ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25aug218

Loan Risk Assessment for Umurenge SACCO using Machine Learning

Mazimpaka Richard[1*]; Dr. Nizeyimana Pacifique[2]; Dr. Kundan Kumar[3]; Mukwende Placide[4]; Nshimiyimana Jerome[5]

[1,2]Master of Science in Big Data analytics, Faculty of IT, Adventist University of Central Africa

Corresponding Author: Mazimpaka Richard[1*]

Publication Date: 2025/08/19

Abstract: Umurenge SACCOs are instrumental in fostering financial inclusion in Rwanda, yet they face significant challenges with high loan default rates that threaten their long-term sustainability. This study develops a predictive model using machine learning techniques to assess loan default risk among SACCO borrowers. Using a real, anonymized dataset of 2,000 loan applications from the Rwanda Cooperative Agency (RCA), we compare six machine learning algorithms: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, AdaBoost, and XGBoost. The study addresses class imbalance through balanced training approaches and evaluates models using accuracy, precision, recall, and F1-score metrics. XGBoost achieved the highest performance with 89.5% accuracy, while Logistic Regression demonstrated optimal balance between performance (86.5% accuracy, 85.2% F1-score) and interpretability, making it suitable for real-world deployment in SACCO environments. Key predictors identified include credit score, past loan repayment behavior, and monthly income. These findings provide a scalable, data-driven approach for SACCOs to transition from intuition-based to evidence-based credit risk assessment, supporting Rwanda's digital transformation goals in financial services.

Keywords: *Credit Risk Assessment, Logistic Regression, SACCOs, Machine Learning, Loan Default Prediction, Financial Inclusion.*

How to Cite: Mazimpaka Richard; Dr. Nizeyimana Pacifique; Dr. Kundan Kumar; Mukwende Placide; Nshimiyimana Jerome (2025) Loan Risk Assessment for Umurenge SACCO using Machine Learning. *International Journal of Innovative Science and Research Technology*, 10(8), 602-609. https://doi.org/10.38124/ijisrt/25aug218

I. INTRODUCTION

Savings and Credit Cooperative Organizations (SACCOs) serve as critical financial institutions in developing economies, particularly in Rwanda where they operate under the community-based model that makes financial services accessible to previously excluded populations. Umurenge SACCOs, established under Rwanda's Vision 2020 development plan, have become the backbone of the government's financial inclusion efforts, with over 400 institutions serving more than three million Rwandans (National Bank of Rwanda, 2023). Despite their success in promoting financial inclusion, these institutions face persistent challenges in credit risk management, with high loan default rates threatening their operational sustainability (Hermes & Lensink, 2011; Gutiérrez-Nieto et al., 2007).

The surge in non-performing loans (NPLs) represents one of the most significant threats to SACCO viability. According to the Rwanda Cooperative Agency (RCA), more than 30 SACCOs reported NPL rates exceeding 20% in 2021, particularly in rural and peri-urban areas where risk assessment capacity remains limited (Rwanda Cooperative Agency, 2022). Traditional credit evaluation methods in SACCOs often rely on subjective judgment, community references, and informal assessment techniques that lack consistency and scalability (Mokhtar et al., 2012; Ledgerwood, 1999).

Machine learning and predictive analytics offer promising solutions for enhancing credit risk assessment in SACCOs (Lessmann et al., 2015; Baesens et al., 2003). The application of data-driven models can help these institutions transition from heuristic-based evaluations to systematic, objective risk assessment frameworks (Thomas, 2009; Schreiner, 2003). Internationally, cooperative financial institutions have successfully implemented machine learning tools for credit scoring, with credit unions in the United States and United Kingdom using automated scoring systems that combine credit bureau data with behavioral insights (Jagtiani & Lemieux, 2019). Similarly, microfinance organizations in China and India have adopted AI-based assessment tools that

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25aug218

leverage mobile data and digital footprints to serve previously excluded populations (Chen & Qian, 2019; Ghosh, 2021).

This study aims to develop and validate a logistic regression-based predictive model for assessing loan default risk among Umurenge SACCO members using historical loan and borrower data. The research employs real, anonymized data from the Rwanda Cooperative Agency and compares multiple machine learning algorithms to identify the most suitable approach for SACCO deployment.

II. METHODOLOGY

This study employs a quantitative approach grounded in supervised machine learning to predict loan default risk using structured data from SACCO loan applications. The methodology follows a systematic, multi-stage process including descriptive analysis, data preprocessing, feature engineering, model training, and evaluation. A comparative modeling strategy was adopted, testing six machine learning algorithms under both baseline and class-balanced scenarios to ensure robust evaluation across different risk categories.

The dataset comprises 2,000 anonymized loan application records aggregated by the Rwanda Cooperative Agency (RCA) from multiple Umurenge SACCOs across Rwanda, collected between 2020 and 2022. The dataset includes four major dimensions: demographic information (age, gender, marital status), financial standing (monthly income, credit score, existing debts), loan characteristics (amount, term, interest rate), and behavioral metrics (past repayment behavior, transaction frequency). The target variable, Risk_Level, classifies borrowers into three categories: Low risk (0), Medium risk (1), and High risk (2).

Initial data inspection revealed missing values in three variables: Monthly_Income, Credit_Score, and Savings_Account_Balance (100 missing entries each). These were imputed using median values to mitigate the impact of skewed distributions. Outlier treatment was performed using the Interquartile Range (IQR) method with additional winsorization at the 5th and 95th percentiles for highly skewed variables. See Figure 1.

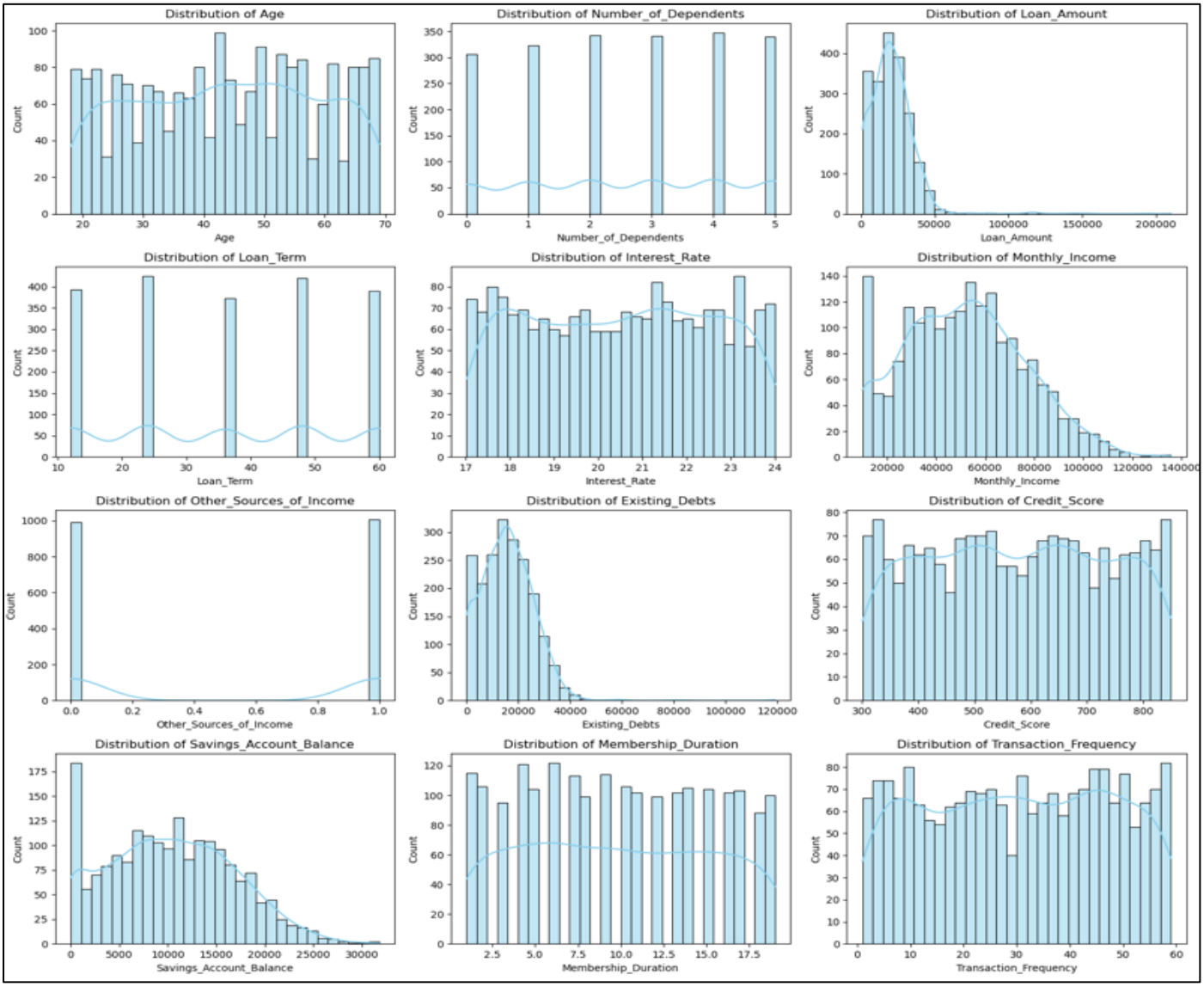

Fig 1 Distribution of Continuous Variables

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25aug218

Categorical variables were encoded using Label Encoder, while numerical features were standardized using Standard Scaler. Feature selection was conducted using Recursive Feature Elimination (RFE) and domain knowledge, resulting in 15 key features including Credit Score, Past_Loan_Repayment_Behavior, Monthly_Income, Savings_Account_Balance, and Loan Amount.

➢ *Six Classification Algorithms were Implemented:*

- *Logistic Regression:*

A linear classification algorithm that estimates the probability of categorical outcomes using the logistic function. For multiclass classification (Hosmer et al., 2013), multinomial logistic regression with softmax probabilities was employed:

$$P(y = k \mid X) = \frac{e^{\beta_k \cdot X}}{\sum_{j=1}^{K} e^{\beta_j \cdot X}}, \quad \text{where } k \in \{0, 1, 2\}$$

Where βk are the coefficients for class k and X is the input feature vector.

- *Decision Tree Classifier:*

A rule-based algorithm that partitions the feature space through decision rules (Quinlan, 1986), limited to maximum depth of 5 to prevent overfitting.

- *Random Forest Classifier:*

An ensemble method aggregating predictions from 100 decision trees trained on different random subsets of data and features (Breiman, 2001).

- *Support Vector Machine (SVM):*

Uses kernel functions to project features into higher-dimensional space for complex pattern recognition (Cortes & Vapnik, 1995), employing RBF kernel:

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right)$$

- *AdaBoost Classifier:*

Sequential ensemble technique that adaptively focuses on misclassified instances through weighted majority voting (Freund & Schapire, 1997).

- *XGBoost Classifier:*

Advanced gradient boosting implementation with L1 and L2 regularization, native missing value handling, and parallelized tree construction (Chen & Guestrin, 2016).

➢ *Model Performance*

Models were evaluated using accuracy, precision, recall, and F1-score metrics. The dataset was split using stratified random sampling (80% training, 20% testing) to maintain representative class proportions. Class imbalance was addressed through balanced training approaches, including class_weight='balanced' parameter for applicable algorithms. See Figure 2

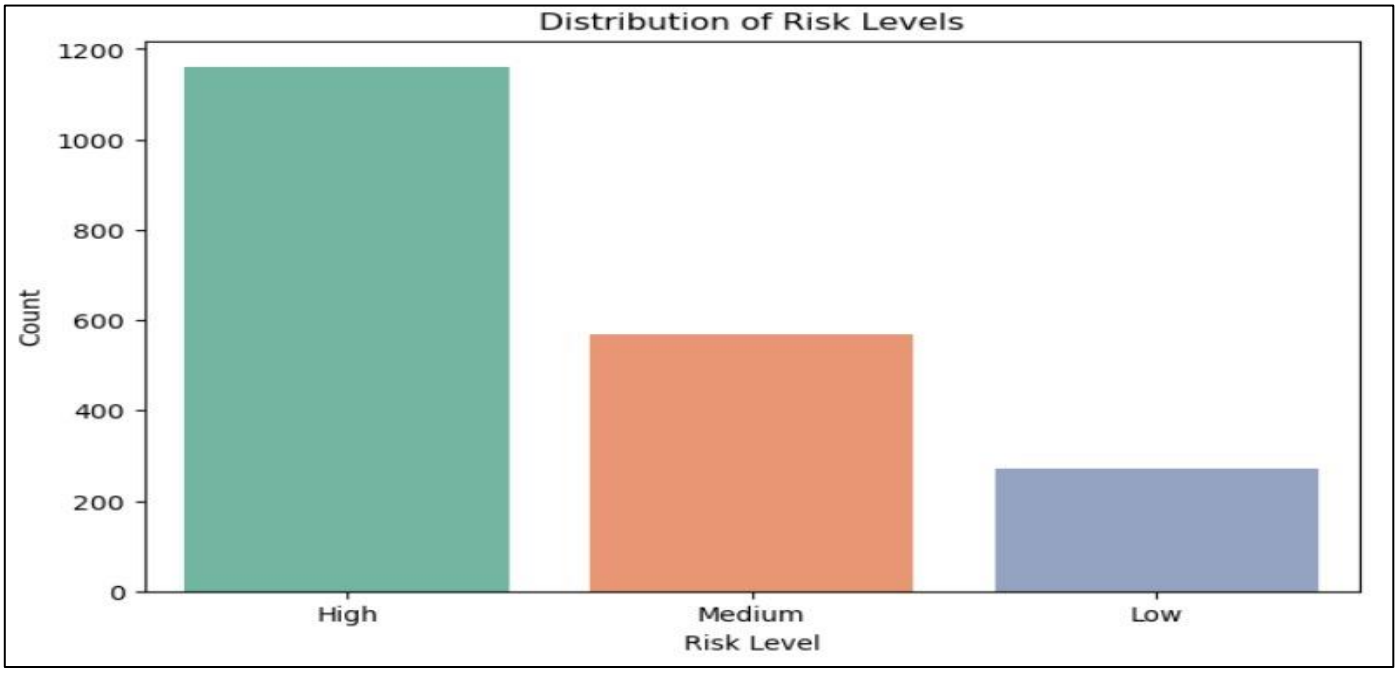

Fig 2 Class Distribution Before and After Balancing

III. DATA ANALYSIS AND RESULTS

➢ *Exploratory Data Analysis*

Initial analysis revealed significant class imbalance with High Risk (58%), Medium Risk (28.5%), and Low Risk (13.5%) borrowers. Financial variables including Loan_Amount, Existing_Debts, and Monthly_Income exhibited right-skewed distributions, while Age and Credit_Score showed near-normal distributions. See Figure 3

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25aug218

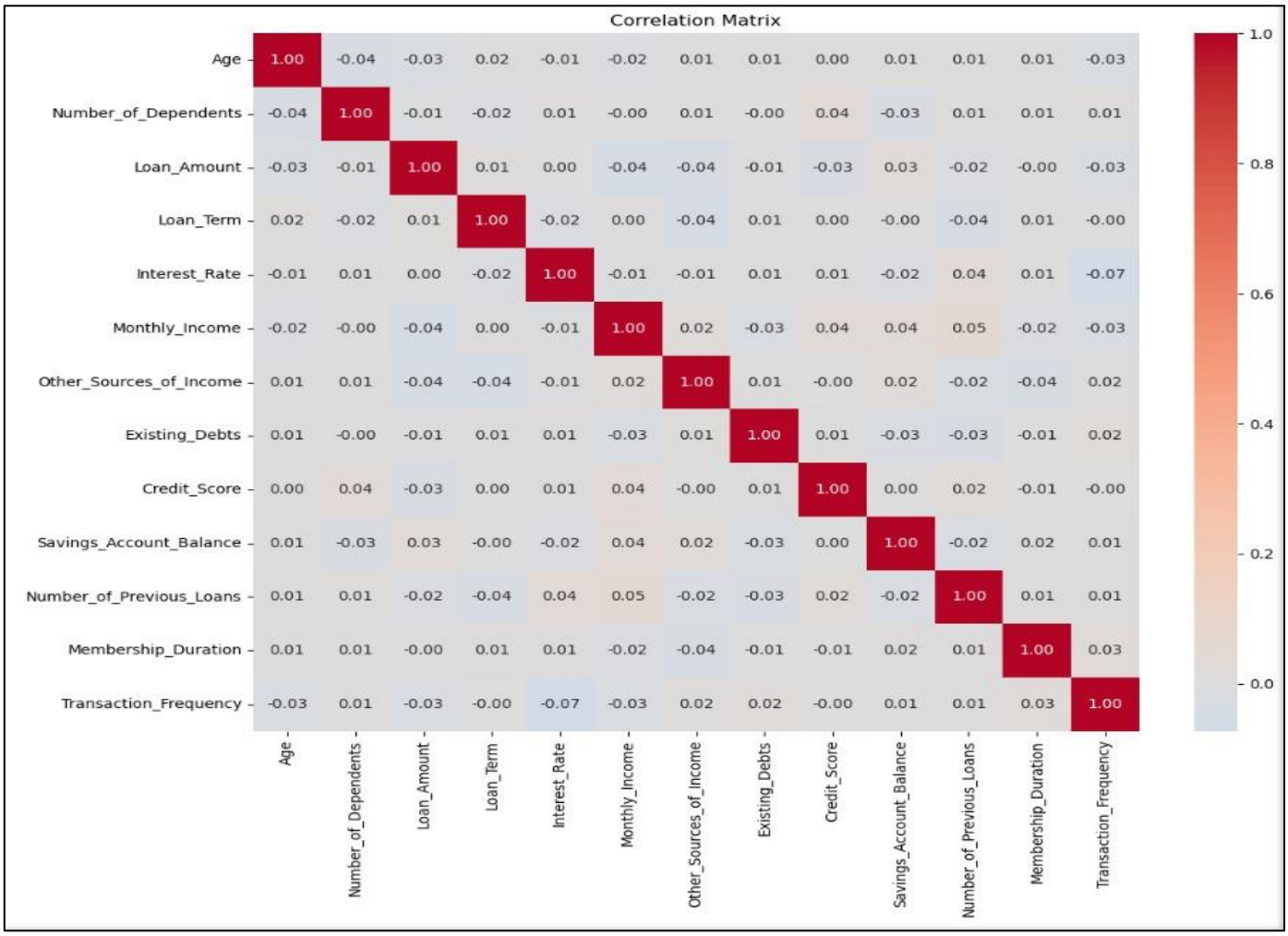

Fig 3 Correlation Matrix of Key Variables

The correlation analysis indicated strong relationships between credit score and risk level, as well as between past repayment behavior and default probability. Monthly income showed negative correlation with risk level, while loan-to-income ratio demonstrated positive correlation with default risk.

➢ *Feature Importance Analysis*

Feature importance analysis using XGBoost identified the most significant predictors of loan default: See Figure 4.

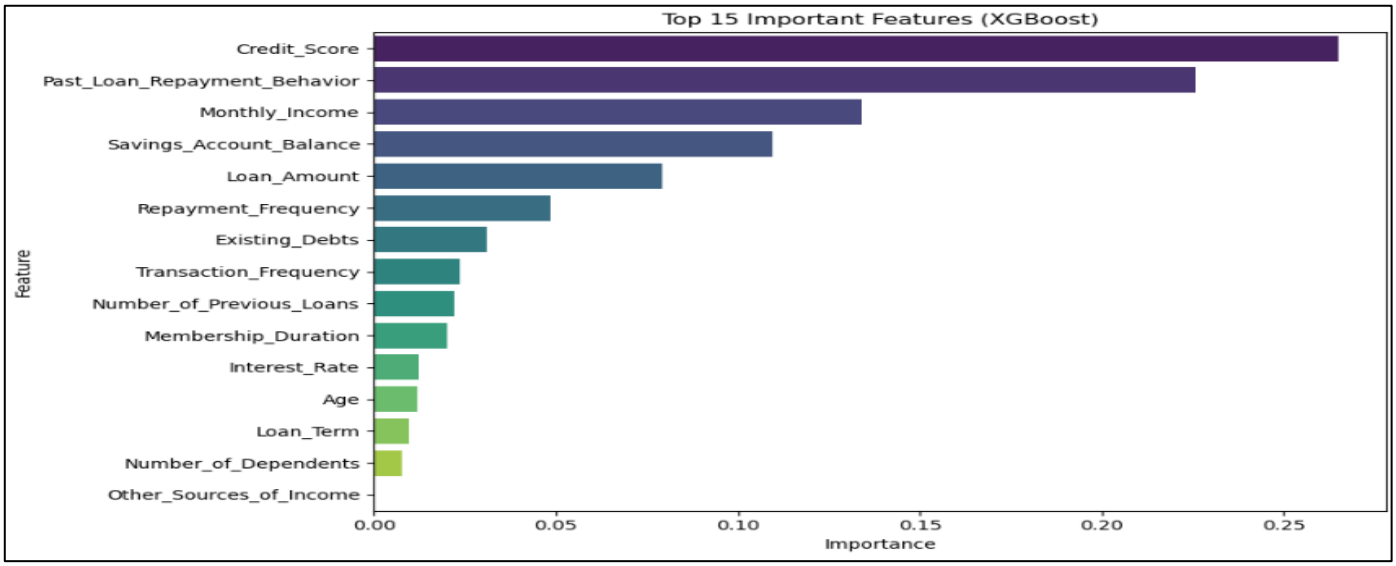

Fig 4 Feature Importance Rankings

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25aug218

- *The Top Five Predictors were:*

✓ Credit_Score
✓ Past_Loan_Repayment_Behavior
✓ Monthly_Income
✓ Savings_Account_Balance
✓ Loan_Amount

➢ *Model Performance Comparison*

Table 1 Model Performance Comparison

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.865	0.8605	0.8406	0.8485
Decision Tree	0.8275	0.7994	0.8211	0.8082
Random Forest	0.7775	0.8169	0.6332	0.6673
SVM	0.8625	0.8612	0.8107	0.8292
AdaBoost	0.7150	0.5121	0.5468	0.5181
XGBoost	0.8950	0.8759	0.8770	0.8761

Table 2 After Class Balancing

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.865	0.8361	0.8737	0.8526
Decision Tree	0.780	0.7357	0.7843	0.7509
Random Forest	0.8525	0.8012	0.8442	0.8151
SVM	0.835	0.7935	0.8393	0.8115
AdaBoost	0.7150	0.5121	0.5468	0.5181
XGBoost	0.8950	0.8759	0.8770	0.8761

➢ *Confusion Matrix Analysis*

Confusion matrices were generated for the top-performing models to assess classification accuracy across risk categories. See Figure 5,6 and 7.

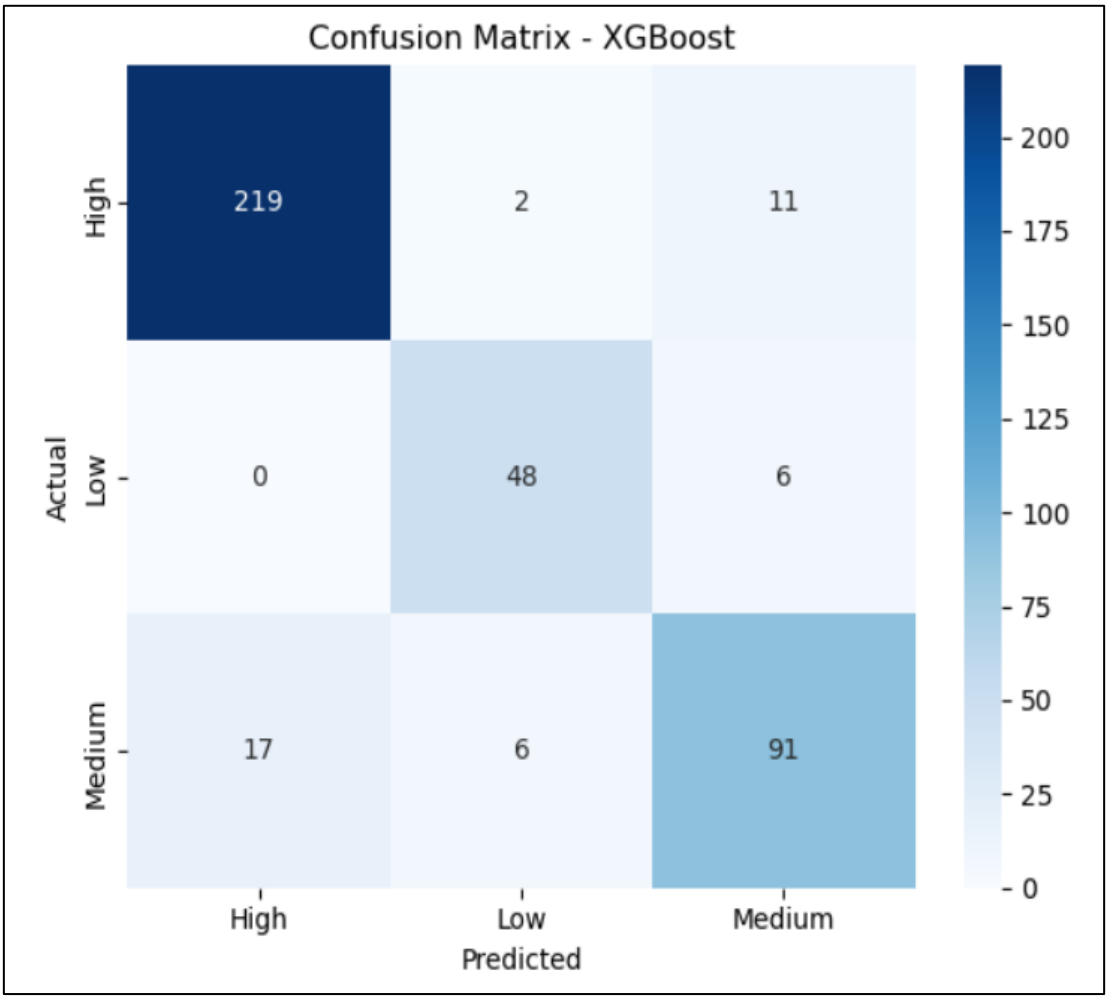

Fig 5 Confusion Matrix for XGBoost Model

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25aug218

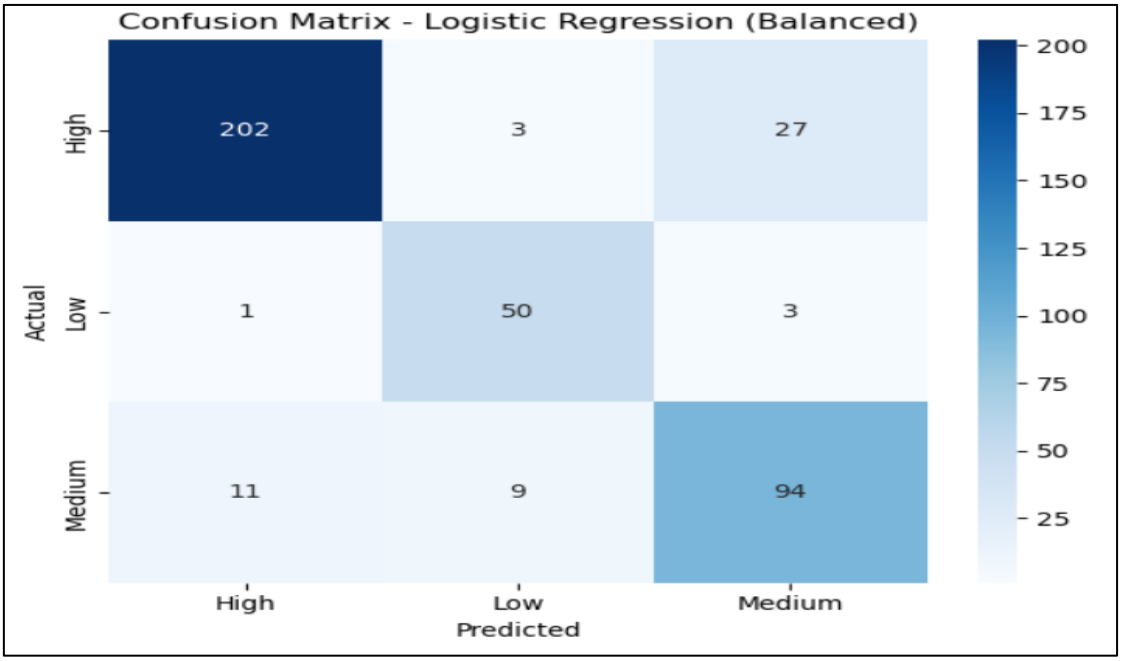

Fig 6 Confusion Matrix for Logistic Regression (Balanced)

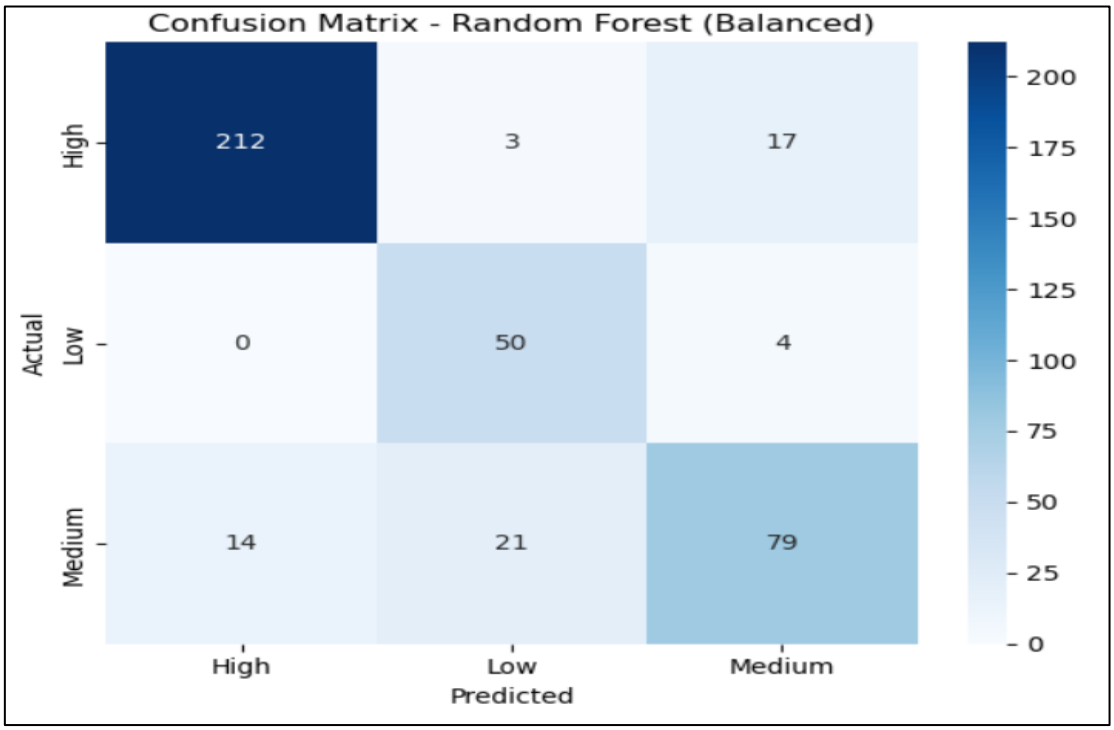

Fig 7 Confusion Matrix for Random Forest (Balanced)

XGBoost demonstrated superior performance in correctly identifying high-risk borrowers, while the balanced Logistic Regression model showed more equitable performance across all risk categories.

IV. CONCLUSION AND RECOMMENDATIONS

The empirical findings demonstrate the feasibility and effectiveness of implementing machine learning models for credit risk assessment in Umurenge SACCOs. While XGBoost achieved the highest predictive accuracy (89.5% accuracy, 87.6% F1-score), Logistic Regression emerged as the optimal choice for deployment due to its balance of performance (86.5% accuracy, 85.2% F1-score), interpretability, and operational simplicity.

The study identified critical predictors of loan default, including credit score, past repayment behavior, and monthly income, providing actionable insights for SACCO risk management. The successful application of class balancing techniques significantly improved model performance for

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25aug218

minority risk categories, ensuring more equitable risk assessment.

➢ *Key Recommendations Include:*

- *Model Integration:*

SACCOs should integrate the logistic regression model into loan evaluation systems as a decision support tool while maintaining human oversight for final lending decisions.

- *Staff Training:*

Comprehensive training programs should equip SACCO personnel with data literacy skills and model interpretation capabilities to build trust and enhance operational effectiveness.

- *Continuous Data Collection:*

Investment in digital record-keeping systems will enable continuous model retraining and adaptation to changing borrower behaviors and economic conditions.

- *Ethical Implementation:*

Transparent communication regarding predictive model use must ensure borrower trust and regulatory compliance while avoiding discriminatory practices (Abdou & Pointon, 2011; Verbraken et al., 2014).

- *Future Research Directions:*

Advanced ensemble methods, incorporation of alternative data sources (mobile money transactions, psychometric assessments), and real-time scoring systems represent promising areas for further development.

This research contributes to the growing body of literature on machine learning applications in cooperative banking and provides a practical framework for enhancing financial inclusion while maintaining institutional sustainability in Rwanda's SACCO sector.

ACKNOWLEDGMENT

The author extends sincere gratitude to **Dr. Kumar Kundan**, academic supervisor, and **Dr. Pacifique NIZEYIMANA**, Director of AUCA Gishushu Campus, for their valuable guidance and continuous encouragement throughout this research.

Special appreciation is extended to the **Ministry of Finance and Economic Planning (MINECOFIN)**, particularly to **Mr. Placide Mukwende**, whose insightful explanations and practical knowledge greatly enhanced the author's understanding of **loan risk assessment and its application within data science**. His input was instrumental in shaping the analytical approach of this study.

The author also acknowledges the **Adventist University of Central Africa (AUCA)** and expresses heartfelt thanks to fellow **MSc students** for their collaboration, support, and inspiration during the course of this work.

REFERENCES

[1]. Abate, G. T., Borzaga, C., & Getnet, K. (2016). Cost-efficiency and outreach of microfinance institutions: Trade-offs and the role of ownership. *Journal of Business Ethics*, 148(3), 647-665.

[2]. Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: A review of the literature. *Intelligent Systems in Accounting, Finance and Management*, 18(2-3), 59-88.

[3]. Akerlof, G. A. (1970). The market for 'lemons': Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, 84(3), 488-500.

[4]. 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.

[5]. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

[6]. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).

[7]. Chen, X., & Qian, W. (2019). AI and big data in China's fintech development. *China Economic Journal*, 12(3), 230-248.

[8]. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297.

[9]. Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1), 119-139.

[10]. Ghosh, S. (2021). FinTech and financial inclusion in India: An overview. *International Journal of Financial Studies*, 9(3), 43.

[11]. Gutiérrez-Nieto, B., Serrano-Cinca, C., & Molinero, C. M. (2007). Microfinance institutions and efficiency. *Omega*, 35(2), 131-142.

[12]. Hermes, N., & Lensink, R. (2011). Microfinance: Its impact, outreach, and sustainability. *World Development*, 39(6), 875-881.

[13]. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Hoboken, NJ: Wiley.

[14]. Jagtiani, J., & Lemieux, C. (2019). The roles of alternative data and machine learning in fintech lending: Evidence from the LendingClub consumer platform. *Financial Management*, 48(4), 1009-1029.

[15]. Kasozi, J., Musisi, C., & Muwanga, J. (2020). Predictive analytics in agricultural SACCO lending: A Ugandan case. *African Journal of Agricultural Economics*, 15(2), 45-62.

[16]. Ledgerwood, J. (1999). *Microfinance Handbook: An Institutional and Financial Perspective*. Washington, DC: The World Bank.

[17]. 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.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25aug218

[18]. Mokhtar, S. H., Nartea, G. V., & Gan, C. (2012). Determinants of microcredit loans repayment problem among microfinance borrowers in Malaysia. *International Journal of Business and Social Research*, 2(7), 152-163.
[19]. National Bank of Rwanda. (2023). *Annual Report 2022*. Kigali: BNR Publications.
[20]. Odhiambo, R. (2019). Application of credit scoring models in SACCOs: A case study of rural Kenya. *East African Journal of Financial Studies*, 12(3), 78-95.
[21]. Otieno, P. A., Ogutu, M., & Awino, Z. B. (2020). Adoption of machine learning in microfinance lending decisions in Kenya. *African Journal of Business Management*, 14(5), 123-134.
[22]. Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81-106.
[23]. Rwanda Cooperative Agency. (2022). *SACCO Performance Report 2021*. Kigali: RCA Publications.
[24]. Schreiner, M. (2003). Scoring: The next breakthrough in microcredit? *Consultative Group to Assist the Poor Occasional Paper*, 7, 1-34.
[25]. Thomas, L. C. (2009). *Consumer Credit Models: Pricing, Profit and Portfolios*. Oxford: Oxford University Press.
[26]. Verbraken, T., Bravo, C., Weber, R., & Baesens, B. (2014). Development and application of consumer credit scoring models using profit-based classification measures. *European Journal of Operational Research*, 238(2), 505-513.