Predicting Loan Defaults Using Big Data Analytics and Machine Learning

UCHENNA EMMANUEL EVANS-ANORUO
Department of Applied Statistics and Operations Research, Bowling Green State University, USA

Abstract- ***This research focuses on predicting loan defaults using big data analytics machine learning models applied to a comprehensive loan dataset. The analysis is conducted using R statistical software, enabling data-driven insights for enhanced credit risk management. Three algorithms Random Forest, XGBoost, and Naïve Bayes are implemented to determine the most effective predictive model and identify key risk factors. The study utilized a comprehensive loan dataset sourced from Kaggle which comprised of 148,670 individual loan records, each characterized by 34 features spanning borrower demographics, financial characteristics, and loan specifications. Feature selection followed a multi-stage process designed to optimize model performance while maintaining interpretability. The balanced dataset (73,278) was partitioned using stratified random sampling to ensure representative class distribution. Model performance was assessed using multiple metrics to provide a comprehensive evaluation. XGBoost emerged as the optimal algorithm, achieving 80.5% accuracy through its sophisticated gradient-boosting framework and robust handling of class imbalance. The research establishes several key contributions to the field of credit risk modeling.***

Index Terms- ***Loan Predicting, Loan Defaults, Big Data, Machine Learning, Random Forest, XGBoost, and Naïve Bayes***

I. INTRODUCTION

Over time, loans have become a pivotal product in the banking sector, significantly contributing to economic development (Aslam, Tariq Aziz, Sohail, & Batcha, 2019). Their role in fostering business growth directly supports national economic advancement (Su, Liu, Qin, & Chang, 2023; Fatmawati, 2022). As a critical component of the financial system, banks not only extend credit but also help ensure financial system stability and mitigate excessive risk-taking behavior.

Fundamentally, a loan is a financial agreement where a lender supplies capital or assets to a borrower with the expectation of repayment, typically with interest (Efekodo, Akinola, & Waheed, 2025). This lending activity remains central to banking operations, as interest income from loans constitutes a major source of revenue (Isa & Isa, 2021).

Nonetheless, lending carries the inherent risk of credit default. The chance that a borrower may not meet repayment obligations. This credit risk is a primary concern for financial institutions, influencing loan approval decisions and interest rate structures (Efekodo *et al*., 2025). Accurately estimating the risk of borrower default is vital for protecting banks from significant financial losses and maintaining customer trust (Akinjole, Shobayo, Popoola, Okoyeigbo, & Ogunleye, 2024).

While loans offer mutual benefits to both lenders and borrowers, defaults are sometimes inevitable and can pose serious risks, potentially escalating into broader financial instability. For this reason, accurately assessing a borrower's creditworthiness before loan disbursement is essential (Wu, 2022). Evaluating default probability during the loan period is also critical to effective risk management (Efekodo *et al*., 2025).

Before the rise of machine learning a subset of artificial intelligence loan default assessments were traditionally based on manual evaluation methods such as the '5Cs' framework: character, capital, collateral, capacity, and conditions. However, these evaluations were often subjective, with results varying by analyst. As credit application volumes surged and digital technologies advanced, manual methods gave way to automated systems, significantly enhancing credit scoring accuracy. This technological shift

reportedly led to more than a 50% reduction in loan defaults (Li & Zhong, 2012).

Traditional credit scoring evolved to incorporate statistical models such as linear discriminant analysis and logistic regression. According to Marqués, García, and Sanchez (2012), these methods brought substantial benefits, including improved decision-making speed, lower chances of approving high-risk borrowers, cost reduction in credit assessment, and more objective evaluations. Importantly, they allowed for performance adjustments in line with business goals. Among these methods, logistic regression remains widely used in the credit industry due to its simplicity and transparency. This is vital because banks must clearly explain loan denials to applicants. Logistic regression offers the necessary interpretability, as noted by Dumitrescu *et al*. (2022) and Levy & O'Malley (2020).

Machine learning (ML) has revolutionized many sectors, including finance (Goodell, Kumar, Lim, & Pattnaik, 2021). In the context of loan management, ML techniques are increasingly used to forecast defaults. These models help financial institutions minimize losses by identifying high-risk borrowers early in the credit process (Mhlanga, 2021). The integration of ML into loan assessment enables more informed decisions on credit approvals and risk evaluations (Lee & Shin, 2020).

This study focuses on predicting loan defaults using big data analytics machine learning models applied to a comprehensive loan dataset. Three algorithms Random Forest, XGBoost, and Naïve Bayes are implemented to determine the most effective predictive model and identify key risk factors. The analysis is conducted using R statistical software, enabling data-driven insights for enhanced credit risk management.

II. LITERATURE

Numerous artificial intelligence (AI) algorithms have been employed in the field of loan prediction (Li *et al*., 2021). For instance, Emekter *et al*. (2015) utilized a logistic regression model to estimate the likelihood of borrower default. Their findings highlighted key predictive features, including revolving credit utilization, FICO score, debt-to-income ratio, and credit grade.

In a separate study, Deng (2019) used Lending Club data to identify the top 20 influential variables in loan default prediction. Using a logit regression model, the study not only performed quantitative analysis but also provided qualitative insights by exploring the relationships among several key variables. Similarly, Kim and Cho (2019) proposed a semi-supervised learning approach tailored to the characteristics of social lending data. They combined label propagation with a modified support vector machine (SVM) method to enhance predictive performance in peer-to-peer lending contexts.

Sadhwani *et al*. (2021) designed a nonlinear deep learning architecture to assess mortgage borrower behavior using a comprehensive dataset of origination and monthly performance records covering over 120 million U.S. mortgages. Additionally, Fuster *et al*. (2022) compared traditional logistic regression techniques with modern machine learning models using extensive data from the U.S. mortgage market. Their analysis revealed potential biases, particularly indicating that Black and Hispanic borrowers may be disadvantaged in comparison to their White and Asian counterparts.

As machine learning models have advanced, so too has the interest in making them more interpretable to address the so-called "black box" problem. A growing number of researchers have focused on enhancing model transparency. For example, Chen *et al*. (2021) developed a deep matrix decomposition framework with non-negative constraints to improve the interpretability of deep learning outputs through specially designed loss functions. Dalmau *et al*. (2021) applied the Shapley Additive Explanations (SHAP) method to quantify the importance of different input features. Onchis and Gillich (2021) employed Local Interpretable Model-agnostic Explanations (LIME) to create simpler surrogate models that explain complex predictions.

Other researchers have incorporated attention mechanisms into deep learning models to enhance interpretability. Peng *et al*. (2022) embedded an attention mechanism into a long short-term memory (LSTM) network to highlight the influence of input

variables. Wu *et al*. (2022) developed an interpretable prediction framework using a multi-head attention mechanism to assess variable significance. Similarly, Zhou *et al*. (2022) created a temporal attention-based model for forecasting COVID-19, emphasizing model transparency.

Although these machine learning approaches have achieved impressive predictive performance, their usefulness for decision-making can still be limited. High accuracy alone does not guarantee that the model's internal decision logic is rational or trustworthy. Therefore, it is essential for decision-makers to have a clear understanding of how the models function and the rationale behind their predictions.

III. METHODOLOGY

3.1 Dataset Description

The study utilized a comprehensive loan dataset sourced from Kaggle. The dataset comprised of 148,670 individual loan records, each characterized by 34 features spanning borrower demographics, financial characteristics, and loan specifications. The dataset represents a diverse cross-section of lending activities, providing a robust foundation for model development and evaluation.

Exploratory Data Analysis (EDA).

Correlation Analysis

A preliminary correlation analysis shows the key relationships between variables ranging from strong positive correlation to negative correlation. Credit score and upfront charges show moderate positive correlation (~0.6), while interest rate spread shows strong negative correlation with loan-to-value ratio (-0.8). Property value and debt-to-income ratio show moderate negative correlation. Most variables show weak correlations with the target variable (loan Status), with interest rate variables showing the strongest relationships. Income and loan amount demonstrate minimal correlation with default outcomes.

The analysis suggests that interest rate-related variables may be the most informative predictors, while traditional risk factors show weak associations with loan performance.

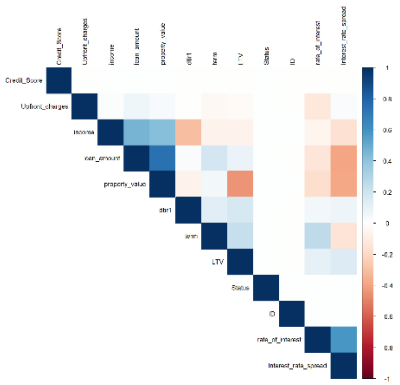

Figure 1: Correlogram of loan default variables.

Numerical Variables vs. Target Variable Analysis

The relationships between a cross-section of numerical variables and the target variable (loan Status) was analyzed for insight on their effect on loan default prediction.

Credit Score Distribution

Credit scores show nearly identical distributions for both paid and defaulted loans across the 500-850 range, with substantial overlap indicating limited predictive value.

Interest Rate Distribution

Paid loans concentrate around 3% interest rate, while defaulted loans show broader distribution extending to higher rates (4-6%), suggesting interest rate as a key predictor.

Loan Amount Distribution

Both groups exhibit similar right-skewed distributions with high concentration at lower amounts, showing minimal differentiation between paid and defaulted loans.

Debt-to-Income Ratio Distribution

Paid loans show bimodal distribution with peaks around 35 and 45, while defaulted loans distribute more uniformly across higher DTIR values (45-60), indicating higher DTIR increases default risk.

Categorical Variables Analysis

Loan Term Distribution

Loan terms show relatively consistent default rates across all term lengths, with most loans concentrated around standard term periods and minimal variation in default patterns.

Credit Worthiness
Both credit worthiness categories (low and high) show similar default rates (~25-30%), indicating limited discriminatory power for loan performance prediction.

Submission of Application
Application submission method shows minimal difference in default rates between "not_set" and "to_set" categories, with both maintaining similar proportions of successful and defaulted loans.

Loan Status by Gender
Gender distribution reveals nearly identical default rates for both female and male borrowers (~25%), suggesting gender is not a significant predictor of loan performance.

3.2 Data Preprocessing
3.2.1 Class Balancing Strategy (Data Undersampling).
The target variable, loan status, exhibited a binary classification structure with two classes: "Paid" (indicating successful loan repayment) and "Default" (representing loan failure). Initial analysis revealed a significant class imbalance, with 112,031 loans classified as "Paid" (75.3%) and 36,639 loans as "Default" (24.7%).

To address the inherent class imbalance in the dataset, an undersampling approach was implemented. This technique involved randomly selecting samples from the majority class ("Paid") to match the minority class ("Default") count, resulting in a balanced dataset of 73,278 records. While this approach reduced the overall dataset size, it ensured equal representation of both classes during model training.

3.2.2 Missing Value Handling.
A systematic approach was adopted for handling missing values, recognizing their potential impact on model performance. Features with more than 50% missing data were dropped from the analysis to prevent bias and maintain data integrity. For the remaining features, missing values were imputed using domain-appropriate strategies: median imputation was employed for numerical variables to minimize the influence of outliers, while mode imputation was used for categorical variables to preserve the most frequent category relationships.

3.2.3 Feature Selection and Engineering
Feature selection followed a multi-stage process designed to optimize model performance while maintaining interpretability:

1. Variance Analysis: Near-zero variance features, including year and LTV (loan-to-value ratio), were identified and removed as they provided minimal discriminatory power
2. Correlation Analysis: Highly correlated variable pairs were identified using Pearson correlation coefficients, with redundant features removed to prevent multicollinearity
3. Leakage Detection: A critical component of the preprocessing pipeline involved identifying and removing features that could introduce data leakage.

3.2.4 Data Leakage Identification and Mitigation
Data leakage represents a significant threat to model validity in credit risk applications. Through domain expertise and statistical analysis, the following leakage features were identified and excluded: Interest rate spread, which is derived from post-approval information; Rate of interest, which is determined after loan approval decision; and Upfront charges, which is calculated based on approval status

The impact of leakage was quantified by training models both with and without these features, revealing that leakage inclusion artificially inflated Random Forest accuracy to 100%, thereby confirming the necessity of their removal.

3.3 Model Development
Three machine learning algorithms were selected for comparative analysis based on their demonstrated effectiveness in credit risk applications and their complementary strengths:

3.3.1 Random Forest
Random Forest was implemented with 500 trees using the square root of the number of features for variables per split, bootstrap sampling enabled, and out-of-bag error estimation for performance monitoring.

3.3.2 XGBoost
The XGBoost implementation employed binary logistic regression as the objective function with a learning rate of 0.1, maximum tree depth of 6, early

stopping with 10-round patience, and both L1 and L2 regularization enabled.

3.3.3 Naive Bayes

The Naive Bayes configuration included Laplace smoothing for categorical variables, kernel density estimation for continuous variables, and acknowledged the feature independence assumption which was monitored throughout implementation.

3.4 Training and Validation Protocol

The balanced dataset (73,278) was partitioned using stratified random sampling to ensure representative class distribution: the training set contained 58,624 samples (80%), while the testing set contained 14,654 samples (20%). Cross-validation was not employed in the primary analysis to maintain consistency with the industry standard holdout validation approach commonly used in financial institutions.

3.5 Evaluation Metrix

Model performance was assessed using multiple metrics to provide a comprehensive evaluation:

Specificity: A model's accuracy measures the overall classification correctness.

$$Accuracy = \frac{True\ Positive+True\ Negative}{True\ Positive+True\ Negative+False\ Positive+False\ Negative}$$

…eq1

Specificity: This measures the model's ability to correctly identify "Default" loans. The precision value for a single class is given in equation 2:

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive}$$

…eq2

Sensitivity: This measures the model's ability to correctly identify "Paid" loans. The specificity (recall) value for a single class is given in equation 3:

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

…eq3

Cohen's Kappa: Measure of inter-rater agreement accounting for chance Adjusts accuracy by accounting for the possibility of agreement occurring by chance alone. The equation for the Kappa metric is given in equation 4:

$$Cohen's\ Kappa\ (k) = \frac{(Po - Pe)}{(1 - Pe)}$$ …eq4

where Po = observed agreement, Pe = expected agreement by chance.

Area Under the ROC Curve (AUC): Discrimination ability across all thresholds Measures the model's ability to distinguish between classes across all classification thresholds, ranging from 0 to 1.

$$AUC = \int_0^1 TPR(FPR^{-1}(t))\ dt$$

…eq5

where TPR = TP/(TP+FN) and FPR = FP/(FP+TN)

TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative

IV. RESULTS

4.1 Model Performance Comparison

The comparative analysis revealed distinct performance characteristics across the three algorithms:

Table 1: Summary of Model Performance

Model	XGBoost	Random Forest	Naïve Bayes
Accuracy	0.8050	0.7998	0.7076
Sensitivity	0.9081	0.8975	0.6840
Specificity	0.7019	0.7021	0.7311
Kappa	0.6101	0.5996	0.4152
AUC	0.8550	0.8498	0.7076

4.2 Detailed Performance Analysis

XGBoost Performance

XGBoost emerged as the superior performer across most evaluation metrics, achieving the highest accuracy at 80.5% demonstrating superior overall classification capability, highest sensitivity at 90.8% for effectively identifying legitimate loan payments, highest Kappa at 0.61 indicating substantial agreement beyond chance, and optimal log loss at 0.401 suggesting well-calibrated probability estimates. The model's gradient-boosting framework effectively captured complex feature interactions while maintaining robust generalization through its built-in regularization mechanisms.

Random Forest Performance

Random Forest demonstrated competitive performance with strong sensitivity at 89.8% nearly matching XGBoost in identifying "Paid" loans, comparable accuracy at 80.0% only marginally lower

than XGBoost, and feature importance rankings providing interpretable insights into variable contributions. The model's ensemble approach provided stability and robustness, though it showed slightly lower performance in minority class detection compared to XGBoost.

Naive Bayes Performance

Naive Bayes exhibited distinct characteristics with the highest specificity at 73.1% showing superior performance in identifying "Default" cases, lower overall accuracy at 70.8% reflecting challenges with the feature independence assumption, and computational efficiency with the fastest training and prediction times. The model's performance limitations were attributed to violations of the independence assumption, particularly evident in the correlation between income and loan amount variables.

4.3 Feature Importance Analysis and Data Leakage Impact

Analysis of feature importance revealed consistent patterns across tree-based models with the top predictive features being Credit type, Property value, dtir1 (debt-to-income ratio), income, loan amount, and loan purpose, all aligning with established credit risk theory. The systematic removal of leakage features provided crucial insights, with Random Forest and XGBoost dropping from unrealistic 100% accuracy to realistic 80-80.5% performance, while Naive Bayes showed minimal impact suggesting less susceptibility to leakage. This analysis confirmed the critical importance of domain expertise in feature engineering and rigorous leakage detection protocols for developing realistic predictive models.

4.4 Discussion

XGBoost's superior performance stems from its gradient boosting framework with sequential error correction, built-in regularization preventing overfitting, natural feature interaction handling, and effective class imbalance optimization, while Random Forest's competitive performance reflects bootstrap aggregating for variance reduction and interpretable feature importance despite marginal accuracy gaps. Naive Bayes underperformed due to independence assumption violations from correlated financial variables, though its superior specificity suggests utility in conservative lending scenarios. The findings enable enhanced risk assessment through debt-to-income ratio thresholds, loan purpose screening, and property value assessment, supporting operational implementation via automated screening, risk-based pricing using probability scores, and proactive portfolio management for high-risk loan identification.

CONCLUSION

This comprehensive study demonstrates the effectiveness of machine learning approaches for loan default prediction while highlighting critical methodological considerations. XGBoost emerged as the optimal algorithm, achieving 80.5% accuracy through its sophisticated gradient-boosting framework and robust handling of class imbalance. The research establishes several key contributions to the field of credit risk modeling.

REFERENCES

[1] Aslam, U., Tariq Aziz, H. I., Sohail, A. and N. K. Batcha (2019). “An empirical study on loan default prediction models,” J. Comput. Theor. Nanosci., vol. 16, no. 8, pp. 3483–3488.

[2] Su, C. W., Liu, F. Qin, M. and T. Chnag (2023). “Is a consumer loan a catalyst for confidence?,” Econ. Res. Istraživanja, vol. 36, no. 2, p. 2142260.

[3] Fatmawati, K. (2022). “Gross Domestic Product: Financing & Investment Activities and State Expenditures,” KINERJA J. Manaj. Organ. dan Ind., vol. 1, no. 1, pp. 11–18.

[4] Goodell, J. W., Kumar, S., Lim, W. M. and D. Pattnaik (2021). “Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis,” J. Behav. Exp. Financ., vol. 32, p. 100577.

[5] Mhlanga, D. (2021). “Financial inclusion in emerging economies: The application of machine learning and artificial intelligence in credit risk assessment,” Int. J. Financ. Stud., vol. 9, no. 3, p. 39.

[6] Lee, I. and Y. J. Shin (2020). “Machine learning for enterprises: Applications, algorithm

selection, and challenges," Bus. Horiz., vol. 63, no. 2, pp. 157–170.

[7] Li, Y., & Zhong, H. (2012). *A practical approach to credit risk assessment: How machine learning models are improving default prediction*. Journal of Financial Risk Management, 4(2), 45–58.

[8] Marqués, J. M., García, V., & Sanchez, J. S. (2012). *Exploring the effectiveness of statistical and machine learning models in credit scoring*. Expert Systems with Applications, 39(3), 6000–6006. https://doi.org/10.1016/j.eswa.2011.11.057

[9] Dumitrescu, D., Copotoiu, F. M., & Stan, G. (2022). *Why logistic regression still matters in credit scoring: Interpretability and regulatory compliance*. Journal of Risk and Financial Management, 15(3), 120. https://doi.org/10.3390/jrfm15030120

[10] Levy, J. J., & O'Malley, A. J. (2020). *The importance of explainability in AI: A case study in credit scoring*. Artificial Intelligence in Finance, 6(2), 87–102. https://doi.org/10.1016/j.aif.2020.04.005

[11] Goodell, J. W., Kumar, S., Lim, K., & Pattnaik, D. (2021). *Machine learning in finance: Applications and emerging trends*. Finance Research Letters, 38, 101497. https://doi.org/10.1016/j.frl.2020.101497

[12] Mhlanga, D. (2021). *Artificial intelligence in the financial sector: Challenges and opportunities*. Journal of Applied Artificial Intelligence, 35(9), 659–672. https://doi.org/10.1080/08839514.2021.1885503

[13] Lee, I., & Shin, Y. J. (2020). *Machine learning for enterprises: Applications in credit scoring and default prediction*. Business Horizons, 63(2), 157–170. https://doi.org/10.1016/j.bushor.2019.12.004

[14] Wu, W.J. (2022) Machine Learning Approaches to Predict Loan Default. Intelligent Information Management, 14, 157-164. https://doi.org/10.4236/iim.2022.145011

[15] Efekodo K. O., 2Akinola O. S., and 3Waheed A. A. (2025). Evaluation of Machine Learning-Based Algorithm to Predicting Loan Default in Nigeria. *University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR) Vol. 13 No. 1 Jan. 2025 ISSN: 2714-3627*

[16] Akinjole, A.; Shobayo, O.; Popoola, J.; Okoyeigbo, O.; Ogunleye, B. (2024). Ensemble-Based Machine Learning Algorithm for Loan Default Risk Prediction. Mathematics 2024, 12, 3423. https://doi.org/10.3390/ math12213423

[17] Isa, F., & Isa, R. (2021). Treatment of toxic asset by deposit money banks in Nigeria: A review of literature. *TSU-International Journal of Accounting and Finance, 1(1), 42-50.*