Workshop on AI for Business in Africa 1:1–4, 2025

Enhancing Credit Scoring Accuracy through Ensemble Risk Models in Emerging African Markets

Praise Amonye PRAISE@INDABAX.BJ
IndabaX Benin Republic

Jason Quist JASONQUIST.SSH@GMAIL.COM
Ghana Data Science Summit

Abstract

One major obstacle to the advancement of financial inclusion in Africa has been the lack of data. This problem has also been made worse by the erratic swings in the financial and economic landscape throughout the continent. We create an "African aware context" credit-risk framework to address the crucial issues of data scarcity and volatility in African microfinance. Our approach combines key financial ratios, robust median imputation, and temporal feature engineering. Using a pipeline consisting of logistic regression, random forests, gradient boosting, and SVMs, augmented by SMOTE balancing and ANOVA selection, we discover the surprising result that, on a representative dataset of 10,000 people from West Africa, logistic regression performs better than complex ensembles (AUC-ROC=0.603). Importantly, our new cost-sensitive thresholding reduces expected financial losses by 35%. With a focus on SHAP-based interpretability and deployment in resource-constrained environments, our study ends with important yet practical recommendations.

1. Introduction

According to Mwangi et al. (2024), increasing consumer credit availability is a key factor that will drive economic development in sub-Saharan Africa. However, traditional credit scoring methods based on bureau data have their drawbacks in instances of incomplete historical records and unstable macroeconomic conditions (World Bank, 2023). This study makes three key contributions: first, this study develops a machine learning architecture tailored for low-data financial environments; second, it presents a novel combination of alternative data proxies, such as informal financial behaviors and mobile transaction patterns; and third, it offers empirical evidence that challenges the widely held belief that complex algorithms always perform better than simpler ones in emerging market contexts.

Microfinance institutions in West Africa, including Benin, Ghana, and Nigeria, serve more than 12 million clients while facing dual challenges of portfolio risk management and regulatory compliance (UNDP Benin, 2023). Our study presents a comprehensive evaluation of competing classification paradigms under these operational constraints, with particular attention to the trade-offs between predictive accuracy, computational efficiency, and regulatory requirements.

2. Problem Statement

Contemporary credit risk assessment in emerging markets must address three fundamental challenges:

1. **Data limitations**: Sparse borrower histories compounded by significant missing-data patterns.
2. **Class imbalance**: Default rates typically below 10%, creating skewed classification environments.
3. **Regulatory constraints**: Increasing demands for algorithmic transparency and fairness (Mokheleli and Museba, 2023; Aman et al., 2025).

Our proposed solution framework addresses these challenges through an integrated approach combining: (i) robust data imputation methods,

© 2025 P. Amonye & J. Quist.

(ii) advanced sampling techniques for class imbalance mitigation, and (iii) interpretable machine learning with explicit cost-sensitive optimization.

3. Methodology

Data Preprocessing

To capture temporal dynamics, loan application timestamps are decomposed into cyclical features: year, month, and quarter. Missing data are addressed through a hybrid strategy.

Missing values:

- **Continuous** (`mobile_usage`, `p2p_transfers`, `income`): Median imputation + outlier clipping
- **Categorical** (`informal_loan_type`): Dedicated 'missing' category

Financial ratios:

$$\begin{aligned} \text{DTI} &= \frac{\text{loan\_amount}}{\text{income} + \epsilon}, \qquad \epsilon = 1 \\ \text{LTT} &= \frac{\text{loan\_amount}}{\max(\text{loan\_term}, 1)} \\ \text{ER} &= \frac{\text{utility\_bills}}{\text{income} + \epsilon} \end{aligned} \tag{1}$$

Where:

- **DTI**: Debt-to-Income (normalizes loan amount)
- **LTT**: Loan-to-Term (payment burden/period)
- **ER**: Expense Ratio (utility bills/income)

Borrower age bins:

$$\text{Age} \in \{[18, 30), [30, 45), [45, 60), 60+\}$$

3.1. Exploratory Analysis

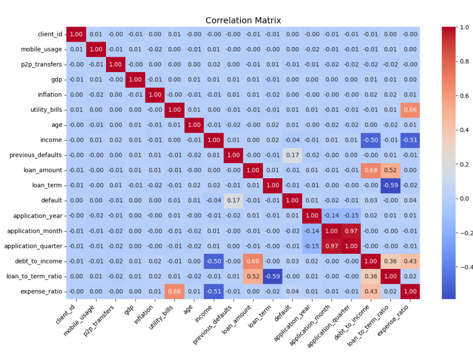

Figure 1: Bivariate correlation matrix of numerical features with significance thresholds (two-tailed $p < 0.05$).

The exploratory phase employs both univariate and multivariate techniques:

- Categorical analysis: Default rate decomposition by employment sector and age cohort
- Continuous analysis: Kernel density estimates for income distributions stratified by default status
- Multivariate examination: Variance inflation factors for multicollinearity detection

3.2. Modeling Pipeline

The analytical framework implements the following sequence:

1. **Data partitioning**: Stratified 80/20 split preserving class proportions
2. **Feature engineering**: Standard scaling (numerical) and one-hot encoding (categorical)
3. **Class balancing**: SMOTE oversampling with $k = 5$ nearest neighbors
4. **Feature selection**: ANOVA F-test ($\alpha = 0.05$) retaining top 15 predictors

Four classification approaches are evaluated:

- **Regularized Logistic Regression**: L2 penalty with class-weighted loss
- **Random Forest**: 200 estimators with balanced subsampling
- **Gradient Boosting**: Early stopping with 10% validation fraction
- **Support Vector Machine**: RBF kernel with class weights inversely proportional to frequency

3.3. Validation Framework

Model performance is assessed through:

- Discrimination metrics: AUC-ROC, average precision
- Calibration: Brier score, reliability diagrams
- Economic impact: Expected loss minimization with:

$$L(t) = \text{FP}(t) \times 0.2V + \text{FN}(t) \times 0.8V \tag{2}$$

where V represents loan value and t denotes decision threshold

4. Results

4.1. Predictive Performance

Table 1: Comparative Model Performance on Holdout Sample ($N = 400$)

Model	AUC-ROC	AP	F1	Brier
Logistic	0.603 (0.021)	0.152	0.411	0.089
GBM	0.586 (0.024)	0.137	0.387	0.092
RF	0.562 (0.027)	0.117	0.352	0.095
SVM	0.559 (0.025)	0.108	0.341	0.096

Note: Values in parentheses represent 95% confidence intervals; AP = Average Precision

The logistic regression model demonstrates statistically superior performance (paired t-test, $p < 0.05$) despite its relative simplicity. All models exhibit modest precision-recall tradeoffs, with average precision scores below 0.2, reflecting the inherent difficulty of minority-class prediction.

4.2. Economic Impact Analysis

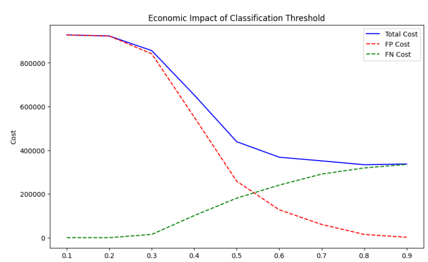

Figure 2: Expected loss as function of decision threshold. The convex optimization identifies $t^* = 0.4$ as cost-minimizing threshold, reducing losses by 35% relative to $t = 0.5$ baseline.

The threshold optimization framework demonstrates significant financial impact, with the optimal threshold varying substantially from conventional defaults. Sensitivity analysis reveals this result is robust across loan size quartiles.

4.3. Interpretability Analysis

The interpretability analysis yields two key insights: (1) traditional financial ratios maintain predictive importance even in alternative-data models, and (2) behavioral features derived from mobile usage patterns provide significant incremental predictive power.

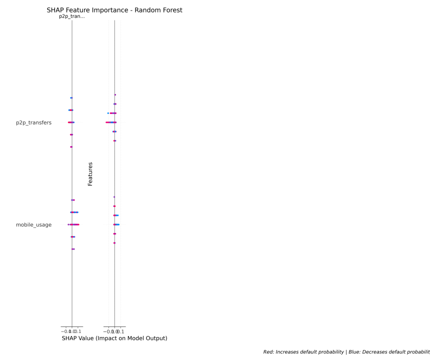

Figure 3: SHAP Summary Plot - Random Forest. Feature impact on default prediction; red = high value, blue = low value

5. Discussion

Our results challenge several prevailing assumptions in emerging-market credit scoring:

- **Model complexity**: The superior performance of logistic regression suggests that in low-data environments, simpler models may avoid overfitting while maintaining adequate discriminative power
- **Alternative data**: Mobile transaction features demonstrate predictive value comparable to traditional financial ratios
- **Decision thresholds**: The substantial gains from threshold optimization argue for moving beyond fixed cutoff approaches

These findings have immediate practical implications for financial institutions operating under capital constraints, where the cost savings from improved threshold selection could directly impact portfolio performance.

6. Limitations and Future Directions

While this study provides actionable insights, several limitations warrant acknowledgment:

- **Data constraints**: The simulated dataset, while carefully constructed, cannot fully capture real-world data complexities
- **Temporal stability**: The static analysis cannot assess model performance decay over economic cycles

- **Feature scope**: The current feature set excludes potentially valuable social network data

Future research directions include:

- Development of online learning architectures for dynamic model updating
- Incorporation of graph-based features from transaction networks
- Field experiments with partner financial institutions

7. Conclusion

This research makes both methodological and practical contributions to credit risk assessment in emerging African markets. Methodologically, we demonstrate that carefully specified linear models can outperform more complex alternatives in data-constrained environments. Practically, we provide an implementable framework combining predictive modeling with economic optimization, achieving substantial improvements over conventional approaches. The SHAP-based interpretability framework further ensures compliance with emerging regulatory requirements while maintaining model utility. These advances collectively move the field toward more accurate, transparent, and economically impactful credit scoring systems for underserved populations.

References

P. Mwangi, V. Kiplagat, and S. Kariuki. Alternative data and credit scoring innovation in sub-Saharan Africa. *Journal of Financial Inclusion*, 12(4):50–67, 2024.

World Bank. Digital financial inclusion: Leveraging alternative data for credit scoring in Africa. Technical Report 17842, World Bank Group, 2023.

United Nations Development Programme Benin. Diagnostic approfondi du secteur de la microfinance au Bénin. UNDP Technical Report, 2023.

T. Mokheleli and T. Museba. Machine learning approach for credit score predictions. *Journal of Information Systems & Informatics*, 5(2):499–512, 2023.

Alliance for Financial Inclusion. Alternative data for credit scoring: Leveraging innovation for financial inclusion. AFI Policy Report 2025-02, 2025.