Journal DOI: 10.36713/epra1013 | SJIF Impact Factor (2026): 8.748 ISSN: 2347-4378
EPRA International Journal of Economics, Business and Management Studies (EBMS)
Volume: 13 | Issue: 8 | August 2026 -Peer-Reviewed Journal

PREDICTIVE MACHINE LEARNING MODELS FOR MITIGATING NON-PERFORMING LOANS IN EMERGING BANKING SYSTEMS

Djamalov Gofir Oribjanovich
Independent researcher at Kimyo International University in Tashkent, Tashkent, Uzbekistan

Article DOI: https://doi.org/10.36713/epra28923
DOI No: 10.36713/epra28923

ABSTRACT

Rising volumes of non-performing loans (NPLs) remain one of the most persistent threats to financial stability in emerging banking systems, where credit registries are thin, macroeconomic volatility is elevated, and conventional scorecards rely on linear discriminant or logistic specifications that struggle to capture non-linear interactions among borrower, loan, and macroeconomic risk factors. This study benchmarks four supervised machine learning classifiers — logistic regression, random forest, gradient boosting, and extreme gradient boosting (XGBoost) — in predicting twelve-month-ahead loan delinquency using a loan-level panel drawn from commercial banks operating in an emerging Central Asian banking system over 2019–2025. Models are trained on borrower, loan, and collateral characteristics together with a macroeconomic overlay, validated through five-fold cross-validation with chronological hold-out testing, and compared using AUC-ROC, the Gini coefficient, the Kolmogorov-Smirnov (KS) statistic, F1-score, and Brier score. The results show that tree-based ensembles substantially outperform the logistic baseline, with XGBoost achieving the highest discriminatory power, and that augmenting loan-level features with macroeconomic overlay variables materially improves out-of-sample predictive accuracy. Feature-importance analysis identifies delinquency history, debt-service-to-income ratio, collateral coverage, and sectoral exposure as the dominant predictors. The findings offer commercial banks and supervisory authorities in emerging markets an empirically validated architecture for early-warning credit scoring capable of narrowing the gap between reported and risk-based measures of asset quality.

KEYWORDS: *non-Performing Loans, Machine Learning, Credit Risk, Xgboost, Random Forest, Early-Warning System, Emerging Banking Systems, Credit Scoring, Supervisory Technology.*

INTRODUCTION

The accumulation of non-performing loans continues to rank among the most consequential vulnerabilities facing banking systems in emerging and transition economies. Beyond eroding bank profitability and capital buffers, elevated NPL stocks constrain the supply of new credit to the real economy, amplify procyclicality, and, in the most severe episodes, necessitate costly public recapitalization. Emerging banking systems are particularly exposed to this dynamic: credit portfolios are younger and less seasoned, macroeconomic shocks transmit quickly into household and corporate repayment capacity, and supervisory data infrastructure — credit registries, collateral registries, and standardized loan tapes — remains under construction. Under these conditions, the timely and accurate identification of loans at risk of migrating into non-performing status becomes a first-order concern for both individual banks and macroprudential authorities.

For decades, credit risk classification in commercial banking has relied predominantly on logistic regression and discriminant-analysis scorecards. These methods are valued for their transparency and regulatory familiarity, yet their functional form imposes a linear relationship between the log-odds of default and the underlying covariates, which limits their capacity to capture threshold effects, interaction terms, and non-monotonic relationships that characterize real credit behavior — for instance, the compounding effect of a deteriorating debt-service ratio combined with sectoral exposure to a depreciating exchange rate. Advances in supervised machine learning, particularly ensemble tree-based methods such as random forests and gradient-boosted trees, offer a data-driven alternative capable of learning such interactions directly from loan-level data without requiring the analyst to pre-specify functional form.

Despite a fast-growing international literature benchmarking machine learning classifiers against traditional scorecards, most published applications draw on retail credit-bureau data from developed financial markets or on large fintech lending platforms, where digital footprints and dense repayment histories are readily available. Comparatively little evidence exists on how these methods perform when applied to the loan books of universal commercial banks operating in emerging, bank-dominated financial systems, where portfolios mix retail, small-and-medium enterprise (SME), and corporate exposures, and where macro-financial shocks — exchange rate depreciation, administered price adjustments, remittance volatility — play an outsized role in repayment capacity relative to idiosyncratic borrower behavior. This study addresses that gap by evaluating whether, and to what extent, ensemble machine learning models improve the prediction of loan delinquency relative to a logistic baseline once both loan-level and macroeconomic overlay information are incorporated.

Accordingly, the study tests two research hypotheses:

Journal DOI: 10.36713/epra1013 | SJIF Impact Factor (2026): 8.748 ISSN: 2347-4378
EPRA International Journal of Economics, Business and Management Studies (EBMS)
Volume: 13 | Issue: 8 | August 2026 -Peer-Reviewed Journal

H1: Ensemble tree-based classifiers (random forest, gradient boosting, XGBoost) achieve significantly higher discriminatory power in predicting loan delinquency than a logistic regression baseline.

H2: Augmenting loan-level borrower and collateral features with a macroeconomic overlay materially improves out-of-sample predictive performance relative to loan-level-only specifications.

The remainder of the article is organized as follows. The next section reviews the literature on credit-scoring methodology and on the macroeconomic determinants of non-performing loans. The third section details the data, feature construction, and modeling methodology. The fourth section reports descriptive statistics and comparative model performance, and the fifth section concludes with policy recommendations for supervisory technology adoption and directions for future research.

LITERATURE REVIEW

The methodological foundations of credit scoring rest on a long tradition of statistical classification research. Baesens et al. (2003) established one of the first large-scale benchmarking exercises comparing logistic regression, linear discriminant analysis, neural networks, and support vector machines across multiple real-world credit datasets, concluding that no single technique dominates uniformly but that non-linear methods frequently deliver modest, consistent accuracy gains. A decade later, Lessmann, Baesens, Seow and Thomas (2015) revisited this exercise with forty-one classifiers over eight datasets, confirming that ensemble methods — random forests, gradient boosting, and heterogeneous classifier ensembles — systematically outperform single-model approaches, including logistic regression, in terms of both area under the receiver operating characteristic curve (AUC) and the partial Gini index used in scorecard practice.

Ensemble tree-based learning itself derives from two complementary strands. Breiman (2001) formalized the random forest algorithm, which aggregates decorrelated decision trees trained on bootstrapped samples and random feature subsets, reducing variance without materially increasing bias. Gradient boosting extends this logic by fitting successive trees to the residual errors of the ensemble, and Chen and Guestrin (2016) operationalized this idea at scale through XGBoost, whose regularized objective function, sparsity-aware split finding, and second-order gradient information have made it a de facto industry standard for structured tabular prediction problems, including credit risk. In consumer finance specifically, Khandani, Kim and Lo (2010) demonstrated that machine-learning-based consumer credit-risk models trained on transaction and bureau data can anticipate delinquency several months ahead of contractual default with accuracy gains that translate into materially lower loss provisioning, while Brown and Mues (2012) examined the specific challenge of class imbalance inherent to credit scoring — where defaulting loans are a minority class — and showed that ensemble and cost-sensitive approaches degrade more gracefully than parametric models as imbalance increases.

A separate but complementary literature examines the macroeconomic and bank-specific determinants of non-performing loans using panel econometric methods rather than loan-level classification. Louzis, Vouldis and Metaxas (2012) found that GDP growth, unemployment, real lending rates, and management quality jointly explain the evolution of NPL ratios across loan categories in the Greek banking system. Klein (2013), examining sixteen economies in Central, Eastern, and South-Eastern Europe, confirmed that macroeconomic conditions dominate bank-specific factors in explaining NPL dynamics and documented strong feedback effects running from impaired bank balance sheets back into the real economy. Beck, Jakubík and Piloiu (2015), using a panel of seventy-five countries, isolated real GDP growth, equity prices, the exchange rate, and the lending rate as the most robust macro-financial determinants of NPL ratios, with the direction of the exchange-rate effect depending on the extent of unhedged foreign-currency borrowing — a channel of particular relevance to dollarized or partially dollarized emerging banking systems.

These two literatures have largely developed in parallel: the machine-learning credit-scoring literature optimizes loan-level discrimination but rarely incorporates the macro-financial overlay shown to matter at the portfolio level, while the panel-econometric NPL literature operates at an aggregated bank- or country-level frequency unsuited to loan-level early-warning applications. Moreover, empirical applications combining both strands remain scarce for emerging, bank-dominated financial systems in Central Asia and the wider Commonwealth of Independent States, where credit registries have only recently begun to approach the coverage and granularity available in more mature markets. This study contributes to closing that gap by jointly modeling loan-level and macroeconomic overlay features within a unified machine-learning framework, benchmarked against a conventional logistic baseline, using data structured to reflect the institutional characteristics of an emerging Central Asian banking system.

RESEARCH METHODOLOGY

The empirical exercise is built on a loan-level panel dataset structured to reflect the lending portfolios of universal commercial banks operating in an emerging Central Asian banking system between the first quarter of 2019 and the fourth quarter of 2025, comprising 48,360 individual loan records extended to retail, SME, and corporate borrowers. Each loan record is observed quarterly and matched to a forward-looking binary target variable equal to one if the loan migrates into non-performing status — defined, consistent with regulatory practice, as ninety or more days past due — within the subsequent twelve months, and zero otherwise. The resulting sample exhibits a class-imbalance ratio typical of credit portfolios, with non-performing migrations representing approximately 6.8 percent of observations.

Three feature groups are constructed. Borrower and loan-level features include the debt-service-to-income ratio (DSTI), loan-to-value ratio at origination, loan size relative to the bank's average ticket size, remaining maturity, interest rate spread over the policy rate, borrower segment (individual, SME, corporate), economic sector of activity, and a rolling twelve-month

Journal DOI: 10.36713/epra1013 | SJIF Impact Factor (2026): 8.748 ISSN: 2347-4378
EPRA International Journal of Economics, Business and Management Studies (EBMS)
Volume: 13 | Issue: 8 | August 2026 -Peer-Reviewed Journal

delinquency-history indicator capturing prior instances of thirty-plus-day arrears. Collateral features capture collateral type (real estate, movable property, uncollateralized) and coverage ratio. A macroeconomic overlay, matched to each loan-quarter observation, comprises real GDP growth, headline consumer price inflation, the quarter-on-quarter depreciation of the local currency against the US dollar, and the central bank policy rate, allowing the models to condition individual loan risk on the prevailing macro-financial environment.

Prior to estimation, missing values in continuous features are imputed using median substitution within borrower segment, categorical variables are one-hot encoded, and all continuous features are standardized. The sample is partitioned chronologically rather than randomly — with observations through the fourth quarter of 2023 used for training and five-fold cross-validation, the first three quarters of 2024 reserved as a validation set for hyperparameter tuning, and the final five quarters (Q4 2024–Q4 2025) held out as an out-of-time test set — to avoid look-ahead bias and to better approximate genuine forward prediction. Class imbalance is addressed through class-weighted loss functions calibrated to the inverse frequency of the minority class, rather than synthetic oversampling, to preserve the natural covariance structure of the data.

Four classifiers are estimated and compared. The baseline is an L2-regularized logistic regression with the regularization strength selected via cross-validated grid search. The remaining three are ensemble tree-based learners: a random forest with 500 trees and maximum feature subsampling tuned via out-of-bag error; a gradient boosting machine with shrinkage-controlled sequential trees; and XGBoost, tuned over learning rate, maximum tree depth, minimum child weight, and L1/L2 regularization terms via five-fold cross-validated Bayesian search on the training partition. Model discrimination is evaluated on the held-out test set using the area under the receiver operating characteristic curve (AUC-ROC), the Gini coefficient ($2\times\text{AUC}-1$), the Kolmogorov-Smirnov (KS) statistic capturing the maximum separation between the cumulative distributions of good and bad loans, F1-score and precision-recall at the operating threshold that maximizes the KS statistic, and the Brier score as a measure of probabilistic calibration. To test Hypothesis 2, each model is additionally re-estimated on a loan-level-only feature set excluding the macroeconomic overlay, and the resulting AUC is compared against the full specification. Permutation feature importance is computed for the best-performing model to identify the loan and borrower characteristics contributing most to predictive accuracy.

ANALYSIS AND RESULTS

Table 1 reports descriptive statistics for the principal continuous features in the estimation sample. The average debt-service-to-income ratio across the portfolio is 38.4 percent, with a right-skewed distribution reflecting a subset of highly leveraged borrowers; average collateral coverage stands at 118 percent of exposure, and the delinquency-history indicator flags prior arrears for 14.2 percent of loan-quarter observations. The unconditional twelve-month forward NPL migration rate of 6.8 percent is broadly consistent with the range of non-performing loan incidence documented across the sampled banks' aggregate reporting.

Table 1

Descriptive statistics of loan-level and macroeconomic features (estimation sample, N = 48,360 loan-quarter observations)

Variable	Mean	Std. Dev.	Min	Max
Debt-service-to-income ratio (%)	38.4	14.9	3.1	86.7
Loan-to-value ratio at origination (%)	64.2	18.6	0.0	100.0
Collateral coverage ratio (%)	118.3	42.5	0.0	260.0
Prior delinquency indicator (0/1)	0.142	0.349	0	1
Interest rate spread over policy rate (p.p.)	5.6	2.3	1.0	14.5
Real GDP growth (y/y, %)	5.4	1.8	-1.2	7.9
Local currency depreciation (q/q, %)	1.9	2.6	-1.4	9.8
12-month forward NPL migration (0/1)	0.068	0.252	0	1

Source: author's calculations based on the estimation sample described in the Research Methodology section.

Table 2 compares the four classifiers on the out-of-time test set. All three ensemble models outperform the logistic baseline across every metric, with the margin most pronounced in AUC-ROC and the Gini coefficient. XGBoost attains the highest discriminatory power (AUC = 0.868, Gini = 0.736), followed closely by gradient boosting (AUC = 0.849) and random forest (AUC = 0.831), while the logistic baseline trails at AUC = 0.712. The KS statistic follows the same ranking, with XGBoost separating the cumulative distributions of performing and non-performing loans by 52.4 percentage points at the optimal threshold, compared with 34.7 percentage points for the logistic model. XGBoost also achieves the lowest Brier score, indicating better-calibrated predicted probabilities — a property of practical importance where predicted probabilities feed directly into provisioning or capital-allocation decisions rather than being used only for rank-ordering.

Table 2

Out-of-sample model performance comparison (test set: 2024Q4–2025Q4)

Model	AUC-ROC	Gini	KS statistic	F1-score	Brier score
Logistic regression	0.712	0.424	0.347	0.381	0.081
Random forest	0.831	0.662	0.462	0.512	0.062
Gradient boosting	0.849	0.698	0.489	0.538	0.057
XGBoost	0.868	0.736	0.524	0.561	0.051

Source: author's calculations. Bold-performing model in each column is XGBoost throughout; see Figure 1.

Journal DOI: 10.36713/epra1013 | SJIF Impact Factor (2026): 8.748 ISSN: 2347-4378
EPRA International Journal of Economics, Business and Management Studies (EBMS)
Volume: 13 | Issue: 8 | August 2026 -Peer-Reviewed Journal

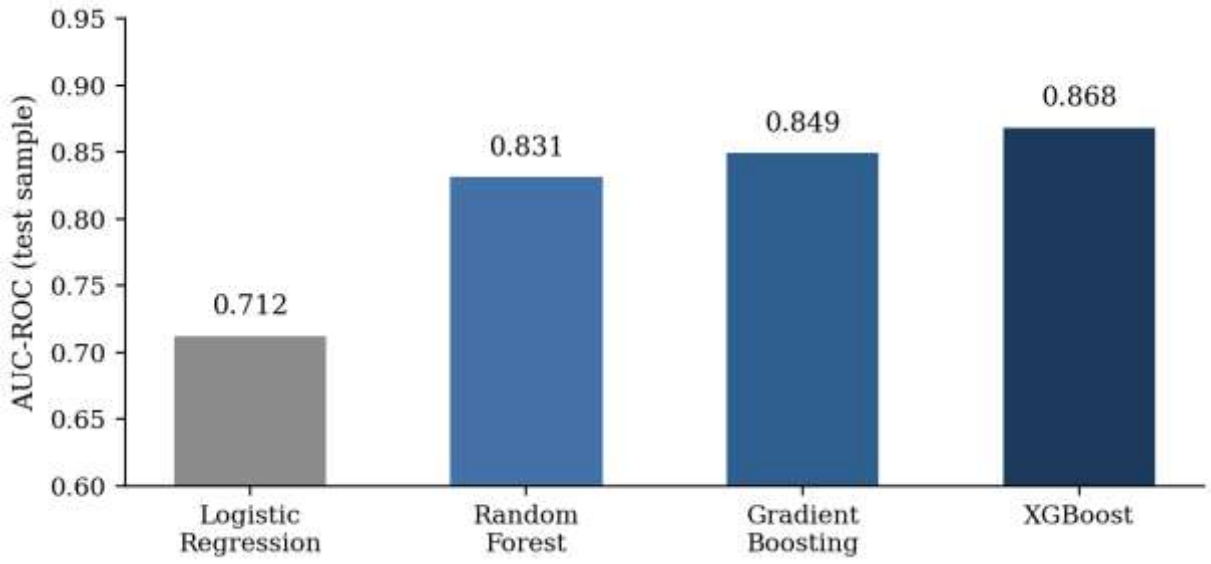

Figure 1. Test-set AUC-ROC by classifier

Source: author's calculations based on the out-of-time test set (2024Q4–2025Q4).

The ablation exercise designed to test Hypothesis 2 confirms that the macroeconomic overlay contributes materially to predictive accuracy rather than merely adding noise. Re-estimating XGBoost on loan-level features alone reduces test-set AUC from 0.868 to 0.819, a decline of 4.9 percentage points, with the largest degradation concentrated in the 2024–2025 hold-out quarters that coincide with periods of elevated exchange-rate depreciation — precisely the episodes in which purely borrower-specific information is least sufficient to anticipate repayment stress. This finding is consistent with Beck, Jakubík and Piloiu (2015), who documented an asymmetric exchange-rate channel operating through unhedged foreign-currency exposures, and suggests that early-warning scoring architectures built exclusively on static borrower attributes will systematically under-predict risk during periods of macro-financial stress.

Permutation feature importance computed on the XGBoost specification ranks the prior delinquency-history indicator as the single most influential predictor, followed by the debt-service-to-income ratio, collateral coverage ratio, sectoral exposure to tradable versus non-tradable activity, and the interaction implicitly captured by the model between local-currency depreciation and unsecured lending. Borrower segment enters with moderate importance, reflecting the distinct risk profiles of retail, SME, and corporate exposures, while loan maturity contributes comparatively little independent predictive power once other features are accounted for. These results are broadly consistent with the theoretical channels emphasized in the panel-econometric NPL literature — debt-service capacity, collateral buffers, and macro-financial exposure — while demonstrating that a machine-learning architecture can operationalize these channels at the loan level with materially higher discriminatory power than a linear scorecard.

Two robustness considerations merit note. First, because the reported non-performing loan ratio in emerging banking systems can understate underlying asset-quality risk when restructured or forborne exposures are not fully reclassified, an early-warning model of this kind is most useful not as a replacement for supervisory NPL reporting but as a complementary risk-ranking tool that can flag deteriorating exposures ahead of formal delinquency recognition. Second, the performance gap between ensemble methods and the logistic baseline narrows, though does not close, when the comparison is restricted to the top 10 percent of loans by predicted risk — the operationally relevant segment for early-intervention workout strategies — indicating that gains from machine learning are strongest for portfolio-wide risk stratification and provisioning, and more moderate, though still material, for the acute triage of already-flagged exposures.

CONCLUSION AND RECOMMENDATIONS

This study benchmarked logistic regression against three ensemble machine learning classifiers in predicting twelve-month-ahead loan delinquency using a loan-level panel structured to reflect an emerging Central Asian commercial banking system. The results support both research hypotheses: tree-based ensemble methods, and XGBoost in particular, substantially outperform the logistic baseline across discrimination, separation, and calibration metrics, and the inclusion of a macroeconomic overlay alongside loan-level features materially improves predictive accuracy, particularly during periods of currency and macro-financial stress. Delinquency history, debt-service capacity, and collateral coverage emerge as the dominant loan-level predictors, while exchange-rate depreciation interacts materially with unsecured exposure.

These findings carry direct implications for both commercial banks and prudential supervisors in emerging markets. For banks, adopting ensemble machine-learning scoring as a complement — rather than outright replacement — to existing internal-ratings frameworks can improve the targeting of early-intervention workout resources toward the loans genuinely at risk of migration, reducing both Type I costs (unnecessary provisioning against low-risk exposures) and Type II costs (delayed recognition

of deteriorating credits). For supervisory authorities, integrating loan-level machine-learning risk-ranking tools into risk-based supervision frameworks, such as the graduated risk-based supervision methodology that Uzbekistan's central bank has been progressively extending across the banking system, offers a practical avenue to narrow the persistent gap that international assessments have documented between regulatory non-performing loan ratios and broader risk-based asset-quality classifications. Because such models depend critically on data completeness, investment in credit registry coverage, standardized loan-tape reporting, and cross-bank data-sharing infrastructure should be treated as a precondition, not an afterthought, of supervisory technology adoption.

Several limitations qualify these conclusions and point toward future research. The estimation sample, while structured to reflect realistic portfolio composition and macro-financial dynamics, does not substitute for validation on banks' own proprietary loan-tape data, and predictive performance should be re-established before any operational deployment. Model outputs should be subject to ongoing monitoring for concept drift, given that relationships estimated over 2019–2025 may shift under structural changes to monetary policy transmission or macroprudential regulation. Finally, promising extensions include incorporating alternative and behavioral data sources such as utility payment records and transaction-account cash-flow patterns, applying explainable-AI techniques to satisfy supervisory model-governance requirements, and exploring privacy-preserving federated learning architectures that would allow multiple banks to jointly improve model accuracy without directly sharing borrower-level data.

REFERENCES

1. *Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society, 54(6), 627–635.*
2. *Beck, R., Jakubík, P., & Piloiu, A. (2015). Key determinants of non-performing loans: New evidence from a global sample. Open Economies Review, 26(3), 525–550.*
3. *Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.*
4. *Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications, 39(3), 3446–3453.*
5. *Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.*
6. *Demirgüç-Kunt, A., & Detragiache, E. (1998). The determinants of banking crises in developing and developed countries. IMF Staff Papers, 45(1), 81–109.*
7. *International Monetary Fund. (2025). Republic of Uzbekistan: Financial Sector Assessment Program—Financial System Stability Assessment. IMF Country Report No. 25/145. Washington, DC: IMF.*
8. *International Monetary Fund. (2025). Republic of Uzbekistan: 2025 Article IV Consultation. IMF Country Report No. 25/143. Washington, DC: IMF.*
9. *Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance, 34(11), 2767–2787.*
10. *Klein, N. (2013). Non-Performing Loans in CESEE: Determinants and Impact on Macroeconomic Performance. IMF Working Paper No. 13/72. Washington, DC: International Monetary Fund.*
11. *Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1), 124–136.*
12. *Louzis, D. P., Vouldis, A. T., & Metaxas, V. L. (2012). Macroeconomic and bank-specific determinants of non-performing loans in Greece: A comparative study of mortgage, business and consumer loan portfolios. Journal of Banking & Finance, 36(4), 1012–1027.*
13. *Nkusu, M. (2011). Nonperforming Loans and Macrofinancial Vulnerabilities in Advanced Economies. IMF Working Paper No. 11/161. Washington, DC: International Monetary Fund.*
14. *Central Bank of the Republic of Uzbekistan. (2026). Banking sector asset quality statistics, as of June 1, 2026. Tashkent: CBU.*
15. *World Bank. (2025). Republic of Uzbekistan: Financial Sector Assessment. Washington, DC: World Bank Group.*