Ratio Mathematatica Volume 56, 2026

Exploring Markov chain and Random Forest to predict loan default in rural banks

Joseph Kwabina Arhinful Johnson*
Ellis Agyeman†
Samuel Kwame Okai‡

Abstract

This study investigates loan default risks in rural banks in Ghana's by developing a hybrid predictive model that combines Markov chains and Random Forest machine learning. The Markov chain analysis examines the transition of loans between risk categories (Current, OLEM, Sub-standard, Doubtful, and Loss) based on Bank of Ghana (BoG) standards, revealing the dynamic nature of credit risk over time. The Random Forest model then classifies loans as default or non-default, identifying key features, with Markov chain transition probabilities serving as a crucial input. Using a dataset of 25,981 loan records from a rural bank, the study finds that a significant portion of borrowers operate in non-traditional sectors, with 14% of loans classified as Losses. The Markov analysis indicates stability in the Current and Loss states, while other risk categories exhibit notable transitions. The Random Forest classifier, utilizing features like Days in Arrears and Interest Charge Type, achieves 93% accuracy and strong F1 scores (0.945 for non-defaults and 0.865 for defaults). The model's Receiver Operating Characteristic and Area Under the Curve (ROC- AUC) score of 0.90 demonstrates its excellent ability to distinguish between defaulting and non-defaulting loans. The model identifies loan duration and arrears as primary predictors of default. Scenario simulations confirm that default risk increases as the loan ages and arrears accumulate. Overall, the hybrid model offers superior predictive accuracy and granular risk insights, enhancing credit risk

*Takoradi Technical University, Takoradi, Ghana; joseph.arhinful@ttu.edu.gh.
† Takoradi Technical University, Takoradi, Ghana; pk.akroma@gmail.com.
‡Takoradi Technical University, Takoradi, Ghana; oksaisamuel415@gmail.com.

management for rural banks and supporting financial inclusion in underserved communities.
Keywords: Markov chain, Random Forest, Loan Default§

1. Introduction

Individuals worldwide rely on bank credit to address short-term liquidity gaps and pursue long-term goals (Vermoesen et al., 2013). In a competitive and volatile financial environment, borrowing has become essential for households and firms, while lending remains a primary driver of bank profitability (Ekpu & Paloni, 2016). Interest income from loans constitutes a significant share of banks' revenues, but it also exposes them to credit risk—the possibility that borrowers will fail to meet contractual repayment obligations (Brown & Moles, 2014).

To mitigate this risk, lenders must rigorously assess applicants before disbursing funds. Traditionally, loan officers have used the "5 Cs" of credit—character, capital, capacity, collateral, and conditions—to gauge borrower quality (Oseni, 2023). However, this judgment-driven approach is limited by human bias and inconsistency, particularly with increasing loan (Abraham, 2023).

Many banks have employed specialized analysts to review each case and assign a numerical credit score that summarizes default risk (Golin & Delhaise, 2013). These scores, derived from variables like past repayment behavior and income stability, serve as probabilistic estimates of a borrower's ability and willingness to repay on time (Skiba & Tobacman, 2008).

Rural banks play a crucial role in promoting financial inclusion, particularly among low-income populations in rural communities. However, they face high loan default rates due to limitations in traditional credit assessment models and the lack of reliable predictive tools (Agyapong, 2023). Previous studies have primarily utilized traditional financial models, which often fail to capture the dynamic and temporal nature of loan defaults.

This study aims to develop and evaluate a hybrid model combining Markov chain and Random Forest techniques to predict loan defaults particularly for rural banks. By considering transitional probabilities and identifying key influencing factors, this research will fill existing gaps and offer a novel approach to credit risk management.

The hybrid framework is used to integrate the Random Forest and Markov chain models, with the computed transition probabilities serving as a crucial input for the

[1] Received on January 01, 2025. Accepted on May 31, 2025. Published on June 30, 2026.
DOI: 10.23755/rm.v56i0.1744. ISSN 1592-7415; e-ISSN 2282-8214.

machine learning classifier. The first step in this method is to model the stochastic movement of loans between five regulatory risk states: Current, OLEM, Sub-Standard, Doubtful, and Loss using a Discrete-Time Markov Chain (DTMC).

In this hybrid architecture, the Random Forest algorithm uses the transition probabilities obtained from the Markov analysis as input features in order to compute a comprehensive default risk score (RS) in conjunction with other borrower-specific covariates such as Days in Arrears and Interest Charge Type. The transition probabilities assist the model account for the dynamic and temporal aspect of credit risk, which standard static models frequently fail to capture. This score is the weighted sum of these features.

The work advances from merely forecasting the next likely state to offering a binary classification of whether or not a loan will eventually fail by including these transition-based insights into the predictive model. This combination improves the accuracy of identifying high-risk accounts before they enter the "Loss" state by enabling the bank to use borrower-level risk scores in conjunction with long-term transition dynamics for daily decision-making.

The hybrid model will address the limitations of traditional credit assessment models and provide a more accurate prediction of loan defaults. By leveraging the strengths of both Markov chains and Random Forest algorithms, this study will contribute to the development of more effective credit risk management strategies for rural banks.

2. Objectives of the study

The primary objective of this study is to develop a Markov Chain model for predicting loan defaults at rural banks. Consequently, the study seeks to achieve the following specific objectives:

1. Develop a Markov chain clustering model to predict loan defaults effectively.
2. Identify key features influencing the loan default prediction using Random Forest.
3. Improve prediction using optimized hyperparameter tuning.

3. Materials and methods

This study employed a quantitative research approach with a descriptive and exploratory design to analyze historical loan repayment data and predict default outcomes using the Markov Chain, K-Means Clustering, and Random Forest model. The design enabled the identification of trends and patterns in clients' loan repayment behavior and facilitated the application of the Markov Chain to model state transitions.

The study was conducted at a rural bank in Ghana's Western Region, serving low-income populations. The bank offers various financial services, including microloans and SME financing. The population comprised 25,981 loan clients with complete repayment histories. The study used a census approach, analyzing data from all clients.
Data processing involved coding, editing, and cleaning to ensure accuracy. Loan repayment behavior was categorized into five states: Current, OLEM, Sub-Standard, Doubtful, and Loss/Default. The Markov chain model estimated transition probabilities, and steady-state probabilities assessed long-term behavior patterns.
Analytical tools included Python and Microsoft Excel. The study addressed research objectives through probabilistic modeling, with justifications based on theoretical and empirical literature. This approach provided a structured framework for predictive modeling and enhanced reproducibility.

3.1 Discrete-Time Markov Chain (DTMC)

A stochastic process $\{X_t\} t \geq 0,$ where X_t represents the state of the process at time t, is called a DTMC if it satisfies the following conditions:

1. State Space (S): The process takes values from a finite or countable set of loan status. These states are defined as

$$S = \{S_1, S_2, S_3, S_4, S_5\} \tag{1}$$

where the five states are {current, Olem, Sub-standard, Doubtful, and loss} respectively.

2. The Markov property

$$P(S_{n+1=j} \mid S_{n=i}, S_{n-1} = k, \dots S_{0=X0}) = P(S_{n+1=j} \mid S_{n=i}) \tag{2}$$

for all states i and j being members in Equation (1) $(i, j \in S)$ at all times $(t \geq 0)$.

3.1.1 Transition Probabilities (P)

The matrix (P_{ij}) contains the probabilities of transitioning from state i to j within a one step as indicated in Equation (3).

$$P = [P_{ij}] \tag{3}$$

From Equation (3):

$$\hat{P}_{ij} = \frac{n_{ij}}{\sum_{k \in S} n_{ik}} \tag{4}$$

Where n_{ij} is the observed count or number of transitions from state i to j over the study period, and $\sum_{k \in S} n_{ik}$ is the total number of transitions which originates from state i.

However, P_{ij} satisfies the condition of probability distribution:

$$P_{ij} \geq 0 \tag{5}$$

$$\sum_{j \in S} P_{ij} = 1 \tag{6}$$

for all $i, j \in S$.

3.1.2 Random Forest

Random Forest is an ensemble learning method. It builds many decision trees and combines their predictions to improve accuracy and avoid over fitting.

3.1.2.1 Prediction equations (Ensemble Prediction)

Let x be any input vector containing loan features like days in arrest, total age of loan, etc. The ensemble of B decision trees $\{h_b(x)\}_{b=1}^{B}$ produces a final classification $\hat{y}$ through majority vote given as:

$$\hat{y} = \operatorname{mod}\{h_1(x), h_2(x), \ldots, h_B(x)\} \tag{7}$$

where $h_b(x)$ is the prediction from the b^{th} decision tree.

For Regression, the prediction is the average of all tree predictions given as

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} h_b(S) \tag{8}$$

3.1.2.2 Prediction Equations (Risk Score - RS)

At this point, a customer's loan risk score (RS) is calculated as a weighted linear combination of its features given as:

$$RS = \sum_{i=1}^{n} \upsilon_i \tau_i \tag{9}$$

where, n is the total number of the influencing features. Example is $n = 10$ for the top ten identified features. Also, υ_i is the importance weight of feature i. An example is the weight of 0.3003 for the number of days in arrears, and τ_i is the normalized value of feature i for a specific borrower.

3.1.2.3 Prediction Equations (Probability of Default - PD)

We further mapped the RS to a probability scale using a logistic (sigmoid) function given as:

$$P(Default) = \frac{1}{1 + e^{-RS}} \tag{10}$$

where, e is the Euler's constant approximately 2.718, and RS is the calculated RS from the Random Forest model in Equation (9).

3.1.2.4 Prediction Equations (Time-Based Risk Evolution)

The time-based risk evolution representing a power-law growth model is presented as:

$$Risk(t) = B_r \times (1+D_a)^{\alpha}(1+\beta L_a)^{\beta} \tag{11}$$

where, $Risk(t)$ is the initial base risk level at time t. B_r is the initial base risk level, D_a is the days in arrear, L_a is the total age of the loan in months, and α, β are the scaling exponents (scaling factors) that represents the sensitivity of risk arrears and loan respectively.

3.2 Model evaluation metrics

This section evaluates the performance and accuracy of the models. This helps to measure the ability of the predictions, understand the probability distribution, classify data correctly, and improve the model's reliability and robustness.

These are achieved with the F1 score, the Receiver Operating Characteristic, and the Area Under the Curve (ROC-AUC). The F1 score is given in equation 3.20 as

$$F1 = \frac{2pr}{p+r} \tag{12}$$

with,

$$p = \frac{TP}{TP+FP} \tag{13}$$

and,

$$r = \frac{TP}{TP+FN} \tag{14}$$

Where the accuracy is given by

$$\frac{TP+TN}{TP+TN+FP+FN} \tag{15}$$

The ROC-AUC score is given as the integral of the TP Rate (TPR) with respect to the FP Rate (FPR) in Equation (16) as:

$$\int_0^1 TPR\left(FPR^{-1}(x)\right)dx \tag{16}$$

From the above, p is the precision, r is the recall, TP is true positives (correctly predicted defaults), TN is the true negatives (correctly predicted non-defaults), FP is the false positives (non-defaults wrongly predicted defaults), and FN is the false negatives (defaults wrongly predicted as non-defaults). Also, x represents the varying decision thresholds of the classifiers.

It is worth noting that there is a notable class imbalance in the dataset, with non-defaults making up 85% of the total sample and defaults making up only 15%. Since the high headline score (91%) mostly reflects the model's ability in managing the dominant non-default class rather than its marginal success in detecting the minority class, this distribution may make overall accuracy a little distorted indicator. Additionally, a larger False Positive (FP) Rate and the intrinsic challenge of correctly classifying less frequent positive examples are shown by the lower F1 score for defaults (0.82) compared to non-defaults (0.93).

3.3 Ethical considerations

This study strictly adhered to ethical principles. Permission to access loan records was obtained from the rural bank's management, and informed consent was secured. All data was anonymized to protect client identities, and confidentiality was maintained through encrypted storage and restricted access. The study received ethical approval from the Takoradi Technical University Ethics Review Board, ensuring compliance with institutional and national standards. Participants' rights were respected throughout the research process.

4. Results and Discussion

This section reports on the sources of data and the discussions of the descriptive nature of the data for this study

4.1 Data description

The dataset consists of 25,981 loan accounts from a rural bank, featuring borrower profiles, repayment behavior, regulatory classification, loan characteristics, and default status. The dataset includes both categorical and numerical variables.

From Table 1 across the 25,981 loan records, the mean approved (and disbursed) facility size was approximately GHS 6,528 (SD = 15,199), although values ranged widely from 200 to about GHS 795,325. Contractual tenor averaged 11.44 months (SD = 16.11) with a maximum of 289 months, indicating a heavily right-skewed distribution. The typical loan was young: the median loan age was 4 months, well below the median contractual duration of 6 months. Days-in-arrears was highly skewed (M = 337, SD = 840) but the median was 0, reflecting the predominance of current accounts.

Table 1. Descriptive statistics for continuous variables

Variable	N	M	SD	Min	Max
Loan Amount Approved	25981.0	6528.5	15199.75	100.00	795324.81
Loan Amount Disbursed	25981.0	6528.5	15199.75	100.00	795324.81
Number of Payments Agreed	25981.0	11.44	16.11	0.00	289.0
Loan Principal Balance (Without Interest)	25981.0	4512.0	21526.32	-3000.0	3010616.51
Days in Arrears	25981.0	337.06	840.26	0.0	3248.0
BoG Provision	25981.0	253.29	1194.11	0.0	52559.64
Interest Received	25981.0	11.89	251.9	0.0	18915.11
loan_status_code	25981.0	0.67	1.43	0.0	4.0
loan_duration_months	25981.0	21.71	28.52	-2.0	292.0
loan_age_months	25981.0	16.02	30.39	0.0	1511.0
is_individual	25981.0	0.97	0.18	0.0	1.0

In Table 2 most facilities were held by individuals (96.5%), while a smaller corporate segment accounted for 3.5%. Borrowers were primarily male (68.5%), with females representing 22.5% and 8.9% unspecified. The book concentrated on the "Miscellaneous" economic sector (77.2%), followed by services (11.7%) and commerce/finance (4.9%). Regarding credit quality, 80.3% of exposures were classified as "Current," whereas 14.1% were in "Loss," with the remainder distributed among "OLEM," "Doubtful," and "Sub-standard."

Table 2. Frequency distribution of selected categorical variables

Variable	Category	N	%
Customer Type	Individual	25082	96.5
Customer Type	Corporate	899	3.5
Gender	Male	17805	68.5
Gender	Female	5851	22.5
Gender	Unspecified	2325	8.9
Economic Sector	Miscellaneous	20060	77.2
Economic Sector	Service	3033	11.7
Economic Sector	Commerce & Finance	1267	4.9
Economic Sector	Agriculture Forestry & Fishing	1099	4.2

Variable	Category	N	%
Economic Sector	Manufacturing	325	1.3
Economic Sector	Construction	191	0.7
Economic Sector	Cottage Industries	6	0.0
BOG Loan Classification	Current	20869	80.3
BOG Loan Classification	OLEM	559	2.2
BOG Loan Classification	Sub-Standard	419	1.6
BOG Loan Classification	Doubtful	467	1.8
BOG Loan Classification	Loss	3667	14.1

The target variable is 'is default', a binary label, coded as 1 = Yes, indicating that the loan eventually defaulted, while 0 = No indicates non-defaulted. The analysis revealed that 4,000 loans (approximately 15% of the total sample) were classified as defaults. This suggests that most (85%, 21981) loans remained non-defaulted.

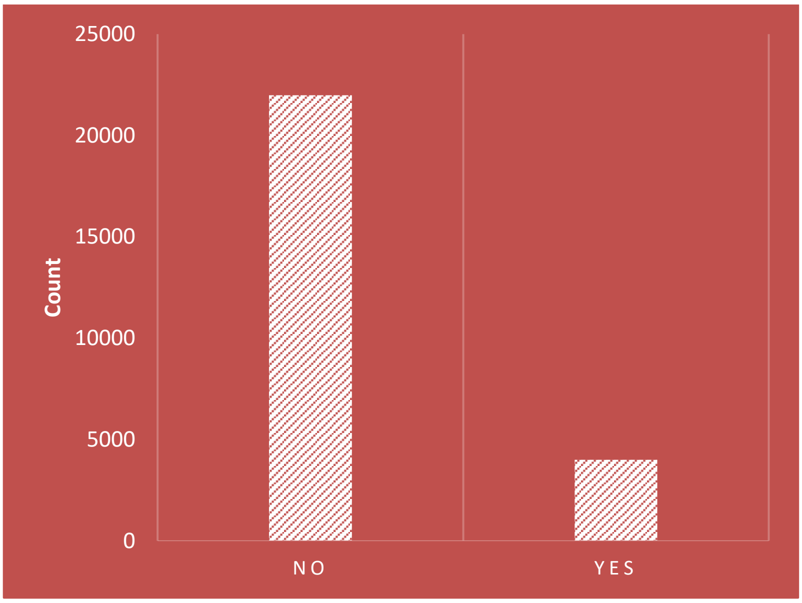

Figure 1 Graph of Default Versus Non-Default

4.2 Markov Chain modeling of Loan status

Loan repayment behavior was modeled as a discrete-time Markov Chain, where each state represents a BOG loan classification: S = {Current, OLEM, Sub-Standard, Doubtful, Loss}.

4.2.1 Distribution of Bank of Ghana (BoG) Loan classification.

Figure 2 shows the percentage of long-term loans in each category, assuming the process continues indefinitely. Current loans (80.32%) make up most of the loan portfolio, indicating good repayment performance. The presence of loans classified as OLEM (2.15%), Sub-standard (1.61%), Doubtful (1.80%), and Losses exceeding ten percent (14.11%) suggests that this Rural Bank's portfolio has some credit concerns.

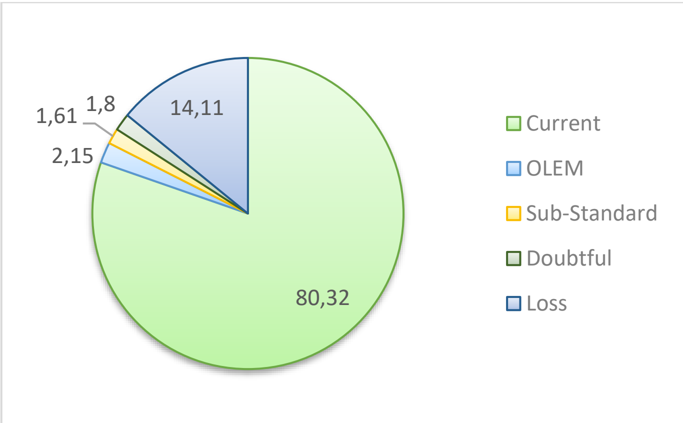

Figure 2: Pie Chart of the Steady State Probabilities (%)

4.2.2 Markov Chain transition Probability Matrix

The transition matrix displaying the probabilities of moving from one state or loan status to another within a single time step is illustrated as a heatmap, as shown in Figure 4.3. In this figure, darker blue represents a higher probability.

Figure 3 shows how the Bank's loans move between risk categories over time. Loans in the current category are highly stable - about 94% of the loans stay current from one month to the next. Only a small fraction slips into early warning or risky categories like OLEM or Loss, a good sign for the bank's overall loan portfolio health.

Interestingly, loans in the OLEM and Sub-Standard categories show a surprisingly high likelihood of reverting to current status, at nearly 74% and 63%, respectively. This may be due to borrowers catching up on payments, stringent monitoring, follow-ups with borrowers, or administrative updates that reclassify loans. However, these loans are not without risk—approximately 16.6% of OLEM loans and 15.8% of Sub-Standard loans still deteriorate into more severe categories such as Doubtful and Loss.

Figure 3 also indicates that the outcomes are more mixed among loans already classified in the Doubtful category. Just over half (56%) recover to current status, which seems optimistic, but a significant portion—more than 11%—ultimately falls into the Loss category, which is concerning.

Finally, loans in the Loss category rarely recover- over 90% stay in the Loss status, confirming that once a loan is written off, it seldom returns.

The diagonal elements in Figure 3 represent the probability of remaining in the same state. For example, the probability of remaining in the "Current (94.47%)" indicates a high chance of staying current, as reflected by the probability of remaining in the "Current" state of 0.9447. This implies that loans currently making their payments on schedule are likely to continue doing so. Furthermore, the chance of remaining Doubtful at 15.42% is comparatively low, with a probability of 0.1542. This suggests that deemed questionable loans are more likely to transition to different states, such as current or loss. In contrast, the probability of remaining in Loss is 90.92%, indicating a strong likelihood (0.9092) of staying in the "Loss" state. This implies that loans declared uncollectible are probably going to remain in this condition. Additionally, the probability of continuing to be in OLEM (9.48%) is relatively low at 0.0948. This suggests an increased likelihood of loans categorized as "OLEM" changing to another state. Finally, the diagonal point for Sub-Standard (11.22%) indicates a reasonable chance of remaining in the "Sub-Standard" state, with a probability of 0.1122. This further suggests that the likelihood of loans categorized as "Sub-Standard" moving to another state is slightly higher.

By comparing the diagonal elements shown in Figure 3, we see a greater degree of stability in the "Current" and "Loss" states, as evidenced by the significantly higher odds of remaining in these states compared to the others. Given the relatively low odds of staying in the "Doubtful," "OLEM," and "Sub-Standard" states, it is likely that loans in these categories will eventually transition to another state.

Further insight into the diagonal entries of TPM, as shown in Figure 6, reveals that the high likelihood of remaining in the "Current" stage suggests that loans currently up to date are repaying their debts effectively. Since these loans are more likely to transition to another state, the comparatively low odds of remaining in the "Doubtful," "OLEM," and "Sub-Standard" states indicate the need for close monitoring. Meanwhile, the high probability of remaining in the "Loss" condition highlights how crucial it is to manage credit risk effectively to minimize the chance of loan default.

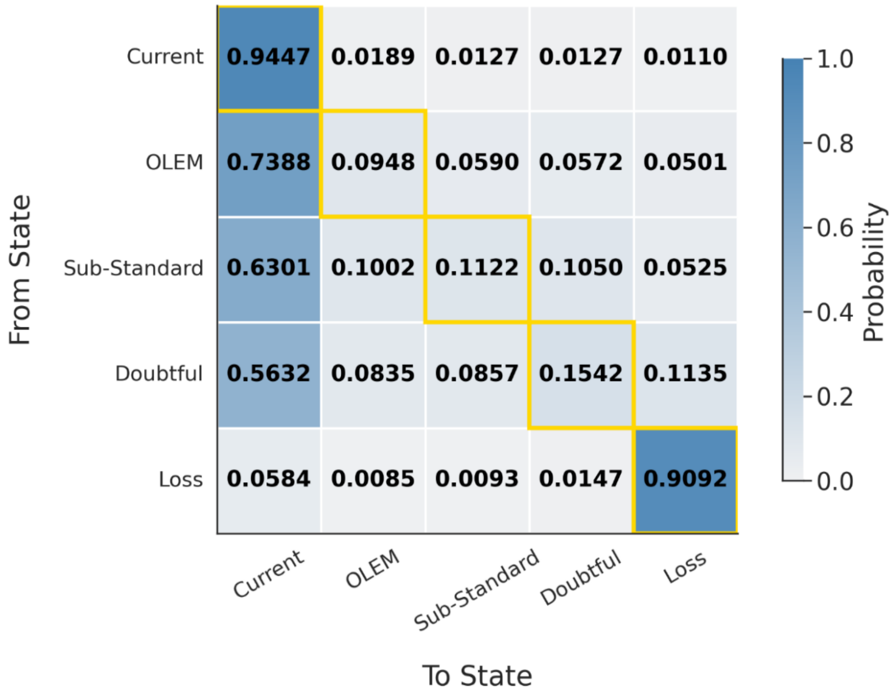

Figure 3 Heatmap of the Markov Chain Transitional Probability Matrix (TPM)

4.2.3 Cluster Analysis

The purpose of the clustering analysing in this research framework is to enable the bank to concentrate its monitoring and intervention strategies on particular segments that have similar repayment tendencies by using the K-Means clustering analysis to group loan statuses according to the similarity of their transition patterns.
Table 3 shows the cluster assignments for the K-Means. The cluster (3-Means cluster) analysis helps the Bank focus on similar-behaving groups of loans, making it easier to design monitoring and intervention strategies.
Table 3 is based on the similarity of the movement patterns of the loan status categories. The loan statuses have been categorized into clusters. “Current” is in group (2) because of its unique behavior from the others. However, "Doubtful," "OLEM," and "Sub-Standard" are comparable and lumped together into one cluster (0), while "Loss" is also distinct and clustered as (1).

Table 3: Cluster Assignment of the K-Means

Loan Status	Cluster
Current	2

Loan Status	Cluster
Doubtful	0
Loss	1
OLEM	0
Sub-Standard	0

4.2.4 Customers Risk Scores

Again, Figure 4 displays customers' risk scores according to their present loan status. The customer risk score is the probability that a loan in a given current status will transition to any of the "worse" (risk) states in the next period. Figure 4 shows the likelihood of a customer with "Current, OLEM, Sub-Standard, Doubtful, and Loss" statuses deteriorating as 5.53%, 26.12%, 39.16%, 43.68%, and 94.16%, respectively. This implies that a customer in "Current" status with a risk score of 0.055 has a 5.5% chance of transitioning to a risk status in the next period.

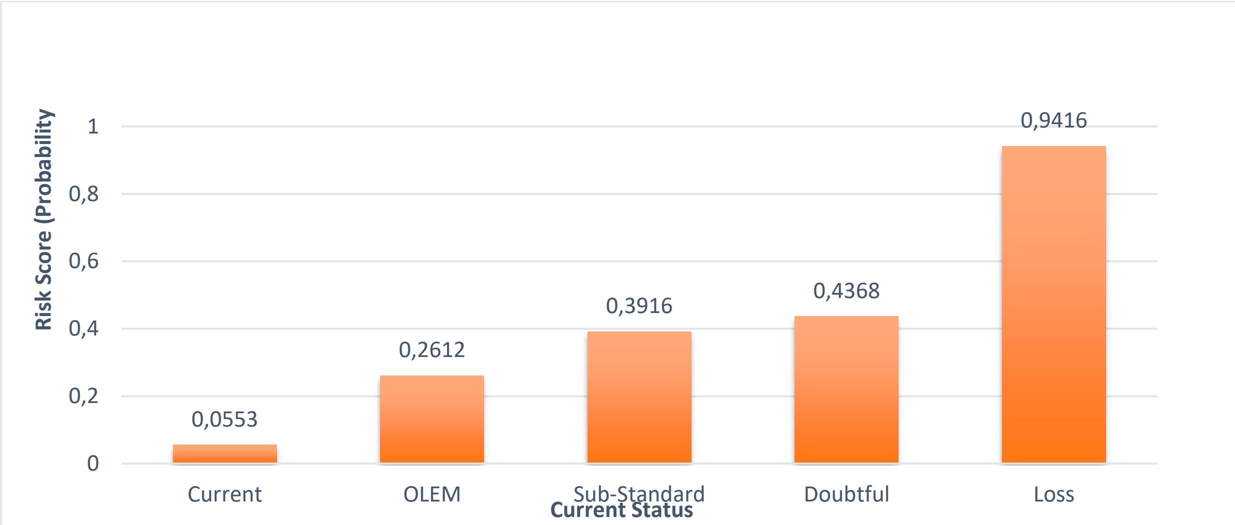

Figure 4: Risk Score by Current Loan Status

Also, Figure 5 shows the expected time to default. Expected time to default is the average number of periods it will take for a customer in a given status to reach the "default" (absorbing) state, assuming the Markov process continues.

Figure 5 shows that, a "Current" loan is expected to take about 69 months before potential default, while an "OLEM" loan has 66 months before potential default, with a "Sub-Standard" loan having about 63 months before defaulting, and a "Doubtful" loan has about 60 months (5 years) before potential default, and

This means that, if a customer's loan moves from "Current" to "OLEM", their expected time to default decreases by about 3 months (from 69 to 66 months).

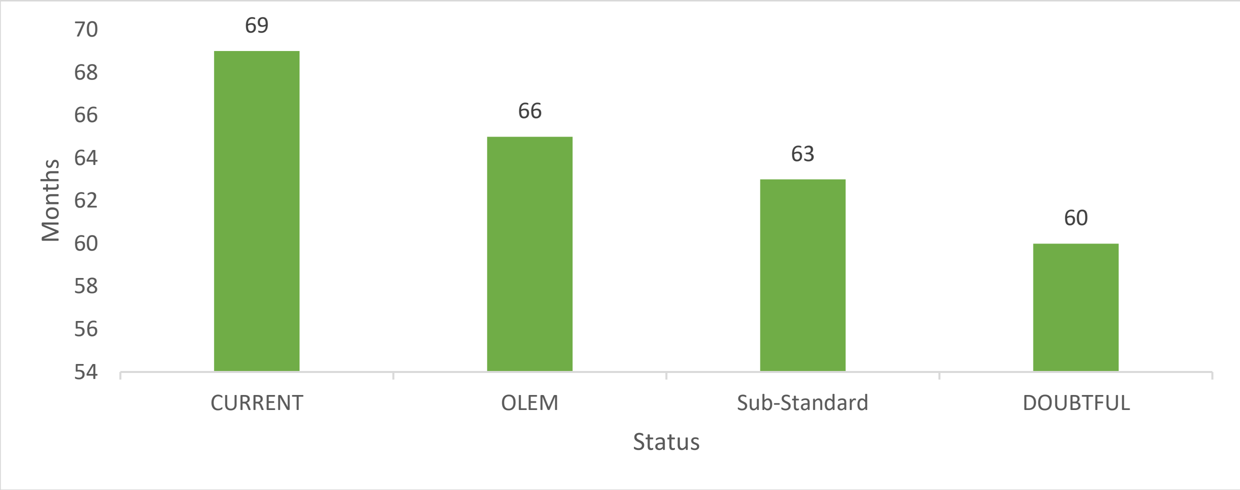

Figure 5: Expected Time of Default (Months)

Combining Figures 4 and 5 reveals the following insights:
1. For loans classified as "Current," the risk of deterioration is 5.53%, with an expected stable period of 5 years and 9 months, necessitating regular monitoring.
2. If the loan transitions to "OLEM," the risk increases to 26.12%, with the expected stable period decreasing to 5.5 years, requiring early intervention.
3. For loans classified as "Doubtful," the risk jumps to 43.68%, with the expected stable period dropping to 5 years, warranting immediate intervention.
These findings highlight the importance of timely risk assessment and intervention to mitigate potential losses.

4.3 Key Feature for Loan Default Prediction.
The objective was achieved by training a Random Forest model and evaluating its performance using various metrics. Default probabilities were simulated under different scenarios, and results were visualized using heatmaps and Receiver Operating Characteristic (ROC) curves.
Table 4 and Figure 6 present the top 10 influential features, accounting for approximately 98% of the model's predictive power. The top features are:
1. Days in arrears (0.300343)
2. IFRS classification (0.154330)
3. BoG loan classification (0.135914)
The least influential feature among the top 10 is the loan principal balance (excluding interest), with a weight of 0.010238.

Table 4: Key Variables Influencing Loan Default Prediction

Feature	Importance
Days in Arrears	0.300344
IFRS Classification	0.154330
BOG Loan Classification	0.135914
Interest Charge Type	0.108712
Loan Duration (Months)	0.078287
Loan Age (Months)	0.066854
Frequency of Payments	0.045482
Number of Payments Agreed	0.045155
BoG Provision	0.030630
Loan Principal Balance (without interest)	0.010238

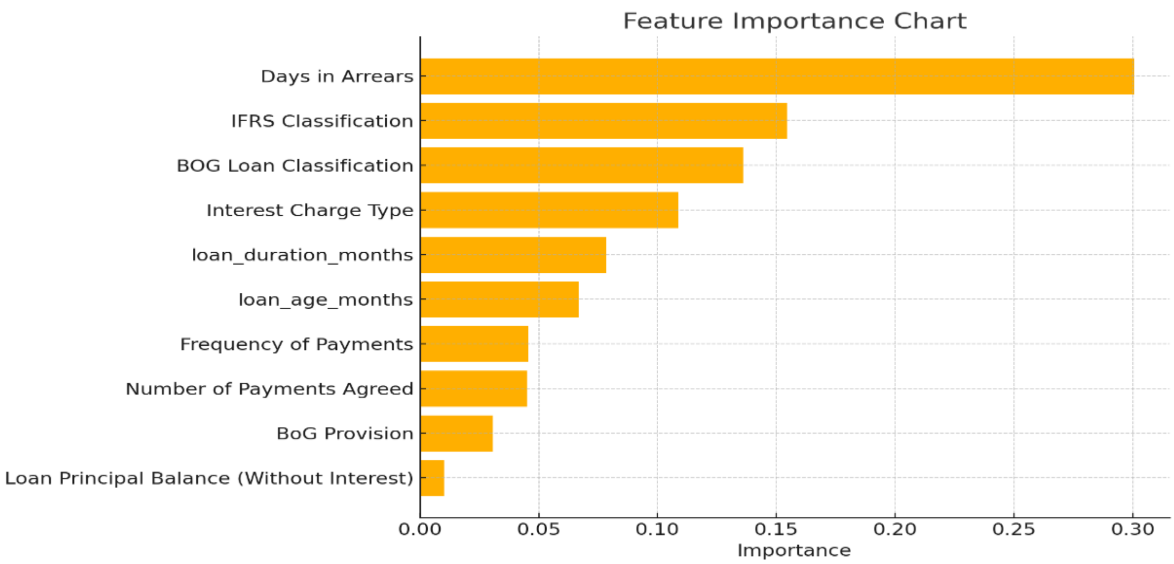

Figure 6: Graph of the Covariate Importance.

4.4 Model Performance Metrics Random Forest

The classifier demonstrated a commendable overall accuracy of 91%, indicating that it correctly classified 91% of the total 5,197 instances in the test dataset (See Table 5 and Figure 7). However, a closer examination of the class-specific metrics reveals nuanced performance across the two categories.

For the negative class (False, n = 4300), the model achieved a precision of 0.94, a recall of 0.93, and an F1-score of 0.93 as indicated on Table 5. These high scores indicate that the model was highly effective in identifying and classifying instances of the negative class. Precision, in this context, refers to the proportion of predicted negatives that were actually correct, while recall represents the proportion of actual negatives that were correctly predicted. The F1-score, a harmonic mean of precision and recall, confirms that the model maintained a balanced performance for this majority class.

In contrast, the performance metrics for the positive class (True, n=897n=897) were slightly lower. The model recorded a precision of 0.81, a recall of 0.84, and an F1-score of 0.82. Although these values are still indicative of strong performance, they suggest that the model is marginally less effective at identifying the minority class. The lower precision implies a higher false positive rate for the positive class, while the recall rate reveals that the model correctly identified 84% of all true positives.

To address class imbalance and provide a more comprehensive assessment, macro and weighted averages were also calculated. The macro-average values, which treat both classes equally regardless of their frequency, yielded precision, recall, and F1-scores of 0.87, 0.88, and 0.88, respectively. These scores indicate a reasonably balanced performance across the two classes. Meanwhile, the weighted average, which adjusts for the number of instances in each class, reported precision, recall, and F1-scores of 0.91 each—consistent with the overall accuracy and further affirming the model's strength in handling the dominant class.

Table 5: Random Forest Performance Metrics

	Precision	**Recall**	**F1-Score**	**Support**
False	0.94	0.93	0.93	4300
TRUE	0.81	0.84	0.82	897
ROC AUC Score			0.85	5197
Accuracy			0.91	5197
Macro Avg	0.87	0.88	0.88	5197
Weighted Avg	0.91	0.91	0.91	5197

The random forest demonstrates significant additional predictive value, while the ROC AUC offers a more thorough perspective on classification skill at all decision thresholds.

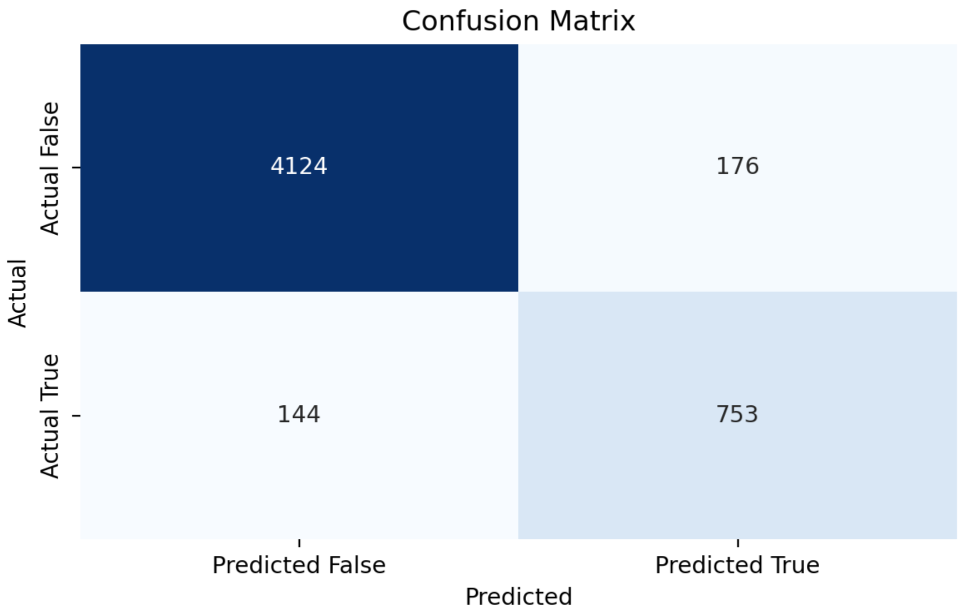

Figure 7: Confusion Matrix of the Random Forest Prediction

Additionally, simulations for different scenarios were conducted using the developed model. Figure 8 displays the scenario simulation heatmap illustrating how the probability of default varies as "Days in Arrears" and "Loan Age (months)" change for a typical customer at the Bank. The darker red in Figure 8 indicates a higher probability of default. As the "Days in Arrears" increases, there is also a sharp rise in the risk of default, especially for loans with a longer duration (age) in months. For instance, the risk increases after 2 months in arrears, but particularly after 3 months in arrears.

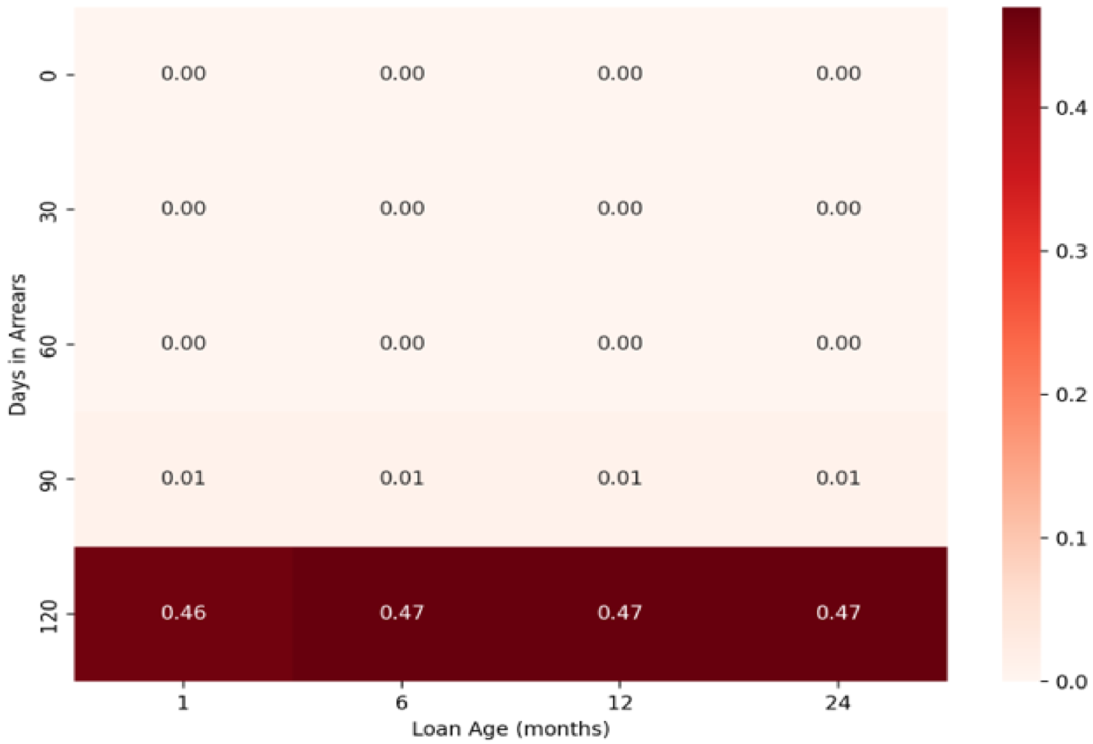

Figure 8: Simulated Default Probability by Days in Arrears and Loan Age (Months)

4.5 Improved Prediction Using Hyperparameter Tuning

From Table 6, after hyperparameter tuning, the random forest model exhibited robust and well-balanced predictive performance. Overall accuracy reached 0.93, indicating that 93% of cases were accurately classified across the ten cross-validation folds. The receiver operating characteristic area under the curve (ROC-AUC) was 0.90, reflecting excellent global discrimination between defaulting and non-defaulting loans (Hosmer, Lemeshow, & Sturdivant, 2013).

Class-specific results show a modest trade-off between majority and minority groups. For non-defaults, precision and recall were 0.95 and 0.94, respectively, resulting in an F1-score of 0.94. For defaults—arguably the more critical category—precision improved to 0.85 and recall to 0.88, yielding an F1-score of 0.86. These values indicate that the tuned model correctly identifies 88% of true defaults while limiting false-positive alerts to about 15%.

Macro-averaged metrics (precision = 0.90; recall = 0.91; F1 = 0.90) further confirm balanced performance when each class is weighted equally, whereas weighted averages (≈ 0.93) mirror headline accuracy by accounting for class prevalence (defaults ≈ 17 %).

Table 6: Optimized Model Performance Metrics

	Precision	Recall	F1-Score	Support
False	0.950	0.940	0.945	4300
TRUE	0.850	0.880	0.865	897
ROC AUC Score			0.90	5197
Accuracy			0.93	5197
Macro Avg	0.90	0.91	0.90	5197
Weighted Avg	0.93	0.93	0.93	5197

The hyper-parameter-optimized random forest model yields a sharply rising default probability curve as payment delinquency extends (see Table 7). Accounts that are current (0–30 days past due) exhibit virtually no risk of default; the estimated PD is 0%. Risk rises significantly once arrears reach 60 days, with an 11.6% chance of default. The slope steepens between 60 and 90 days, more than doubling the PD to 28.5%. At 120 days past due, the probability of default escalates to 85.5%, indicating that loans which are four months in arrears are very likely to become charge-offs.

Table 7: Daily Loan Risk Assessment

Days in Arrears	Default Probability (%)
0	0.00
30	0.00
60	11.60
90	28.52
120	85.49

5. Summary, Conclusion and Recommendations

This section considers the summary, the conclusions and the recommendation of the study.

5.1 Summary

The primary goal of this study was to create an integrated credit-risk framework for a rural bank by combining Markov chain clustering analysis with a machine-learning classifier. The developed model from this study results indicated that (a) most loans stayed in the Current state or, once written off, remained in Loss; (b) intermediate states showed significant two-way mobility; and (c) a tuned random forest model achieved high predictive accuracy (overall = 0.93, ROC AUC = 0.90) while identifying 88% of actual defaults. These findings support previous research suggesting that early-stage delinquency is highly useful for predicting defaults (e.g., Brown & Moles, 2021) and expand the literature by confirming this pattern in the context of a rural bank.

The Markov results further showed that the expected time to default varies significantly by initial grade, decreasing from about 66 months for OLEM accounts to 63 months for Sub-Standard accounts. This gradient highlights the importance of detailed provisioning under IFRS 9, as lifetime expected credit losses differ greatly across BoG buckets. Additionally, scenario simulations based on the tuned random forest revealed a non-linear increase in default probability once arrears exceeded 60 days (PD = 11.6%) and nearly certain loss at 120 days (PD = 85.5%). These time-based insights stress that the best intervention window is before the 90-day mark, when default probability rises sharply.

Additionally, we created a Markov-chain map that illustrates the progression of loans from “Current” to “Loss”. The developed model gives a realistic and practical guide for action. The model additionally provided a coherent, data-driven solution for rural bank credit management.

The deductions made from the results of the developed model are; most loans are positioned at the secure extremes of the spectrum. These loans either remain healthy or, in cases of default, they become difficult to recover. The perilous zone is found in the middle. This implies that, around the two-month overdue threshold, the likelihood of complete default significantly increases. The model provided a framework that enhances the bank’s risk-management capability beyond conventional rule-of-thumb methods**.** This enables the manager to identify high-risk customers that had to be contacted over time without the manager relying on gut feelings. The model further unraveled that early outreach to customers, a payment plan offer, and a reminder through any form of communication sent 30 to 60 days past due are very crucial in the bank’s strategy to mitigate the risk of default by the customers.

In addition, there are three indicators of concern in this study, which were the primary causes of default. These primary causes are, the days past due, the official IFRS stage, and the Bank of Ghana’s designated loan-quality classification.

Finally, the model showed an accuracy level of 93%, indicating how efficient the model could be applied in the banking sector by capturing the long-term transition dynamics and the borrower-level risk scores, which are suitable for daily decision-making.

5.2 Conclusions

In this study, a hybrid Markov Chain and Random Forest model for loan default in rural banks in Ghana have been developed. Also, the feature importance of the model has been determined to be ten major variables including, days in arrears, IFRS classification, BoG classification, interest charge type, etc. Finally, the hyperparameter tuning of the model has been established.

5.3 Practical Implications

From the outcomes of the study, the following are observed as the practical implications associated to the study. On provisioning; the transition-based expected-loss estimates help improve the accuracy of Stage 2 and Stage 3 classifications, lowering the chances of under- or over-reserving. Also, on the early-warning triggers; it was seen that the tuned classifier achieves 88% recall for defaults, enabling automated alerts when a borrower's predicted PD exceeds a management-defined threshold (e.g., 20%). Moreover, on resource allocation; collection efforts can be tiered for low-touch reminders for 0–30 days overdue, personalized outreach at 31–59 days, and full workout escalation at 60 days or more. Finally, on pricing and product design; the study unraveled that, risk-based interest-rate add-ons can be calibrated using point-in-time PDs generated by the random forest, promoting a more sustainable credit portfolio.

5.4 Recommendations for Future Research

Researchers might expand this work by (a) incorporating macroeconomic leading indicators into the feature set to enable macro scenario forecasting, (b) comparing the Markov–random forest hybrid with gradient boosting or deep learning alternatives, and (c) conducting a multi-bank study to evaluate model transferability across various rural institutions. Additionally, exploring explainable AI techniques (e.g., SHAP values) could help bridge the gap between model accuracy and regulatory transparency.

References

[1] Abraham, H. R. (2023). Appraisal Discrimination: Five Lessons for Litigators. *SMU L. Rev.*, *76*, 205.

[2] Agyapong, D. (2023). *Implications of Monitoring and Evaluation Systems for Small and Medium-Sized Enterprises in Selected Metropolis in Ghana Implications of Monitoring and Evaluation Systems for Small and Medium-Sized Enterprises in Selected Metropolis in Ghana* (Doctoral dissertation, University of Cape Coast).

[3] Brown, K., & Moles, P. (2014). Credit risk management. *K. Brown & P. Moles, Credit Risk Management*, *16*.

[4] Ekpu, V., & Paloni, A. (2016). Business lending and bank profitability in the UK. *Studies in Economics and Finance*, *33*(2), 302-319.

[5] Golin, J., & Delhaise, P. (2013). *The bank credit analysis handbook: a guide for analysts, bankers and investors*. John Wiley & Sons.

[6] Oseni, E. (2023). Assessment of the five cs of credit in the lending requirements of the Nigerian commercial banks. *International Journal of Economics and Financial Issues*, *13*(4), 58.
[7] Skiba, P. M., & Tobacman, J. (2008). Payday loans, uncertainty and discounting: Explaining patterns of borrowing, repayment, and default. *Vanderbilt Law and Economics Research Paper*, (08-33).
[8] Vermoesen, V., Deloof, M., & Laveren, E. (2013). Long-term debt maturity and financing constraints of SMEs during the global financial crisis. *Small Business Economics*, *41*(2), 433-448.