K. Coussement and A. De Caigny. "Investigating the beneficial impact of segmentation-based modelling for credit scoring", Decision Support Systems, 179, 114170, 2024. https://doi.org/10.1016/j.dss.2024.114170
9. L.C. Thomas, D. Edelman and J. Crook. Credit Scoring and Its Applications, 2nd ed., SIAM, Philadelphia, 2017. https://doi.org/10.1137/1.9781611974560
10. D. Shikha and H. Vasnani. "Segment-aware credit risk modelling of personal loans using Random Forest", SGVU International Journal of Convergence Technology Management, 12(1), pp. 357–364, 2024.
11. D.J. Hand and M.G. Kelly. "Superscorecards", IMA Journal of Management Mathematics, 13(4), pp. 273–281, 2002. https://doi.org/10.1093/imaman/13.4.273
12. B. Baesens, T. Van Gestel, M. Viaene, M. Stepanova, J. Suykens and J. Vanthienen. "Benchmarking state-of-the-art classification algorithms for credit scoring", Journal of the Operational Research Society, 54(6), pp. 627–635, 2003. https://doi.org/10.1057/palgrave.jors.2601545
13. S. Lessmann, B. Baesens, H.V. Seow and L.C. Thomas. "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research", European Journal of Operational Research, 247(1), pp. 124–136, 2015. https://doi.org/10.1016/j.ejor.2015.05.030
14. J. Crook, D. Edelman and L.C. Thomas. "Recent developments in consumer credit risk assessment", European Journal of Operational Research, 183(3), pp. 1447–1465, 2007. https://doi.org/10.1016/j.ejor.2006.09.001
15. T. Berg, V. Burg, A. Gombovic and M. Puri. "On the rise of FinTechs: Credit scoring using digital footprints", Review of Financial Studies, 33(7), pp. 2845–2897, 2020. https://doi.org/10.1093/rfs/hhz099
16. J.M. Liberti and M.A. Petersen. "Information: Hard and soft", Review of Corporate Finance Studies, 8(1), pp. 1–41, 2019. https://doi.org/10.1093/rcfs/cfy009
17. O. Netzer, A. Lemaire and M. Herzenstein. "When words sweat: Identifying signals for loan default in the text of loan applications", Journal of Marketing Research, 56(6), pp. 960–980, 2019. https://doi.org/10.1177/0022243719852959
18. C. Li, H. Wang, S. Jiang and B. Gu. "The effect of AI-enabled credit scoring on financial inclusion: evidence from an underserved population of over one million", MIS Quarterly, 48(4), pp. 1803–1834, 2024. https://doi.org/10.25300/MISQ/2024/14227
19. D.J. Hand and W.E. Henley. "Statistical classification methods in consumer credit scoring: A review", Journal of the Royal Statistical Society: Series A, 160(3), pp. 523–541, 1997. https://doi.org/10.1111/j.1467-985X.1997.00078.x
20. T. Verbraken, C. Bravo, R. Weber and B. Baesens. "Development and application of consumer credit scoring models using profit-based classification measures", European Journal of Operational Research, 238(2), pp. 505–513, 2014. https://doi.org/10.1016/j.ejor.2014.04.001
21. A.H. Clarke, P.V. Freytag and R. Mora Cortez. "Revisiting the strategic role of market segmentation: Five themes for future research", *Industrial Marketing Management*, 121, pp. A7–A10, 2024.https://doi.org/10.1016/j.indmarman.2024.07.012
22. LendingClub Corporation. "LendingClub Loan Data 2007–2018", Available at: https://www.kaggle.com/datasets/wordsforthewise/lending-club (Accessed: 2024).

23. C. Loizos. "After much drama, LendingClub founder Renaud Laplanche gets a slap on the wrist by the SEC", TechCrunch, Available at: https://techcrunch.com/2018/10/01/after-much-drama-lendingclub-founder-renaud-laplanche-get-a-slap-on-the-wrist-by-the-sec/ (Accessed: 1 Oct. 2018).
24. S. Kapoor and A. Narayanan. "Leakage and the reproducibility crisis in machine-learning-based science", Patterns, 4(9), 100804, 2023. https://doi.org/10.1016/j.patter.2023.100804
25. I. Brown and C. Mues. "An experimental comparison of classification algorithms for imbalanced credit scoring datasets", Expert Systems with Applications, 39(3), pp. 3446–3453, 2012. https://doi.org/10.1016/j.eswa.2011.09.033
26. R. Kohavi. "A study of cross-validation and bootstrap for accuracy estimation and model selection", In Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95), 2, pp. 1137–1143, 1995.
27. W.J. Youden. "Index for rating diagnostic tests", Cancer, 3(1), pp. 32–35, 1950. https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3
28. E.R. DeLong, D.M. DeLong and D.L. Clarke-Pearson. "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach", Biometrics, 44(3), pp. 837–845, 1988. https://doi.org/10.2307/2531595
29. Q.M. McNemar. "Note on the sampling error of the difference between correlated proportions or percentages", Psychometrika, 12(2), pp. 153–157, 1947. https://doi.org/10.1007/BF02295996
30. J.D. Turiel and T. Aste. "Peer-to-peer loan acceptance and default prediction with artificial intelligence", Royal Society Open Science, 7(6), 191649, 2020. https://doi.org/10.1098/rsos.191649
31. T.T. Nguyen, T.H. Nguyen and V.T. Pham. "Comparative analysis of boosting algorithms for predicting personal default", Cogent Economics and Finance, 13(1), 2465971, 2025. https://doi.org/10.1080/23322039.2025.2465971
32. M. Ala'raj, M.F. Abbod and M. Radi. "A novel credit card default prediction model using behavioural and transactional LSTM approach", Applied Soft Computing, 117, 108356, 2022. https://doi.org/10.1016/j.asoc.2021.108356
33. S. Wang and X. Zhang**,** "Research on Credit Default Prediction Model Based on TabNet-Stacking", *Entropy*, 26(10), 861, 2024). https://doi.org/10.3390/e26100861
34. M.J. Ariza-Garzón, J. Arroyo, A. Caparrini and M.J. Segovia-Vargas. "Explainability of a machine learning scoring model in peer-to-peer lending", IEEE Access, 8, pp. 64873–64890, 2020. https://doi.org/10.1109/ACCESS.2020.2984412
35. R. Hlongwane, K. Ramabao and W. Mongwe. "A novel framework for enhancing transparency in credit scoring: Leveraging Shapley values for interpretable credit scorecards", PLOS ONE, 19(8), e0308718, 2024. https://doi.org/10.1371/journal.pone.0308718

Biographical Sketch

Deep Shikha is a PhD research scholar in Computer Science at Suresh Gyan Vihar University, Jaipur, India. Her research interests include applications of machine learning in credit risk modelling, including borrower segmentation, leakage-free validation and explainable AI for financial institutions. She has previously published a peer-reviewed paper on related topics.

Himanshu Vasnani is a faculty member in the Department of Mechanical Engineering, Suresh Gyan Vihar University, Jaipur, India. His research interests are in interdisciplinary applications of data driven methods and he guides research scholars on applied machine learning problems in engineering and management domains.

Amit Kumar Verma is a practitioner and researcher of Arrivo Edu Private Limited, Ahmedabad, Gujarat, India. He is a PhD scholar from Suresh Gyan Vihar University. His research interest is applied AI, including stock market prediction, credit risk modelling and explainable hybrid artificial intelligence (AI). He has skills in Fintech , ML.