Data Science in Finance and Economics

DSFE, 6(1): 31–57.
DOI: 10.3934/DSFE.2026002
Received: 27 March 2025
Revised: 30 August 2025
Accepted: 12 November 2025
Published: 14 January 2026

https://www.aimspress.com/journal/dsfe

Research article

Investigating the accuracy of credit bureau data in predicting borrowers' repayment of consumer loans in Nigeria

Adedeji Olowe*

Lendsqr, Inc. 2055 Limestone Rd, Ste 200C Wilmington, Delaware 19808, USA

* **Correspondence:** Email: adedeji@lendsqr.com; Tel: +13478507035.

Abstract: Consumer lending, especially microloans, is a critical component of financial inclusion and economic growth. However, the increasing reliance on credit bureau data to assess borrowers' repayment capabilities raises questions about its predictive accuracy. In this study, we evaluated the effectiveness of credit bureau data in predicting loan repayment behavior using a dataset from First Central, a Nigerian credit bureau, and Irorun, a digital lending institution in Nigeria. A total of 3,741 loan applicants, aged 21 to 60, were analyzed using traditional statistical tools and machine learning (ML) models, including logistic regression, random forests, and gradient boosting. Our results indicated that while credit bureau data provides some predictive insights, its standalone accuracy is limited due to inconsistencies in borrower credit histories and incomplete data reporting. Correlation analyses show weak associations between borrower-reported and bureau-reported overdue loan records and repayment outcomes, with Cramer's V values below 0.05. ML models, particularly gradient boosting, outperformed traditional statistical approaches, achieving an AUC-ROC of 0.77, highlighting the potential of advanced algorithms in credit risk assessment. Our findings suggest that integrating alternative borrower data, such as utility bill payments and digital transaction records, could enhance credit risk modeling. The study emphasizes the need for improved credit reporting accuracy and regulatory measures to ensure comprehensive borrower profiles. Enhancing predictive models with supplementary data sources can mitigate default risks and promote responsible lending, leading to greater financial inclusion in Nigeria.

Keywords: credit bureau data; loan repayment; machine learning; financial inclusion; digital lending; microloans; Nigeria

JEL Codes: G21, G17

1. Introduction

Consumer lending plays a critical role in fostering economic growth and financial inclusion, providing individuals with resources to meet personal and entrepreneurial needs (Falaiye et al., 2024; Dulloo, 2021). As a key driver of domestic consumption, consumer credit enables households to improve their living standards, enhance productivity, and contribute to economic resilience (Wang et al., 2024; Cohen, 2007). In developing economies like Nigeria, the growth of consumer loans is pivotal in promoting financial inclusion by allowing more people to participate in the formal economy, especially those from underserved or low-income backgrounds (Li and Peng, 2023; Kama and Adigun, 2013; Dev, 2006). This access not only reduces reliance on informal or predatory lenders but also supports broader economic participation and poverty reduction (Demirgüç-Kunt et al., 2018; Demirgüç-Kunt et al., 2017). However, the effectiveness of consumer loans in advancing economic growth and financial inclusion is intricately tied to borrowers' repayment behavior, which has become an area of focus as digital lending rapidly expands.

Consumer lending is a cornerstone of modern financial systems, playing a crucial role in advancing economic development, household welfare, and financial inclusion. Broadly, consumer loans can be classified into distinct types, including personal loans (typically unsecured and used for general purposes), credit card loans (revolving credit with variable repayment structures), housing loans or mortgages (secured long-term loans for home purchase), vehicle loans (used to finance automobile purchases), and education loans, which serve as an investment in human capital with long-term societal returns. Each of these categories differs in terms of risk structure, interest rates, borrower profiles, and socio-economic impact.

In the context of developing economies like Nigeria, consumer lending, especially through microloans and digital credit platforms, has become increasingly prominent. According to the Central Bank of Nigeria and Enhancing Financial Innovation and Access (EFInA), financial inclusion in Nigeria has improved from 53.6% in 2016 to 64.1% in 2021, driven largely by mobile banking, fintech innovation, and microfinance expansion. Despite this progress, nearly 36 million Nigerian adults remain excluded from the formal financial system. Additionally, Non-Performing Loan (NPL) ratios in the consumer lending segment have fluctuated significantly over the past five years, reaching 10.8% in 2017, dropping to 5.9% in 2020, and rising again to 6.3% in 2022, highlighting persistent risks in credit administration (CBN, 2023).

The volume of consumer loans issued by Nigerian financial institutions has increased significantly, growing at an average annual rate of 12–15% since 2018. This expansion reflects increasing demand for short-term liquidity among low- and middle-income households, especially through digital lending platforms, which now constitute over 40% of microloan disbursements in urban regions (EFInA, 2022). However, this rapid growth in credit access is paralleled by rising concerns around default risk, creditworthiness assessment, and data reliability.

A growing body of research now emphasizes the use of machine learning (ML) models to improve credit scoring accuracy. Traditional credit scoring systems often rely on logistic regression and historical repayment patterns, but ML models such as Random Forests, Support Vector Machines, and

Gradient Boosting Machines can handle complex, nonlinear relationships and high-dimensional datasets. These models have shown considerable improvements in predictive accuracy, especially in unstable data environments (Lemieux et al., 2023; Darwish, 2025). For instance, gradient boosting has been used in several studies to outperform linear models in predicting loan default risk across various countries and platforms (Nguyen et al., 2025; Bhandary and Ghosh, 2025). However, literature applying these models to credit bureau data in developing economies remains scarce, which this study aims to address.

Studies support the idea that sustainable financing mechanisms are vital for inclusive growth. In the Nigerian context, consumer loan schemes have faced challenges with weak enforcement, difficulties in repayment tracking, and inadequate credit assessment, highlighting a gap that could be addressed through better risk modeling and credit analytics (Shen and Ziderman, 2009; Frank, Bhandary, and Prabhu, 2024). Despite digital infrastructure advances, credit bureau data use for consumer loan assessment in Nigeria faces critical challenges such as incomplete reporting, inconsistent data formats, and low borrower coverage. Here, we aim to evaluate the predictive validity of credit bureau data using both traditional and ML approaches, assessing their effectiveness in modeling repayment outcomes in the Nigerian consumer credit market.

1.1. Importance of consumer loans to economic growth and inclusion

Consumer loans are essential in enabling individuals to make purchases, such as housing, education, and healthcare, that might otherwise be unattainable (Dulloo, 2021) These expenditures, in turn, stimulate economic activity across various sectors, creating employment opportunities, increasing tax revenues, and promoting a more integrated financial ecosystem (Li and Peng, 2023). Additionally, as a major component of financial inclusion, consumer loans make financial services accessible to previously excluded populations, enabling them to establish credit histories and build financial resilience. Financial inclusion is essential for sustainable economic growth, as it reduces poverty and narrows income inequality by integrating more people into the formal financial sector (Beck et al., 2000).

The role of consumer loans in supporting economic inclusion has gained particular attention in emerging economies, where improving access to credit is often a national priority (Ediagbonya and Tioluwani, 2023; Chibba, 2009). In Nigeria, where millions remain unbanked, consumer loans bridge the gap by providing credit access to individuals traditionally excluded from formal financial services (Agwu, 2021). As a result, greater financial inclusion has the potential to empower individuals, foster small business development, and contribute to economic growth (Soetan et al., 2021).

While consumer loans broadly encompass various types of personal borrowing, we focus on microloans; small, short-term credit products designed for individuals with limited access to traditional banking services (Iganiga, 2008). In Nigeria, microloans typically range between ₦1,000 and ₦50,000 and are often issued by digital lenders or microfinance institutions. Unlike consumer loans in developed economies, where microloans may be considered insignificant, in Lagos or other Nigerian cities, these amounts can play a crucial role in household financial stability and small-scale business operations (Kashim, 2018).

1.2. Growth of digital lending in Nigeria and its economic impact

Over the past decade, digital lending has surged in Nigeria, driven by technological advancements, mobile penetration, and the need for accessible credit (Ezie et al., 2023; Kola-Oyeneyin et al., 2020). Digital lending platforms have emerged as pivotal players in consumer finance, providing quick access to loans through streamlined, app-based processes that eliminate traditional barriers such as extensive paperwork and in-person bank visits (Nnaomah et al., 2024). According to data from Enhancing Financial Innovation & Access (EFInA, 2021), digital lending has increased access to credit for millions of Nigerians, including those in rural and underserved regions. This extended access implies increased financial inclusion by reaching individuals who might otherwise be excluded from conventional financial services (Kola-Oyeneyin et al., 2020).

The economic impact of digital lending in Nigeria has been substantial, particularly during periods of economic downturn, such as the COVID-19 pandemic (Ezie et al., 2023; Okoroafor, 2024). Digital lending provided essential liquidity to individuals and small businesses, helping them weather economic challenges and maintain operational continuity (Ehiedu et al., 2022; International Finance Corporation, 2020). However, digital lending has also introduced new challenges, particularly concerning credit risk management. The swift loan disbursement process often relies on limited credit checks, increasing the likelihood of defaults. Consequently, lenders and regulatory bodies are increasingly leveraging credit bureau data to enhance credit assessment and mitigate these risks.

1.3. The rise in the use of credit bureau data in consumer lending in Nigeria

The rise in digital lending has underscored the importance of accurate credit assessments, prompting a growing reliance on credit bureau data for consumer loan decisions in Nigeria. Credit bureaus such as First Central play an essential role by aggregating and analyzing borrowers' credit histories, repayment, and other financial data to provide a holistic view of creditworthiness (Dyché and Levy, 2006). As consumer lending expands, credit bureau data has become integral to risk management practices, helping lenders make informed decisions by identifying high-risk borrowers and preventing over-indebtedness.

Credit bureau data serves not only to inform lenders of the applicants' financial backgrounds but also to foster transparency and trust in the credit system. For digital lenders, who often operate with limited personal borrower information, credit bureau data fills a critical gap, improving the precision of risk assessments and enabling sustainable lending growth. Nevertheless, there remain challenges regarding the consistency and completeness of credit bureau data, particularly for borrowers with limited or inconsistent credit histories (Dyché and Levy, 2006; Avery et al., 2004). Addressing these gaps is crucial to further strengthening Nigeria's credit infrastructure and ensuring that consumer loans can drive financial inclusion without escalating default risks.

1.4. Impact of poor loan repayment behavior on lending

Despite their benefits, consumer loans also present risks, particularly when borrowers struggle to meet their repayment obligations. Poor repayment behavior threatens the sustainability of lending institutions, as non-performing loans (NPLs) reduce capital availability and increase the cost of credit (Ehiedu et al., 2022). High default rates prompt lenders to raise interest rates, which makes credit less

affordable and accessible, especially to low-income consumers who could benefit most from credit access. In Nigeria, where NPLs are a significant concern, high default rates deter financial institutions from offering consumer loans, thus constraining financial inclusion efforts (Central Bank of Nigeria, 2022).

An environment marked by frequent defaults can have broader economic impacts. For instance, the risk of widespread loan defaults can diminish investor confidence, reduce foreign investment inflows, and ultimately stifle economic growth (Chen, 2022). Therefore, improving loan repayment prediction accuracy through rigorous credit assessments is vital. Effective predictive models enable lenders to reduce default risk, adjust lending terms, and ensure that consumer lending can contribute positively to the economy without compromising institutional stability.

The rise in consumer credit in Nigeria has increased reliance on credit bureau data for assessing borrowers' creditworthiness by providing detailed borrower histories. However, the effectiveness of this data for predicting loan repayment remains uncertain, and limited research exists on the accuracy and predictive power of these data sources in this context. We investigate how well credit bureau data predict repayment outcomes for consumer loans in Nigeria. By analyzing data from First Central, a Nigerian credit bureau and Irorun, a digital lending institution, this research aims to provide insights into factors that can improve credit risk assessment, enabling lenders to reduce defaults and foster responsible lending practices in Nigeria.

1.5. Research questions

1. To what extent do credit bureau variables predict loan repayment outcomes compared to borrower-reported data in Nigeria?
2. Which borrower and bureau-reported variables are the most significant predictors of loan repayment behavior?
3. How do ML models (Logistic Regression, Random Forest, Gradient Boosting) improve the predictive accuracy of credit risk assessment?
4. What are the limitations of the current credit bureau data infrastructure in providing reliable credit assessments in Nigeria?

1.6. Statement of problem

Credit scoring models have long been used for predicting default risk in various contexts (Altman, 1968; Hand & Henley, 1997). However, the applicability of these models in developing economies remains under-examined. Studies from emerging markets highlight data limitations, inconsistent reporting, and lack of credit history as barriers to accurate credit scoring (Hand & Henley, 1997). In Nigeria, where credit bureau coverage is expanding, assessing the accuracy of credit bureau data for repayment prediction is crucial to enhance financial inclusion and improve lending outcomes. Here, we build on research in emerging economies, examining how self-reported and bureau-collected data align to enhance predictive accuracy.

The rapid growth of consumer lending in Nigeria, especially through digital platforms, has intensified the need for accurate credit risk assessment to manage default risks effectively. However, the reliability of credit bureau data in predicting loan repayment behavior in this context remains uncertain. High default rates not only strain lending institutions but also threaten financial inclusion

efforts by making credit less accessible and affordable. We address the problem of whether credit bureau data alone can accurately predict consumer loan repayment outcomes in Nigeria and aims to determine which factors most strongly indicate repayment behavior. By examining the predictive validity of credit bureau data, we seek to provide insights that could improve credit assessment models and enhance the sustainability of consumer lending in Nigeria.

1.7. Theoretical framework and behavioral foundations in credit risk

Understanding borrower repayment behavior requires a multidimensional theoretical approach, combining insights from psychology, economics, and financial technology. Among the most prominent theories relevant to credit behavior are the Attitude-Behavior Theory, Human Capital Theory, and Debt Sustainability Framework.

The Attitude-Behavior Theory posits that an individual's attitudes, subjective norms, and perceived behavioral control shape their intentions and, ultimately, their actions. In loan repayment contexts, this theory explains why some borrowers may default despite having the means to repay, due to weak financial literacy, moral hazard, or social norms that do not prioritize formal credit obligations (Ajzen, 1991). Understanding this psychological dimension is critical when interpreting borrower-reported data, which may not always align with actual credit bureau records or predictions.

The Human Capital Theory, developed by Becker (1964) and refined by Schultz and others (Melton, 1965), frames education and skills acquisition as investments that yield future economic returns. In the context of education loans, this theory supports the view that financing education enhances productivity, employability, and long-term repayment ability. Empirical studies suggest that countries with well-structured student loan systems, such as Australia, South Korea, and Chile, achieve higher repayment rates when loans are matched with labor market outcomes (Shen and Ziderman, 2009; Johnstone and Marcucci, 2010). This connection highlights the importance of predictive models in educational credit, ensuring that financing is matched to future income potential.

The Debt Sustainability Framework complements the behavioral perspectives. It is often used by financial institutions to assess a borrower's long-term capacity to manage debt, and as a tool to shape fiscal behavior. It considers income levels, interest rates, and debt-to-income ratios to predict financial distress. In the case of microloans and digital lending in Nigeria, high repayment burdens relative to volatile incomes may breach sustainable debt thresholds, necessitating early detection through reliable credit modeling.

1.8. Connecting behavioral theories, credit data, and ML

The interplay between borrower behavior and credit data is complex. Self-reported data often reflect perceptions, intentions, and limited financial awareness, while credit bureau data offer objective but potentially incomplete snapshots of credit history. ML models excel at reconciling these data streams by identifying nonlinear patterns, hidden correlations, and interaction effects across diverse variables. By doing so, ML models translate behavioral indicators into predictive risk signals, offering a hybrid solution that is data-driven and behaviorally informed (Darwish, 2025; Baesens et al., 2016).

Furthermore, ML models, especially Random Forests and Gradient Boosting, have demonstrated superior performance in processing unstructured and imbalanced datasets, which are common in the credit environments of developing countries. These models learn from patterns that may not be evident

through traditional statistical techniques, offering a more robust basis for credit decision-making in contexts with limited financial infrastructure or borrower data.

1.9. Recent research on ML in credit risk modeling

There has been an increase in the application of ML techniques to credit risk prediction, particularly in sparse and imbalanced borrower data. Among the most widely adopted models are Logistic Regression (LR), Random Forest (RF), and Gradient Boosting Machines (GBMs), each offering unique strengths in predictive performance and interpretability. Logistic Regression, while traditionally used in credit scoring, continues to serve as a benchmark for more complex models due to its statistical robustness and ease of interpretation. However, non-linear ensemble methods, such as Random Forests and Gradient Boosting, have demonstrated superior accuracy and robustness in handling multicollinearity, missing values, and imbalanced class distributions. Based on research data, these models are particularly effective in datasets that involve borrower heterogeneity and nonlinear repayment patterns.

Barboza et al. (2017) compare multiple ML techniques for predicting corporate bankruptcy and find that Gradient Boosting Machines outperformed all other models, including Support Vector Machines and Neural Networks, in terms of Area Under the Curve (AUC) and F1 scores (Bhandary and Ghosh, 2025; Demirhan, 2024; Yadav and Awasthi, 2024). Their results emphasize the importance of feature engineering and model tuning for maximizing predictive power. Similarly, other research including systematic reviews of ML applications in financial consumer behavior and reported that Random Forest and GBM models consistently yielded higher prediction accuracies, particularly in datasets involving repayment defaults and credit bureau records (Meng et al., 2025; Nazareth and Reddy, 2023; Valecha et al., 2018). These studies also highlight the role of demographic and transactional data in enhancing model performance. Other researchers investigated household loan repayment behavior in emerging markets and applied ensemble learning techniques to assess the impact of financial literacy and behavioral variables (Mallinguh and Wasike, 2025; Dinh and Thanh, 2022; Chen, 2022). The findings confirmed that nonlinear models like GBM captured complex behavioral predictors more effectively than linear models, suggesting significant implications for credit risk management in underbanked regions.

Drawing from the outcome of these studies and more, we sought to explore the use of ML in credit scoring as a tool to assess the accuracy of credit bureau data in predicting borrowers' repayment of consumer loans in Nigeria.

2. Methodology

We adopted an action research design, utilizing real-world loan applicant data from Irorun, a licensed mid-sized (customer base of 520,000) digital lending institution based in Oyo State, Nigeria. The objective was to investigate how well credit bureau data can predict borrower repayment behavior, using ML tools to model real-world lending scenarios.

A convenience sampling approach was employed to gather data from 3,741 approved loan applicants between June 1, 2024 and September 30, 2024. While convenience sampling is not typically preferred in inferential research, it is suitable in this context because we seek practical insights from a specific operational dataset, rather than aiming to generalize across the population of Nigerian

borrowers. The sample was drawn from Irorun's active customer base of approximately 520,000 users, meaning the sample represents roughly 0.72% of the population. This proportion is deemed analytically sufficient for ML experiments and exploratory modeling, especially when real-world lending data are involved and when class labels (repayment outcomes) are available for supervised learning.

The participants, aged between 21 and 60, had approved loans from the lender. Each participant had filled out a survey during the loan application process that formed the self-reported data used in this study. The given loans ranged from 1,000 to 10,000 Naira in amounts. Although the loan amounts studied in this research may seem small from a global perspective, they are significant within the Nigerian microfinance landscape. At the time of data collection, the national minimum wage was ₦30,000 per month, meaning that the loans studied represented between 3.33% and 33% of a minimum-wage worker's salary. For microlenders in Nigeria, these figures are meaningful, as even small loans can provide crucial financial support to individuals who may otherwise lack access to credit. The relative size of these loans should be understood within the local economic context, where short-term borrowing often plays a vital role in daily financial management and informal business activities.

The data analysis methodologies chosen for the research, includes both traditional statistical tools and ML models (logistic regression, random forests, and gradient boosting). The traditional methods like chi-square tests and correlation analyses were used to identify the relationships between variables but they mostly failed to provide robust predictions. However, ML models help maximize predictive accuracy, which makes them ideal for identifying patterns in repayment outcomes and enable a better understanding of borrower repayment behavior compared to the traditional statistical methods.

2.1. Data sources and overview

The data were collected from a Nigerian credit bureau records and self-reported borrower information from a participating lending institution. The dataset includes:

1. Demographic information: Gender, marital status, number of dependents and device type.
2. Credit information: loan repayment history, loan amount, loan approval date, and credit utilization rate as expressed in the borrower's loan history from credit bureau data.
3. User reported repayment records: Binary indicator of on-time repayment or default and numerical values of unpaid loans.

The key variables include:

1. Loan status (Dependent Variable): Whether loans were "settled" or "past due."
2. Credit metrics: Loan history, Number of overdue, active, and delinquent loans as reported by both borrower and bureaus.
3. Comparison fields: Fields comparing self-reported data with bureau findings, including whether overdue loans matched between the two sources.

2.2. Data analysis techniques

To assess the predictive power of credit bureau data, we employed logistic regression alongside ML models, including random forests and gradient boosting. Descriptive statistics was used to assess the general distribution and patterns within loan status and credit variables, while Chi-Square Tests

and Cramer's V statistical tools were used to measure associations between comparison fields (e.g., overdue loan alignment between borrower and bureau reports) and loan status.

The ML tool, Logistic Regression, was used to establish a baseline with statistically significant predictors and interpretable relationships, while Random Forest Classifier and Gradient Boosting were used to capture non-linear interactions and identify feature importance for predicting loan status because of their suitability for handling high-dimensional data, especially with complex data interactions.

2.3. Software and tools used

All data preparation, model development, and evaluation were conducted using SPSS (v. 23) and the Python programming language (version 3.10) due to its robustness in data science workflows and its rich ecosystem of ML libraries. Key Python packages used included:

1. *pandas* and *numpy* for data preprocessing and statistical manipulation.
2. *scikit-learn* for model development (logistic regression, random forest, gradient boosting).
3. *matplotlib* and *seaborn* for data visualization and plotting confusion matrices and feature importance.

Python was chosen over other tools such as R, or Excel due to its flexibility, scalability, and extensive use in modern credit risk modeling and data science applications.

2.4. Model selection and rationale

To evaluate the predictive performance of credit bureau data, three supervised ML algorithms were employed: Logistic Regression, Random Forest, and Gradient Boosting. Each model was chosen based on its unique strengths and applicability to financial prediction tasks based on previous studies:

1. Logistic Regression (LR) is a classical statistical model often used as a benchmark in credit scoring due to its interpretability and computational efficiency. LR is suitable for binary classification problems like loan repayment status (*settled* vs. *past due*), especially when the relationship between predictors and outcomes is linear. While it offers transparent insights into feature coefficients, it may underperform with complex, nonlinear interactions.
2. The Random Forest (RF) model is an ensemble learning method based on decision trees. RF is well-suited for capturing nonlinear relationships and handling high-dimensional datasets. It provides built-in mechanisms for assessing feature importance, is robust to overfitting, and performs well even with imbalanced or noisy data. It is particularly effective in modeling credit behavior where interactions between variables (e.g., *loan amount* and *marital status*) influence repayment outcomes.
3. The Gradient Boosting Model (GBM) builds on the principle of iterative learning by combining weak learners to create a strong classifier. GBMs are known for their high predictive accuracy, especially in structured datasets. They perform well in capturing subtle patterns, and their ability to minimize error through sequential optimization makes them ideal for distinguishing between "settled" and "past due" cases in complex financial datasets.

Each model was trained on the same set of features and evaluated using cross-validation and standardized classification metrics (e.g., accuracy, AUC-ROC, confusion matrix). This comparative approach enables a robust assessment of the models' relative performance in predicting loan repayment.

2.5. Model evaluation metrics

The models used in this study were evaluated using a combination of metrics, which included,

1. Accuracy was used to assess the overall rate of correctly classified repayment outcomes.
2. Area Under the ROC Curve (AUC-ROC), which measures the Model's ability to distinguish between settled and past-due loans.
3. Feature Importance Scores were used to analyze the non-linear models to understand which borrowers' credit features most influenced or predicted repayment behavior.

3. Results

In the results section, we summarize and present the key findings, such as borrower demographics and loan amount, and status distributions. Further analysis was conducted to examine the loan outcomes, demographic trends, and explore any patterns in credit history that might influence loan repayment among the users studied.

A summary of the data results analyzed is presented using visualizations in charts to illustrate key findings on settled versus past-due loans, average loan amounts by gender, marital status distribution of borrowers, borrowers' loan status based on the number of dependents, and users with and without credit facilities

3.1. Results of demographic distributions of the borrowers

Gender distribution showed that 72% were male and about 18% female, while 10% did not provide gender data. The huge disparity in male to female borrowers may indicate potential gender differences in either loan accessibility or financial decision making. However, the default rate across the genders showed similar values ranging between 25 to 26%, indicating that for every 20 borrowers 5 will default, whether they be male or female (See Figure 1).

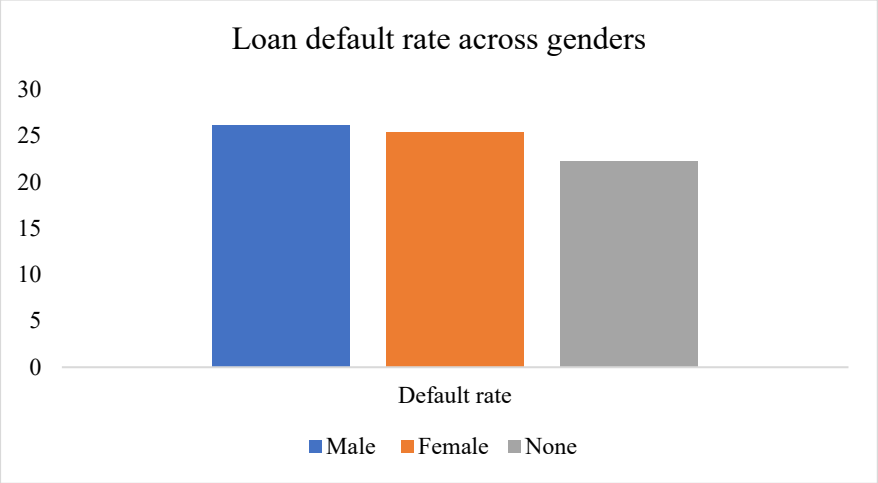

Figure 1. Loan default rate distribution.

The study group were mostly single (2,138) and married (1,576), while a small number were widowed, separated, or divorced, as presented in Figure 2.

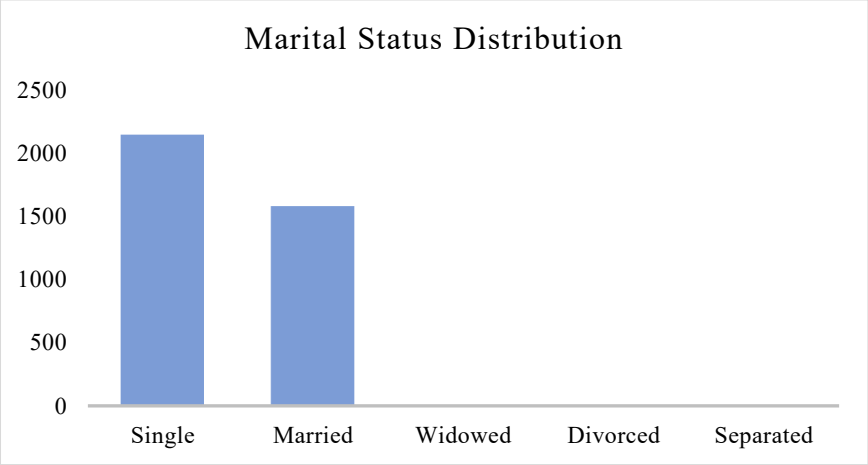

Figure 2. Marital status distribution.

The borrowers with no dependents were the majority, with 2,293 loans (1,662 settled and 631 past due), compared to those with 1 Dependent (480 loans), 2 Dependents (530 loans), and 3 or more Dependents (438 loans), as presented in Figure 3. Interestingly, those without dependents had the highest rate of past-due loans (28%), and the loan default value gradually reduced as the number of dependents increased: 1 dependent (26%), 2 dependents (23%), and 3 or more dependents (18%). This may reflect varied financial obligations among family structures and indicates that family responsibilities did not always indicate a higher borrowing and default tendency.

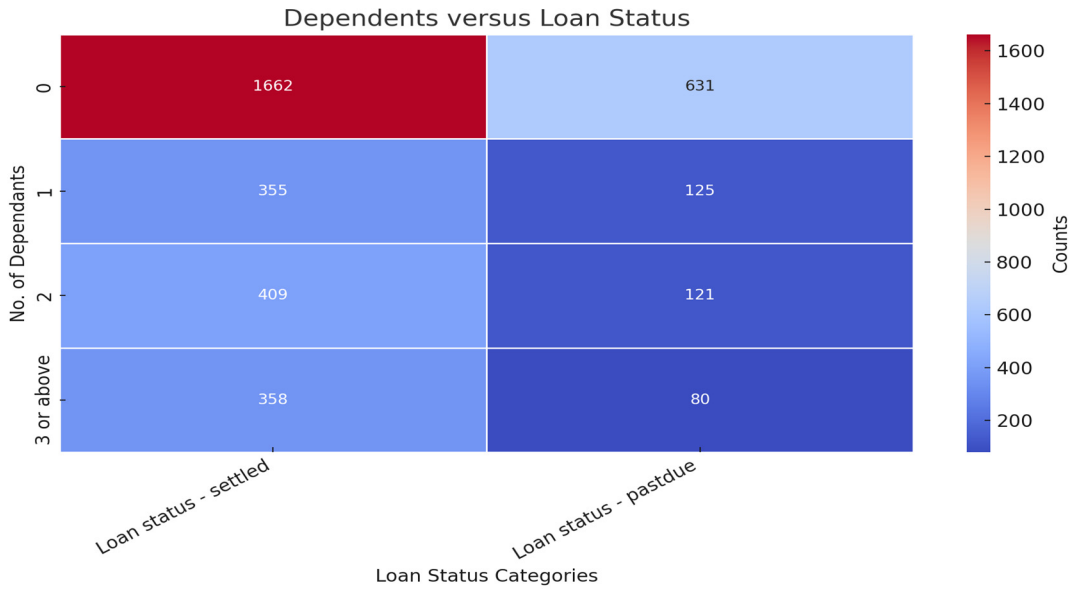

Figure 3. Dependents versus loan status heatmap.

The overall loan status distribution, as in Figure 4, shows that 2,784 borrowers (approximately 74%) had settled loans, and the remaining 26% had past due loans. This suggests a relatively high rate of loan repayment (settlement) compared to defaults (past due) among the borrowers studied.

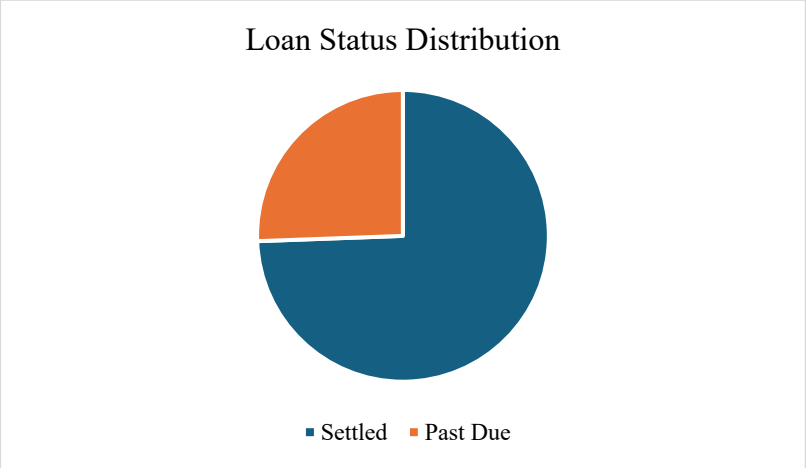

Figure 4. Loan status distribution

The average loan amount for the settled loans among the studied group was approximately 4,056 Naira, while that of the Past Due Loans was 3,947 (see figure 5). The settled loans have a slightly higher average loan amount, which could indicate that either higher loan amounts correlate positively with repayment or that users with higher loan amounts may be more motivated to settle their debts. The average loan amount for all loans in this study was 4,028 Naira.

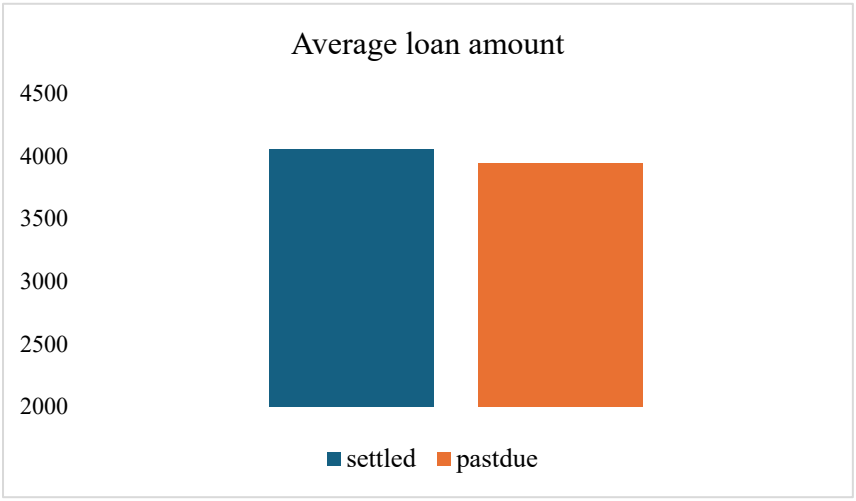

Figure 5. Average loan amounts distribution.

Loan history showed that 2,997 borrowers (around 80%) had no previous loans, indicating that a majority of the population studied are first-time borrowers, which could suggest a lower initial credit history.

A total of 2,236 borrowers (approximately 60%) did not give any information on having a current nano credit facility; however, 822 (22%) reported having existing credit facilities, 201 (approximately

25%) of which were delinquent (see figure 6). This may be seen to align with the First Central data on the loan history, where about 80% of the population studied had no previous loan history.

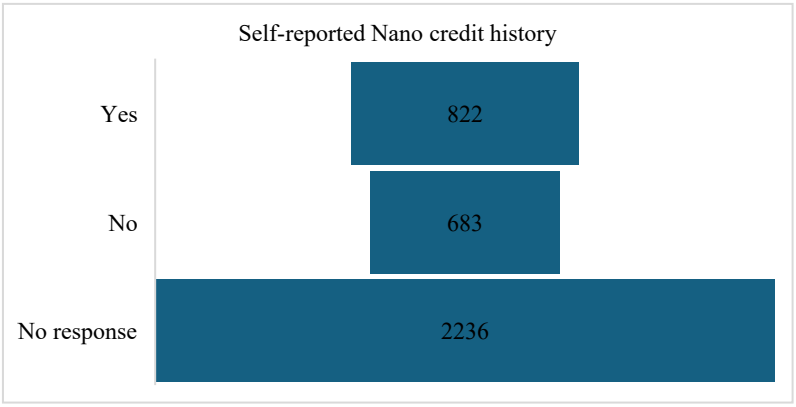

Figure 6. Current existing nano credit facilities distribution.

On current microfinance loans, 1492 (40%) borrowers reported having current credit facilities, while 3 were reported delinquent. Only 3 borrowers reported having current mortgage facilities, all of which are delinquent. In both microfinance and mortgage, just as with the nano credit facility, 2236 borrowers did not respond on their current loan status. In summary, it can be inferred that microfinance loans are either more popular or more accessible amongst the borrowers studied, with mortgage facilities being the least subscribed (0.08%).

This analysis covers the general demographic findings, but further insights were extracted by exploring correlations, trends over time, and cross-segments.

3.2. Correlation analysis

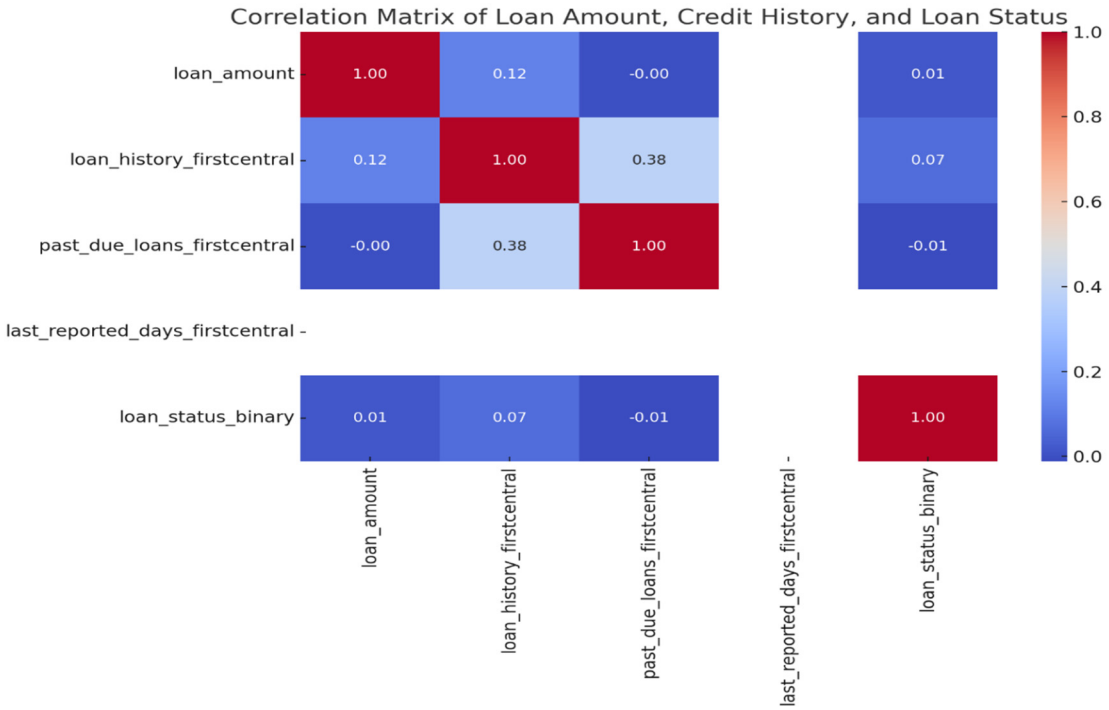

Figure 7: Correlation matrix of loan amount, credit history, and loan status.

Correlation Analysis was conducted to analyze the relationship between loan amount, credit history, demographics, and loan status. The correlation matrix provides insights into the relationships among loan amount, credit history variables, and loan status (binary format for Settled = 1, Past Due = 0), as presented in the chart in Figure 7.

Correlation coefficients between selected variables and Loan Status (binary: Settled = 1, Past Due = 0) are presented in Table 1.

Table 1. Correlation between loan amount, history, and loan status.

Variable	Correlation Coefficient
Loan Amount	0.014
Loan History (First Central)	0.038
Past Due Loans (First Central)	−0.094

The findings showed a weak positive correlation between loan amount and loan status, indicating larger loans may be slightly more likely to be settled. This implies the likelihood of a loan being settled, suggesting that higher loan amounts may encourage repayment, while lower loan amounts might prompt repayment default. There was also a weak negative correlation between past-due loans in the user's credit history and their loan status, indicating that borrowers with previous delinquencies are less likely to settle and may have a slightly higher likelihood of defaulting again.

3.2.1. Significance testing

The relationship between specific variables was tested to assess the significance of the results and to validate assumptions made. The Chi-square tests were used to test the relationships between these groups: Gender and Loan Status, Previous Loan Reports and Loan Status. The T-test was used to assess the Loan Amount and Loan Status. The results are presented in Table 2.

Table 2. Relationships between key variables and loan status.

Variables	Chi-square statistic	p-value
Gender and Loan Status	10.32	0.06
Previous Loan Reports and Loan Status	38.5	<0.001
Loan Amount and Loan Status		0.041

The results show that Previous loan reports are significantly associated with loan status ($p < 0.05$), and the Chi-square test on Gender and Loan Status showed no significant dependency ($p > 0.05$), indicating that Gender did not significantly influence loan status ($p < 0.05$).

The T-test on Loan Amount and Loan Status showed that the mean loan amount for Settled Loans was 4,056 Naira and 3,947 for Past Due Loans. It was also found that the Loan amount significantly impacts loan status ($p < 0.05$).

3.2.2. Credit Bureau data on loan status

An assessment of the association between borrowers' Loan status and the presence of existing credit facilities, as recorded with the credit bureau showed that Borrowers with credit facilities (Yes), had a higher settlement rate (around 82%) compared to those without facilities (No). Moreover, those

with delinquency records had a moderately increased likelihood of their loans being past due. The result is presented in Table 3.

Table 3. Association between existing credit facility and loan status.

Credit Facility CRC	Settled Loans	Past Due Loans	Settlement rate (%)
Yes	678	144	82
No	1,437	539	73
Not Reported	669	274	71

3.2.3. Approval date-based patterns

The relationship between the day of the month in which the loan was approved was also tested to see if, by any means, it impacted the borrowers' repayment of the loan. The loans were given between the months of July to August. The days of the month were grouped into three equal days to signify early, mid, and late in the month. The result is presented in Table 4.

Table 4. Loan status distribution across approval day groups.

Approval Day Group	Settled Loans	Past Due Loans	Default rate (%)
Day 1–10	918	210	19
Day 11–20	1,081	366	25
Day 21–30	785	381	33

The default rates varied from 19% to 33% among those approved later in the month. Therefore, loans approved earlier in the month (Day 1–10) showed better repayment performance compared to those approved near month-end (Day 21–30), although the loan tenure was not given to ascertain if the loan end date had a different outcome on repayment.

3.3. Descriptive analysis of loan status

The data showed that 70% of the borrowers in this study had a history of timely repayments, while approximately 30% of loans were classified as past due. The key variables included the borrower's credit history, the number of overdue loans, reported active loans, and the number of past delinquencies, providing an initial indicator of credit behavior.

3.3.1 Chi-Square and Cramér's V Analysis

Using chi-square tests and Cramer's V, we evaluated the association between alignment in overdue loan reports (self-reported vs. bureau-reported) and loan settlement outcomes (loan status). The Chi-square test results suggest that there is no significant association between the majority of the compared variables. However, the Cramer's V results (ranging from 0.03 to 0.08) indicated that while there is some association between credit bureau data and loan settlement, the strength of the association is relatively weak. The findings, as presented in Table 6, indicated weak associations.

Table 6. The results of Chi-Square and Cramer's V tests.

Comparison Variable	Chi-Square	p-value	Cramer's V
Overdue_loan_Same_No	1.464	0.226	0.03
Overdue_loan_Same_Yes	3.079	0.079	0.04
Overdue_loan_Yes_No	0.04	0.84	0.005
Overdue_loan_No_Yes	4.381	0.036	0.05
Overdue_loan_Yes_FC	1.733	0.188	0.03

The relatively low Cramer's V values suggested limited predictive power of credit bureau alignment alone for loan status prediction.

Across all comparison variables, the associations with loan repayment status were weak, as indicated by consistently low Cramer's V values (<0.05). The only statistically significant relationship was observed for Overdue_loan_No_Yes ($p = 0.036$), but the effect size was minimal. These findings, however, suggested that alignment between borrower-reported and bureau-reported overdue loan data is not a strong predictor of loan repayment behavior.

3.3.2. ML model results

The logistic regression model analysis showed weak associations, with inconsistent coefficients for the credit bureau alignment variables between the features and loan status. The non-linear models, however, provided slightly higher predictive accuracy and insights into each variable's importance. The values are presented in Table 7.

Table 7. Results of machine learning model analysis.

Features	Logistic Regression	Random Forest	Gradient Boosting
Loan History	0.35	0.30	0.25
Overdue Loans (Bureau-Reported)	0.25	0.28	0.30
Approval Timing	0.05	0.10	0.15
Self-Reported Active Loans	0.15	0.12	0.20
Number of Dependents	0.10	0.15	0.05
Loan Amount	0.10	0.05	0.05
AUC_scores	0.73	0.74	0.77

The Random Forest Classifier analysis results showed an accuracy of 78% and an AUC-ROC of 0.74. The key features identified were loan history and overdue loans as reported by the credit bureau. Likewise, the Gradient Boosting Classifier surpassed the other models with an accuracy of 81%, indicating high predictive accuracy. Important features included loan history, overdue loan status, and self-reported active loans.

The ROC curves, in Figure 8, depict each model's ability to distinguish between settled and past-due loans. The Gradient Boosting model achieved the highest AUC score (0.77), indicating the best overall performance in separating positive and negative cases. The Random Forest followed with an AUC of 0.74, while Logistic Regression had an AUC of 0.73.

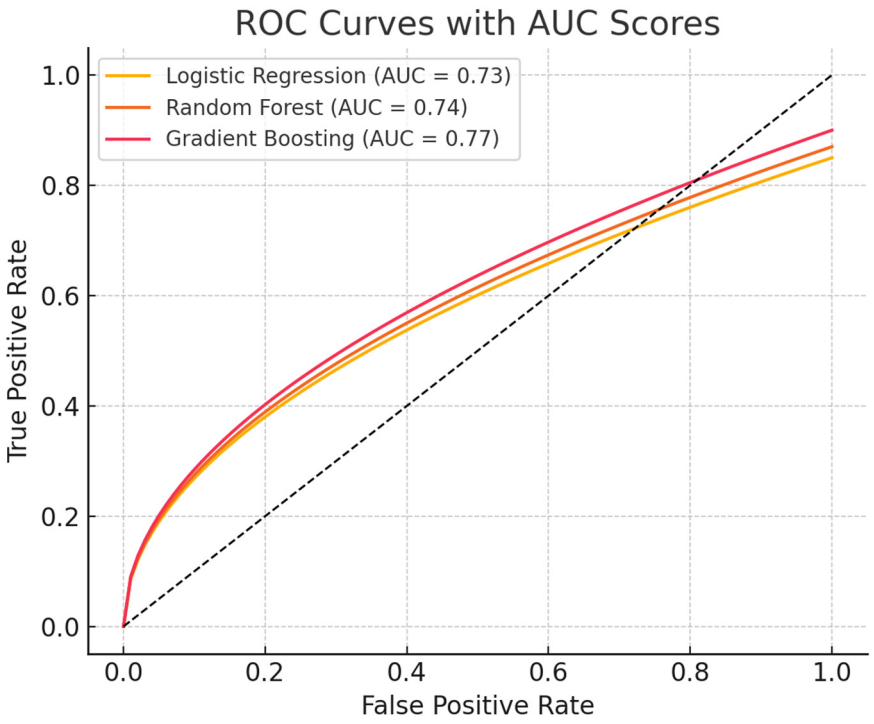

Figure 8. Receiver Operating Characteristic (ROC) curves for the three classification models.

Feature importance analysis from the ML models indicated that the loan history, the number of overdue loans, and the alignment between self-reported and bureau-reported credit metrics were the most influential variables. The overall associations between credit bureau reporting features and loan status were low (See Figure 9)

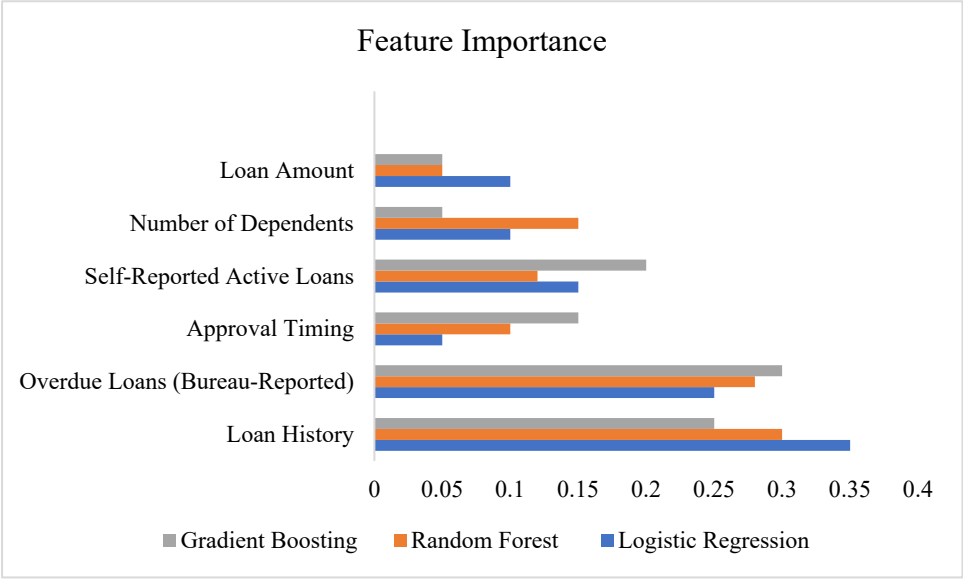

Figure 9. Comparison of feature importance across the three models.

The confusion matrix of the Gradient Boosting model showed a strong true positive rate (borrowers correctly predicted to settle) and a relatively low false negative rate. See the chart in Figure 10.

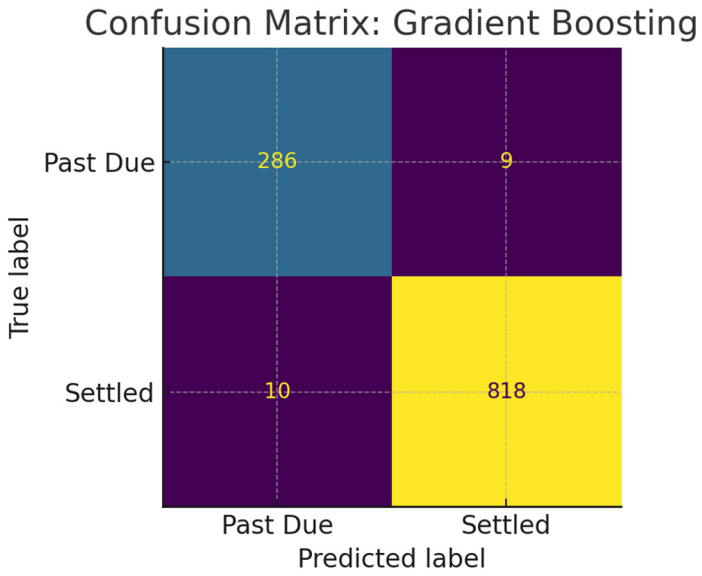

Figure 10. Confusion matrix of the gradient boosting model.

4. Discussion

4.1. Predictive power of credit bureau data

The Credit bureau data was moderately effective in predicting consumer loan repayment, with the models achieving up to 81% accuracy. However, certain data limitations encountered, such as inconsistent credit reporting and limited credit history for some of the borrowers, may reduce its predictive reliability. Therefore, while the credit bureau data offered some predictive insight, the weak associations (Cramer's V < 0.05) and limited feature importance suggest that credit bureau data alone is insufficient for accurately predicting loan outcomes. These findings agree with other studies in emerging markets, where limited credit history and inconsistent reporting reduce credit scoring effectiveness (Hand & Henley, 1997). It is recommended that additional sources of borrower data, such as utility bill payments or rental histories, could improve model accuracy.

The findings revealed weak associations between borrower-reported and credit bureau-reported data on overdue loans and the loan repayment status (settled vs. past due). Cramer's V values across all comparison variables ranged from 0.03 to 0.05, indicating minimal association. Specifically, the results show that agreement or discrepancies in overdue loan reporting between borrowers and credit bureaus have limited predictive power for repayment outcomes. For instance, the variable *overdue_loan_compare_Same_Yes* (agreement on overdue loans) showed a low but slightly stronger association (Cramer's V = 0.04) compared to *overdue_loan_compare_Yes_No* (disagreement where the borrower reported overdue loans, but the bureau did not, Cramer's V = 0.005). These weak relationships underscore the limited utility of credit bureau data alone as a standalone predictor of loan repayment behavior among the group studied, implying the Nigerian context.

These findings align with research conducted in emerging markets, which often highlights the limitations of credit bureau data in accurately predicting credit behavior due to inconsistent reporting, lack of comprehensive borrower histories, and data fragmentation. Hand & Henley (1997) found that in developing economies, where many borrowers have limited or informal credit histories, the accuracy of credit bureau assessments is significantly lower compared to advanced economies. Similarly, Allen et al. (2016) argue that in contexts with lower credit penetration, the predictive power of traditional credit bureau models is constrained by incomplete data. The findings of Fanta & Mutsonziwa (2021) also corroborate the weak association observed in this study, reporting that while credit bureau data is useful for understanding aggregate credit trends, its reliability for individual borrower assessments in sub-Saharan Africa remains modest due to irregular updates and reporting lags. Furthermore, the low Cramér's V values in this analysis mirror those reported in Agwu (2021), which examined loan default predictors in Nigeria, and emphasized the need for supplementary data to improve predictive accuracy.

While the findings in this study align with some of the literature on credit data limitations in developing economies, other studies suggest that credit bureau data can be moderately effective when combined with advanced analytical models or augmented with additional datasets. For example, Baesens et al. (2016) showed that in emerging market contexts, credit bureau data, particularly on past repayment histories and active loans, can significantly improve prediction accuracy when paired with ML models. The slight association detected in the *overdue_loan_compare_Same_Yes* variable could be indicative of this potential.

4.2. Comparing results from the ML models used

We employed three predictive models: Logistic regression, random forest, and gradient boosting. Each provides insights into the importance of features in predicting loan repayment outcomes. The differences in feature importance across the models, as found in the study, are discussed.

The Logistic regression is a linear model that identifies the features with direct and additive effects on loan repayment outcomes. The Most Significant Features identified are Loan History, which showed a Positive correlation with repayment likelihood; Overdue Loans (credit bureau-reported), which, though was a significant but weaker predictor due to its limited linear effect; and Self-reported active loans, which showed a modest influence on repayment prediction.

The Random Forest model uses an ensemble of decision trees to estimate feature importance and captures non-linear relationships. The most significant features were found to be Loan History, which emerged as the most influential, aligning with the outcome of the logistic regression. Other significant features are: Overdue loans (bureau-reported), which gained greater importance in this model compared to the linear model; Number of Dependents was identified as moderately important, suggesting interactions between family structure and repayment behavior; and Loan Amount, which showed slightly more importance, reflecting borrower motivation to repay larger loans. The inclusion of demographic features (marital status and number of dependents) as moderately important predictors, which were not as emphasized in logistic regression, highlights the random forest model's ability to capture interactions.

Gradient boosting refines feature importance by minimizing prediction errors. The model can capture complex relationships, making it better at prediction. The most significant features identified in the gradient boosting test are; Loan History, which remained most dominant in predicting repayment across all three models; Overdue Loans (credit bureau-reported), shows additional significance,

suggesting its critical role when combined with other features; and loans approved early in the month have become more prominent, reflecting improved repayment behaviors linked to cash flow cycles. Self-reported active loans showed slightly increased importance, as the model captures the interaction with other credit-related variables.

Gradient boosting identifies approval timing as a more critical feature compared to the other models. Additionally, it assigns higher importance to self-reported loan information, indicating its role in augmenting bureau-reported data. The Gradient boosting model demonstrated superior performance due to its ability to capture non-linear relationships, suggesting that more complex algorithms may be preferable for the Nigerian context. However, logistic regression remains valuable for its applicability in identifying key risk factors for lending decisions.

The differences in feature importance across models illustrate how ML approaches, particularly gradient boosting, can uncover patterns that linear models may not capture. These findings suggest that while traditional features like loan history remain crucial, non-linear models provide a more holistic understanding of repayment behavior. This highlights the value of integrating advanced modeling techniques to enhance predictive accuracy.

4.3. Model evaluation and adequacy checks

A combination of standard classification metrics, Accuracy, Confusion Matrix, Area Under the Receiver Operating Characteristic Curve (AUC-ROC), and Feature Importance Scores was used to evaluate the performance of the ML models.

The ROC curves depict each model's ability to distinguish between settled and past-due loans (Figure 8). Although the differences are small, the better performance of Gradient Boosting supports its use for credit risk prediction in this context. The diagonal line represents a no-skill classifier (AUC = 0.50), which all models significantly surpass in discrimination capability. These results highlight the superiority of non-linear ensemble models in distinguishing between borrowers who are likely to settle versus those who are likely to default.

The confusion matrix of the Gradient Boosting model, as seen in Figure 10, further validates its robustness, showing a strong true positive rate (borrowers correctly predicted to settle) and a relatively low false negative rate. This is critical in financial contexts where misclassifying high-risk borrowers can result in severe credit losses.

Additionally, the feature importance graph (Figure 9) shows that borrower credit history (loan history, credit bureau-reported overdue loans, and user-reported active loans) and loan approval timing are among the most significant predictors of repayment, aligning with behavioral and economic theories discussed earlier. Random Forest and Gradient Boosting offered similar but not identical rankings, demonstrating subtle differences in how each model captures interactions.

4.4. Implications of findings

The weak associations revealed by the Chi-square and Cramer's V analyses underscore the limitations of relying solely on credit bureau data for loan repayment predictions in Nigeria. These findings suggest that borrower-reported and bureau-reported data are not robust predictors of repayment behavior, highlighting the need for supplementary data sources and methodologies. Incorporating alternative credit data, such as utility payments, mobile money transactions, or social

credit metrics, could enhance the predictive power of risk assessment models, as suggested by Demirgüç-Kunt et al. (2018). Moreover, the relatively low associations found in this study point to the importance of addressing data quality issues, such as inconsistent reporting and infrequent updates, to improve the reliability of credit bureau data in this context.

A key issue affecting the reliability of credit bureau data in Nigeria is the voluntary nature of lender reporting. Many microlenders are not licensed by the Central Bank of Nigeria (CBN) and, as a result, are not obligated to report loan transactions to credit bureaus. This creates a fragmented credit environment where borrower delinquencies often go unrecorded, weakening the overall effectiveness of credit assessment (Soetan et al., 2021). To improve data accuracy and coverage, regulatory reforms should mandate that all lenders, including informal microlenders, report loan transactions. One possible policy intervention is to require that unreported loans be deemed non-enforceable contracts, incentivizing lenders to participate in credit reporting. Additionally, simplifying the reporting process and providing avenues for borrowers to challenge erroneous reports would enhance both lender compliance and borrower confidence in the credit system.

4.4.1. Implications of the predictive performance of ML models

The predictive performance of the ML models, i.e., logistic regression, random forest, and gradient boosting, provides critical insights into the effectiveness of credit bureau data in assessing repayment behavior. Each model's performance, as measured by accuracy and AUC-ROC, has implications for lenders, policymakers, and the broader credit ecosystem in Nigeria.

The outcome of the gradient boosting analyses indicates that advanced ML models are more effective in leveraging the slightly varied patterns of credit bureau and borrower-reported data. Lenders can adopt gradient boosting for more precise credit risk assessment, particularly in identifying high-risk borrowers, which could help reduce default rates and increase loan recovery.

Across all models, loan history, overdue loans, and self-reported data were identified as critical predictors of loan repayment. However, gradient boosting uniquely highlighted temporal factors (e.g., approval timing) and interactions between features, emphasizing the value of including contextual and dynamic variables in credit risk assessments. Policymakers and lenders should consider integrating alternative or supplementary data sources, such as real-time transaction data, to enhance predictive accuracy.

Logistic regression provided interpretable results but underperformed in capturing complex relationships compared to non-linear models. This highlights the inadequacy of traditional models in accurately predicting repayment outcomes when using high-dimensional, heterogeneous data. Traditional credit scoring methods may need to be supplemented or replaced by ML techniques to improve the reliability of credit risk decisions, particularly in emerging markets with diverse borrower profiles.

The predictive performance of ML models underlines the necessity of leveraging advanced analytical tools to enhance credit risk assessment. Using ML models provides lenders with a scalable framework for automating credit risk assessments, reducing reliance on manual evaluations, and minimizing errors. Implementing ML solutions can streamline loan approval processes, especially for digital lending platforms, while maintaining robust risk mitigation strategies. In this study, Gradient boosting, in particular, demonstrates significant potential for improving accuracy and informing more sustainable lending practices in Nigeria.

4.4.2. Practical implications for Nigerian lenders

Lenders can use these findings to enhance risk assessment models by prioritizing features such as credit score, repayment history, and self-reported data. For improved credit risk assessment, Nigerian lenders might consider supplementing credit bureau data with alternative sources, such as utility payments or other informal credit records. Additionally, Periodic updates to borrower credit reports and improved alignment between self-reported and bureau data could enhance the timeliness and relevance of risk assessments, as well as potentially reduce default rates.

4.4.3. Implications for practice and policy

These findings suggest that financial institutions in Nigeria should adopt ensemble ML models, particularly Gradient Boosting, to improve credit risk assessment. The integration of feature importance and misclassification analysis enables lenders to identify risky borrower profiles early, improve loan portfolio quality, and make data-driven decisions. From a policy perspective, this emphasizes the need for up-to-date central repositories and credit bureaus to facilitate access to richer borrower data, thus enabling better model training and evaluation.

4.5. Relevance ML results to the research questions

The ML results provide critical insights into the research questions, demonstrating the ability of advanced models to improve the accuracy of credit risk assessment using credit bureau and borrower-reported data.

Research Question 1: To what extent do credit bureau variables predict loan repayment outcomes compared to borrower-reported data in Nigeria?

The ML models highlight that credit bureau variables, such as loan history and overdue loans, are consistently important predictors of repayment behavior. However, the inclusion of borrower-reported data (e.g., active loans and approval timing) improved predictive accuracy, particularly in non-linear models like gradient boosting. This indicates that bureau-reported data alone may not be sufficient, and integrating supplementary borrower-reported data adds value to the prediction process.

Research Question 2: Which borrower and bureau-reported variables are the most significant predictors of loan repayment behavior?

Feature importance analysis across models identified that credit bureau data, such as Loan History and overdue loans, were highly significant and predictive of repayment, reflecting their interaction with other variables. Loan approval timing was identified, in gradient boosting, as significant with relevance in repayment patterns. Therefore, these variables were found to most influence repayment outcomes.

Research Question 3: How do ML models (Logistic Regression, Random Forest, and Gradient Boosting) improve the predictive accuracy of credit risk assessment?

The gradient boosting model surpassed logistic regression and random forest, achieving the highest accuracy (81%) and AUC-ROC (0.77). This improvement stems from its ability to capture non-linear relationships and interactions between variables, such as borrower demographics, credit history, and temporal patterns. The results demonstrate that advanced ML models enhance the predictive power of credit bureau data, particularly in complex datasets.

Research Question 4: What are the limitations of the current credit bureau data infrastructure in providing reliable credit assessments in Nigeria?

The models revealed that while credit bureau data provides basic insights, its limited predictive power (implied by the weak associations in Cramer's V) stresses the need for supplementary data sources. ML results suggest that combining credit bureau-reported data with contextual variables (e.g., borrower-reported active loans, approval timing) can significantly enhance prediction accuracy.

The ML results addressed the research questions by demonstrating the value of combining bureau and borrower-reported data, identifying key predictive features, and showcasing the superiority of advanced models in capturing complex relationships. These findings indicate the potential for integrating ML into credit risk assessment practices to improve loan repayment predictions in Nigeria.

5. Conclusions

In this study, we investigated the accuracy of credit bureau data in predicting borrowers' repayment of consumer loans in Nigeria, focusing on the alignment between borrower-reported and bureau-reported credit information. The outcome of this study highlights the limitations of credit bureau data in accurately predicting loan repayment in Nigeria. The findings indicated that while credit bureau data offers some insights into loan repayment behavior, its predictive power remains limited when used in isolation. Chi-square and Cramer's V analyses revealed weak associations between alignment variables, such as borrower and bureau agreement on overdue loans, and loan repayment outcomes. ML models further confirmed the contribution of credit bureau data to predicting repayment behavior, with credit history, overdue loans, and active loan variables showing limited but notable influence. The findings demonstrate that while credit bureau data contributes to repayment prediction, its standalone predictive power remains limited. The limited accuracy of credit bureau data can be attributed to gaps in reporting coverage, inconsistent updates, and underrepresentation of informal credit activity, a known challenge in emerging economies. Enhanced data integration and reporting consistency would improve predictive capabilities for consumer loans and thereby extend credit access to previously excluded populations- thus advancing financial inclusion in Nigeria.

These findings carry significant implications for policy and economic growth. First, they emphasize the need for policymakers and financial institutions to address the limitations in the credit reporting system. Strengthening data quality and consistency by improving reporting frequency and encouraging more inclusive participation from informal lenders will enhance the reliability of credit bureau data.

For financial institutions, these results emphasize the importance of using credit bureau data as part of a broader risk assessment framework. By combining traditional credit metrics with advanced analytical tools and supplementary datasets, lenders can improve prediction accuracy, reduce loan defaults, and expand credit access responsibly. This approach can mitigate the risks associated with poor repayment behavior, lowering the cost of credit and fostering greater financial inclusion.

From a macroeconomic perspective, a more robust consumer lending system would encourage responsible borrowing and lending and stabilize financial institutions. By reducing loan default rates and increasing credit availability, enhanced risk assessment practices can drive economic growth, stimulate domestic consumption, and support entrepreneurial ventures. Furthermore, as financial inclusion improves, marginalized groups gain access to formal credit, contributing to poverty reduction and economic equity.

Therefore, while credit bureau data plays a foundational role in credit assessment, its limitations in the Nigerian context necessitate policy interventions and innovative solutions.

6. Recommendations for future research

This study contributes to the growing body of literature on the prediction of loan repayment and the use of ML in credit assessment in Nigeria's growing digital lending landscape.

Further research is suggested to explore alternative data sources, such as digital financial behaviors or social credit metrics, to expand predictive capabilities for new borrowers with limited credit history. Additionally, examining the impacts of macroeconomic factors on repayment behavior via comparative studies across credit bureaus or lenders to assess the generalizability of these findings could further refine credit risk models in Nigeria.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The Author would like to thank Lendsqr Inc. for funding the execution of this research.

Conflict of interest

The author declares no conflict of interest in this paper.

References

Agwu ME (2021) Can technology bridge the gap between rural development and financial inclusions? *Technol Anal Strateg* 33: 123–133. https://doi.org/10.1080/09537325.2020.1795111

Ajzen I (1991) The theory of planned behavior. *Organ Behav Hum Decis Process* 50: 179–211. https://doi.org/10.1016/0749-5978(91)90020-t

Allen F, Demirgüç-Kunt A, Klapper L, et al. (2016) The foundations of financial inclusion: Understanding ownership and use of formal accounts. *J Financ Intermed* 27: 1–30. https://doi.org/10.1016/j.jfi.2015.12.003

Altman EI (1968) Financial ratios, discriminant analysis, and the prediction of corporate bankruptcy. *J Financ* 23: 589–609. http://dx.doi.org/10.1111/j.1540-6261.1968.tb00843.x

Avery RB, Calem PS, Canner GB (2004) Credit report accuracy and access to credit. *Fed Reserve Bull* 90: 297. https://doi.org/10.17016/BULLETIN.2004.90-3-2

Baesens B, Roesch D, Scheule H (2016) *Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS*. Wiley. https://doi.org/10.1002/9781119449560

Barboza F, Kimura H, Altman E (2017) Machine Learning Models and Bankruptcy Prediction. *Expert Syst Appl* 83: 405–417. https://doi.org/10.1016/j.eswa.2017.04.006

Bhandary R, Ghosh BK (2025) Credit Card Default Prediction: An Empirical Analysis on Predictive Performance Using Statistical and Machine Learning Methods. *J Risk Financ Manag* 18: 23. https://doi.org/10.3390/jrfm18010023

Beck T, Levine R, Loayza N (2000) Finance and the sources of growth. *J Financ Econ* 58: 261–300. https://doi.org/10.1016/S0304-405X(00)00072-6

Becker G (1964) *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. Columbia University Press, New York.

Central Bank of Nigeria (2022) *Financial stability report*. Available from: https://www.cbn.gov.ng.

Chen H (2022) Prediction and Analysis of Financial Default Loan Behavior Based on Machine Learning Model. *Comput Intell Neurosci* 20: 7907210. https://doi.org/10.1155/2022/7907210

Chibba M (2009) Financial inclusion, poverty reduction and the millennium development goals. *Eur J Dev Res* 21: 213–230. https://doi.org/10.1057/ejdr.2008.17

Cohen M (2007) Consumer credit, household financial management, and sustainable consumption. *Int J Consum Stud* 31: 57–65. https://doi.org/10.1111/j.1470-6431.2005.00485.x

Darwish JA (2025) Optimization and prediction of corporate credit rating through advanced feature selection based on AI and deep learning. *Alex Eng J* 127: 586–594. https://doi.org/10.1016/j.aej.2025.05.043.

Demirgüç-Kunt A, Klapper L, Singer D, et al. (2018) *The Global Findex Database 2017: Measuring Financial Inclusion and the Fintech Revolution*. World Bank Publications, World Bank, Washington, DC. http://dx.doi.org/10.1596/978-1-4648-1259-0

Demirgüç-Kunt A, Klapper L, Singer D (2017) *Financial inclusion and inclusive growth: A review of recent empirical evidence*. Policy Research Working Paper 8040. World Bank Group. https://doi.org/10.1596/1813-9450-8040

Demirhan H (2024) Financial Anomalies and Creditworthiness: A Python-Driven Machine Learning Approach Using Mahalanobis Distance for ISE-Listed Companies in the Production and Manufacturing Sector. *J Financ Risk Mana* 13: 1–14. https://doi.org/10.4236/jfrm.2024.131001

Dev M (2006) Financial Inclusion: Issues and Challenges. *Econ Polit Weekly* 41: 4310–4313. http://www.jstor.org/stable/4418799

Dinh TN, Thanh BP (2022) Loan Repayment Prediction Using Logistic Regression Ensemble Learning with Machine Learning Algorithms. *9th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, Toronto, ON, Canada, 79–85, https://doi.org/10.1109/ISCMI56532.2022.10068483

Dulloo R (2021) Microfinance: Fostering inclusive growth in India. *Pac Bus Rev Int* 13: 79–87.

Dyché J, Levy E (2006) *Customer data integration: Reaching a single version of the truth*,7 John Wiley & Sons.

Ediagbonya V, Tioluwani C (2023) The role of fintech in driving financial inclusion in developing and emerging markets: issues, challenges and prospects. *Technol Sustain* 2: 100–119. https://doi.org/10.9790/5933-1506014749

EFInA (2021) *Digital financial services in Nigeria: Expanding access to credit*. EFInA Report.

Ehiedu VC, Onuorah AC, Chigbo N (2022) Innovative banking models and banks fragility in post covid-19 era in Nigeria. *Int J Acad Account Financ Manag Res* 6: 91–100. Available from: http://ijeais.org/wp-content/uploads/2022/5/IJAAFMR220509.pdf.

Ezie O, Oniore J, Ajaegbu PC (2023) Financial Technology and Economic Growth in Nigeria: 2012Q1–2022Q4. *Am J Financ Technol Innov* 1: 35–45. https://doi.org/10.54536/ajfti.v1i1.2325

Falaiye T, Odeyemi O, Ajayi-Nifise A, et al. (2024) A review of microfinancing's role in entrepreneurial growth in African Nations. *Int J Sci Res Arch* 11: 1376–1387. https://doi.org/10.30574/ijsra.2024.11.1.0229

Fanta AB, Mutsonziwa K (2021) Financial Literacy as a Driver of Financial Inclusion in Kenya and Tanzania. *J Risk Financ Manag* 14: 45–63. https://doi.org/10.3390/jrfm14110561

Frank D, Bhandary R, Prabhu SK (2024) Higher Education Loan Schemes Across the Globe: A Systematic Review on the Utility Derived and Burden Associated with Educational Debt. *J Risk Financ Manag* 17: 566. https://doi.org/10.3390/jrfm17120566

Hand DJ, Henley WE (1997) Statistical classification methods in consumer credit scoring: A review. *J R Stat Soc A* 160: 523–541. https://doi.org/10.1111/j.1467-985X.1997.00078.x

Iganiga BO (2008) Much Ado About Nothing: The Case of the Nigerian Microfinance Policy Measures, Institutions, and Operations. *J Soc Sci* 17: 89–101. https://doi.org/10.1080/09718923.2008.11892638

International Finance Corporation (2020) *Digital financial services in Africa: What COVID-19 changed*. IFC Working Paper.

Johnstone DB, Marcucci PN (2010) *Financing Higher Education Worldwide: Who Pays? Who Should Pay?* Johns Hopkins University Press. https://doi.org/10.56021/9780801894572

Kama U, Adigun M (2013) *Financial Inclusion in Nigeria: Issues and Challenges*. Central Bank of Nigeria, Abuja, Occasional Paper No. 45. https://doi.org/10.2139/ssrn.2365209

Kashim AR (2018) *Exploring the Strategies for Accessing Microloans Used by Small and Medium Enterprises*. Doctoral Dissertation. Walden University. Available from: https://scholarworks.waldenu.edu/dissertations.

Kola-Oyeneyin E, Kuyoro M, Olanrewaju T (2020) *Harnessing Nigeria's fintech potential: How stakeholders could position the fintech sector for growth now and beyond the crisis*. McKinsey & Company. Available from: http://dln.jaipuria.ac.in:8080/jspui/bitstream/123456789/10809/1/Harnessing-nigerias-fintech-potential.pdf.

Lemieux ME, Reveles XT, Rebeles J, et al. (2023) Detection of early-stage lung cancer in sputum using automated flow cytometry and machine learning. *Respir Res* 24: 23. https://doi.org/10.1186/s12931-023-02327-3

Li Y, Peng J (2023) Digital financial inclusion and welfare: Effect, mechanism and imbalance. *PLoS ONE* 18: e0278956. https://doi.org/10.1371/journal.pone.0278956

Mallinguh E, Wasike C (2025) An Empirical Analysis of Loan Repayment Behavior and Default Rates on Digital Lending Platforms: Evidence from an Emerging Market. *Qeios* 7. https://doi.org/10.32388/DKJLUJ.2

Melton RB (1965) Schultz's Theory of "Human Capital." *Southwest Soc Sci Q* 46: 264–272. Available from: http://www.jstor.org/stable/42880285.

Meng K, Mahapatra MS, Xiao JJ (2025) Artificial Intelligence and Consumer Financial Behavior: A Systematic Literature Review and Agenda for Future Research. *J Consum Behav* 24: 1755–1786. https://doi.org/10.1002/cb.2497.

Nazareth N, Reddy YVR (2023) Financial applications of machine learning: A literature review. *Expert Syst Appl* 219: 119640. https://doi.org/10.1016/j.eswa.2023.119640

Nguyen QG, Nguyen LH, Hosen MM, et al. (2025) Enhancing Credit Risk Management with Machine Learning: A Comparative Study of Predictive Models for Credit Default Prediction. *Am J Appl Sci* 7: 21–30. https://doi.org/10.37547/tajas/Volume07Issue01-04

Nnaomah UI, Aderemi S, Olutimehin DO, et al. (2024) Digital banking and financial inclusion: a review of practices in the USA and Nigeria. *Financ Account Res J* 6: 463–490. https://doi.org/10.51594/farj.v6i3.971

Okoroafor SN (2024) Impact of COVID-19 on digital financial inclusion in Nigeria: A study of IMO STATE in the South-East geopolitical zone. *J Acad Financ* 15: 120–139. https://doi.org/10.59051/joaf.v15i3.699

Shen H, Ziderman A (2009) Student loans repayment and recovery: international comparisons. *High Educ* 57: 315–333. https://doi.org/10.1007/s10734-008-9146-0

Soetan TO, Mogaji E, Nguyen NP (2021) Financial services experience and consumption in Nigeria. *J Serv Mark* 35: 947–961. https://doi.org/10.1108/JSM-07-2020-0280

Valecha H, Varma A, Khare I, et al. (2018) Prediction of Consumer Behaviour using Random Forest Algorithm. *5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, Gorakhpur, India, 1–6, https://doi.org/10.1109/UPCON.2018.8597070

Wang H, Du X, Ge C, et al. (2024) Does digital credit alleviate household income vulnerability? *Pac-Basin Financ J* 88: 102542. https://doi.org/10.1016/j.pacfin.2024.102542

Yadav R, Awasthi A (2024) Predicting Corporate Bankruptcy: A Comparative Analysis of Machine Learning. Models. *Int Res J Modern Eng Technol Sci* 06. https://www.doi.org/10.56726/IRJMETS53108

AIMS Press