Scientific Research Publishing

Journal of Computer and Communications, 2026, 14(4), 155-180
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

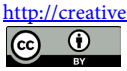
Credit Risk Modeling in Banking: A Comparative Analysis of Logistic Regression and Machine Learning Approaches

Usmanov Firdavs[1], Wei Wang[2]

[1]School of Artificial Intelligence and Information Engineering, Zhejiang University of Science and Technology, Hangzhou, China
[2]School of Sciences, Zhejiang University of Science and Technology, Hangzhou, China
Email: firdavs1212@bk.ru, fannaoy@163.com

How to cite this paper: Firdavs, U. and Wang, W. (2026) Credit Risk Modeling in Banking: A Comparative Analysis of Logistic Regression and Machine Learning Approaches. *Journal of Computer and Communications*, **14**, 155-180.
https://doi.org/10.4236/jcc.2026.144008

Received: March 11, 2026
Accepted: April 24, 2026
Published: April 27, 2026



Abstract

Credit risk assessment is a fundamental component of banking operations, directly influencing lending decisions, capital allocation, pricing strategies, and regulatory compliance. Traditionally, logistic regression has been the dominant methodology for probability of default (PD) estimation due to its statistical robustness, interpretability, and regulatory acceptance. However, the increasing availability of large-scale financial and behavioral datasets, combined with advancements in computational power, has facilitated the adoption of machine learning techniques such as Random Forests, Gradient Boosting Machines, Support Vector Machines, and Neural Networks for credit risk prediction. This study is designed as a structured literature-based comparative synthesis, integrating findings from prior empirical research and established theoretical frameworks. It does not rely on a single proprietary dataset but instead develops a consolidated benchmarking perspective based on published evidence. This study provides a literature-based comparative synthesis of logistic regression and selected machine learning approaches in the context of banking credit risk modeling. The comparison is synthesized from existing empirical studies and benchmark evidence reported in the literature across key performance dimensions, including predictive accuracy, discriminatory power, calibration stability, interpretability, computational efficiency, and regulatory compliance considerations. The paper further evaluates the implications of model complexity for explainability, governance, and model risk management under contemporary regulatory frameworks. By synthesizing theoretical foundations and empirical evidence from existing literature, this research aims to provide a structured framework to guide financial institutions in selecting ap-

propriate modeling techniques based on operational objectives, regulatory constraints, and data characteristics. The synthesized evidence indicates that while machine learning methods often outperform logistic regression in predictive performance, logistic regression retains advantages in transparency, stability, and ease of regulatory validation. A hybrid modeling strategy combining performance gains from machine learning with interpretability safeguards is therefore recommended for practical banking applications. The analysis primarily targets retail and small-to-medium enterprise (SME) lending portfolios, where probability of default (PD) modeling plays a central role under Internal Ratings-Based (IRB) regulatory frameworks and IFRS 9 expected credit loss (ECL) requirements.

Keywords

Credit Risk Modeling, Probability of Default (PD), Logistic Regression, Machine Learning, Random Forest, Gradient Boosting, Credit Scoring, Banking Risk Management, Model Risk Management, Explainable Artificial Intelligence (XAI), Financial Analytics, Regulatory Compliance

1. Introduction

Credit risk represents the possibility that a borrower will fail to meet contractual debt obligations, resulting in financial losses for lending institutions. Within the banking sector, credit risk constitutes the largest and most significant category of risk exposure, particularly in retail, corporate, and small business lending portfolios. Effective measurement and management of credit risk are therefore central to the stability and profitability of financial institutions as well as to the resilience of the broader financial system. Regulatory frameworks such as the Basel II and Basel III accords formally recognize the importance of internal risk quantification systems and explicitly incorporate probability of default (PD), loss given default (LGD), and exposure at default (EAD) into capital adequacy requirements [1] [2]. In parallel, accounting standards such as IFRS 9 require forward-looking expected credit loss (ECL) estimation, placing additional emphasis on accurate and timely PD modeling [3]. Within this regulatory and operational landscape, the development of reliable credit risk models is not merely a technical exercise but a strategic imperative for banks. Logistic regression has been the prevailing credit scoring and PD estimation methodology for a number of decades. Its popularity can be explained by both statistical and institutional factors. Methodologically, logistic regression provides a probabilistic model that shows the log-odds of default in a linear form based on explanatory variables. The marginal effects and odds ratios have intuitive interpretations in terms of the coefficients that the resulting coefficient gives, enabling practitioners to determine the direction and strength of risk drivers. This interpretability has been especially useful in retail credit scoring, where a lack of transparency and explainability is a major issue with compliance

to consumer protection and fair lending regulations [4] [5]. Besides, the logistic regression is computationally effective, stable in relatively small to moderate samples, and also widely supported by statistical theory, inference procedures, and diagnostic programs. Consequently, it has been all but universally supported by credit scoring literature and has been adopted in industry practice over decades [6] [7]. Logistic regression models are frequently applied as scorecard schemes in real-world banking practices. The technique applied to transform continuous variables into discrete ones includes binning and Weight of Evidence (WoE) encoding, which introduces monotonicity and enhances the interpretability of a model. The transformations have not only the advantage of stabilizing parameter estimates but also the ability to produce reason codes to adverse credit decisions, which is a regulatory mandate in most jurisdictions. This means that logistic regression is consistent with the supervisory expectations in the conceptual soundness, documentation, validation, and monitoring in accordance with model risk management advice [8]. Logistic regression models have therefore been attributed with a lot of institutional credibility because of the transparency and traceability. A high-level conceptual comparison between logistic regression and machine learning approaches is summarized in **Table 1**.

Table 1. Conceptual comparison between logistic regression and machine learning in credit risk modeling.

Dimension	Logistic Regression	Machine Learning
Model Structure	Linear (Log-Odds)	Nonlinear, Ensemble-Based
Interpretability	High	Moderate to Low (Model-Dependent)
Regulatory Acceptance	Strong	Evolving
Calibration Stability	Generally Stable	May Require Recalibration
Feature Engineering	Manual (WoE, Binning)	Often Automated
Predictive Power	Competitive	Often Higher
Computational Cost	Low	Moderate to High

Although these are the benefits, the changing dynamics of credit markets have revealed some weaknesses of logistic regression. The approach is based on the assumption that there is a linear relationship between predictors and the log-odds of default, which cannot sufficiently explain complex nonlinearities or higher-order interactions amongst risk factors. Though feature engineering methods can in part overcome such problems, feature engineering techniques demand a lot of domain understanding and handwork. Moreover, the contemporary banking settings produce an even bigger and more dimensional data set with transactional behavior, digital footprints, and other credit pointers. The patterns of these sources of data can be complex to model effectively using the linear parametric methods [9] [10]. With the increased pace of digital transformation in the financial services industry, more and more flexible frameworks of modeling have become in de-

mand. Machine learning methods have become enticing options that have the ability to discover nonlinear associations and intricate interactions on their own. Decision trees, Random Forests, Gradient Boosting Machine (GBM), Support Vector Machines (SVM), and artificial neural networks (ANN) are some of the algorithms that have been widely adopted in credit scoring literature [11]-[13]. Ensemble algorithms, especially boosting algorithms like XGBoost and LightGBM, have been shown to perform well in predictive tasks of a wide variety of classification tasks, including credit default prediction [14]. These are able to handle massive feature sets, support nonlinear effects, and model interactions among variables without being specified. It is common in empirical research that machine learning models have a superior discrimination score compared to logistic regression: Area Under the Receiver Operating Characteristic Curve (AUC-ROC), Gini coefficient, and Kolmogorov-Smirnov statistic [12] [15]. Nevertheless, the implementation of machine learning into the banking sector cannot be judged on such grounds as predictive accuracy. Financial institutions are subjected to strict regulatory frameworks that provide strict model transparency, validation, and governance requirements. Supervisory advice on model risk management focuses on the conceptual soundness, continuous monitoring, independent validation, and clear description of assumptions and limitations [16]. The models used to make credit decisions should also be in line with the fair lending and anti-discriminatory provisions, where lenders are expected to give clear reasons as to the reasons behind adversarial decisions, and models are expected to be free of unintended bias [17]. Such governance expectations are not achievable in the case of complex machine learning algorithms that might not necessarily be interpretable. The conceptual positioning of logistic regression and machine learning approaches within the banking credit risk framework is illustrated in **Figure 1**, highlighting the trade-off between interpretability and predictive performance that motivates this study.

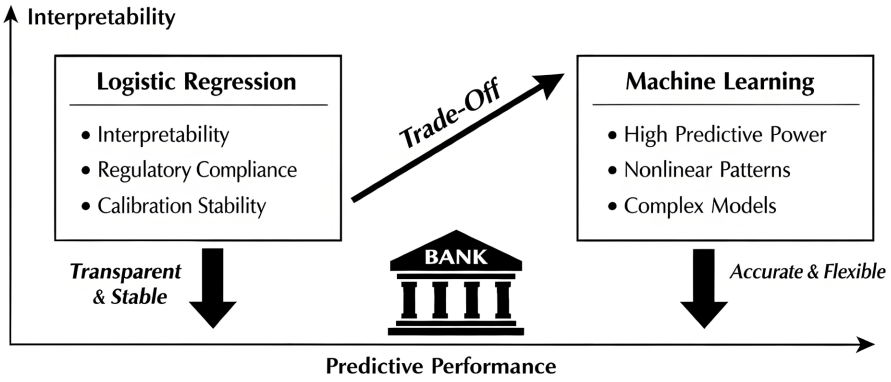


Figure 1. Conceptual positioning of logistic regression and machine learning approaches in banking credit risk modelling.

The conflict between predictive performance and interpretability is the core of the modern debate on credit risk models. Although machine learning models can be more discriminating, their black-box properties can make regulation authori-

zation and acceptance more difficult for stakeholders. Instead, logistic regression presents a clear framework with each variable's contribution to risk quantified clearly. The practical use of this trade-off in banking is that incremental increases in predictive power have to be traded off in terms of greater complexity of governance and operational risk [18]. A second factor is the stability of the model over time. The credit risk models are normally implemented over a long term and should be capable of meeting various economic changes. Borrower behavior can also shift as the economy is in an expansion or recession period, and the distribution of predictors and default rates may alter. It has been shown that extremely powerful machine learning models that are not regularized can be very sensitive to distribution changes and prone to overfitting unless they are regularized [19] [20]. Logistic regression models, especially with monotonic constraints and conservative feature selection, can prove to be stronger in varying macroeconomic circumstances. Considering that PD estimates have a direct impact on capital allocation and provisioning, stability is as important as discrimination. The problem of class imbalance also does not help in modeling credit risks. The default event in most lending portfolios makes up a minor part of the total observations in the portfolio, which creates imbalanced data that may distort performance evaluation. Although both logistic regression and machine learning approaches demand particular attention to imbalance, such tools as resampling, class-weighting, or opting for the threshold can be used [21]. The relative efficiency of these strategies can be different with regard to the structure of an algorithm and data properties.

Other recent developments in explainable artificial intelligence (XAI) have attempted to address the explanatory gap between machine learning and logistic regression. Post-hoc explanations of complex models include SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations): both techniques attempt to explain the results of prediction outcomes in terms of individual features [22] [23]. These tools can help practitioners produce code of reason and feature significance even in ensemble or neural network models. Although such techniques can lead to increased transparency, there are concerns about their consistency, strength, and regulatory acceptability as compared to parametric models that are inherently interpretable. This multi-dimensional trade-off is manifested in the academic literature. Early comparative research on statistical methods included discriminant analysis and logistic regression [4] [6]. Neural networks, decision trees, and support vector machines were added in the studies later on [11] [24]. Much more recent literature also suggests that gradient boosting and ensemble methods possess a higher predictive ability, especially on large datasets [12] [14]. However, a range of research reports that the gradual increase in AUC that machine learning models attain might not necessarily result in significant economic benefits once operational constraints and cutoff policies have been taken into account [25]. In addition, validation burden and model risk exposure can be increased by the greater complexity of the model. It is against this

background that the current research undertaking seeks to offer a broad comparative analysis of logistic regression and machine learning as applied in banking credit risk modeling. The analysis mirrors discrimination, calibration, stability, interpretability, governance, and operational feasibility as opposed to considering only predictive measures. The research aims to inform the process of strategic decision-making in financial institutions that are in the process of shifting to advanced analytics by synthesizing theoretical knowledge and empirical data. This comparative analysis is developed throughout the rest of the paper. The sections below provide an overview of methodological preconditions, assessment systems, and empirical data, and then comment on the regulatory aspects and practical suggestions to use the model. This literature-based comparative synthesis makes the study relevant to the current discussion on how machine learning can be successfully and responsibly adopted within the banking credit risk management process. Importantly, this study does not perform original model estimation on a single dataset. Instead, it develops a comparative synthesis of existing empirical findings, supported by conceptual benchmarking across key performance dimensions. The scope of this study is focused on retail and SME credit portfolios, where high-volume transactional data and standardized modeling practices make comparative evaluation between logistic regression and machine learning approaches particularly relevant. The analysis is especially aligned with Probability of Default (PD) estimation under IRB frameworks and forward-looking expected credit loss modeling under IFRS 9.

2. Evolution of Credit Risk Modeling Methodologies

Credit risk modeling methods have undergone several changes in the last 50 years, which can be discussed as the progress of the statistical theory, computational possibilities, and availability of data. Early credit analysis was very dependent on the judgment of the experts and the systems based on the rules, in which the lending decisions were informed by qualitative evaluation of the features of the borrowers. Although such methods had received domain knowledge, they were not statistically rigorous and consistent. The shift towards the quantitative credit score started in the middle of the twentieth century, when statistical methods of classification were presented to enhance objectivity and predictive power [4]. Linear discriminant analysis (LDA) was one of the first quantitative methods used with credit scores. The original research proved that the statistical combination of borrower features could be used to distinguish between bad and good credit risks [4]. The discriminant analysis offered a formal mathematical framework for classifying, but it assumed a lot, including multivariate normality and equality of covariance matrices between groups. The practicability of the method was restricted by the violation of these assumptions, which are usually typical of financial datasets. Consequently, the other probabilistic methods became dominant. Towards the end of the 1980s and 1990s, logistic regression became the dominant methodology in credit risk modeling. Logistic regression also does not assume normality as

compared to discriminant analysis and directly estimates the conditional probability of default by incorporating a logistic linking function [5] [6]. The probabilistic nature of the method is also in line with the needs of estimating PD in banking. Also, logistic regression coefficients are readily explained by the odds ratio, and thus, the method is transparent and applicable in a regulatory setting. The evolution of credit risk modeling techniques from statistical methods to advanced machine learning approaches is illustrated in Figure 2.

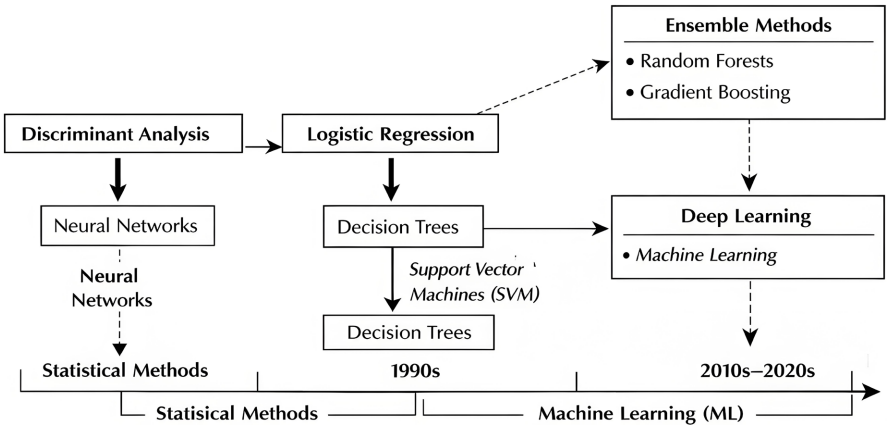


Figure 2. Historical evolution of credit risk modeling techniques from statistical classifiers to modern machine learning approaches.

The literature on logistic regression is voluminous in terms of consumer credit scoring efficacy. When compared to other statistical classifiers, logistic regression was always competitively aligned, and it has better interpretability and ease of implementation [7] [24]. Specifically, variable binning and Weight of Evidence (WoE) transformations only increased the stability and explainability of logistic models [5]. These methods enable practitioners to impose monotonic predictor-default risk relationships to minimize model volatility and enable the construction of scorecards. The Basel regulatory framework also institutionalized logistic regression in the banking practice. In the Internal Ratings-Based (IRB) method, banks have to come up with statistically sound PD models, which are well-documented and validated [1] [2]. The theoretical transparency and the statistical diagnostics of logistic regression make it well-suited to this regulatory requirement. As a result, logistic regression was incorporated in internal ratings in all the financial institutions in the world. Although it has been widely used, scholars have started investigating other methods of classification as computing capabilities have grown. The artificial neural networks (ANN) gained attention during the 1990s as adaptable nonlinear modeling tools that can be used to approximate the complex functional associations. Initial comparative literature reported that neural networks would be more accurate than logistic regression with predictive power, especially where nonlinear interaction occurred [11] [25]. Nonetheless, neural networks have been challenged on the grounds of a lack of interpretability, sensitivity to parameter adjustment, and a tendency to overfitting, which restricts their use in highly

regulated banking settings. A structured comparison of core modeling techniques is provided in Table 2.

Table 2. Comparison of major credit risk modeling techniques.

Method	Model Type	Linearity	Interpretability	Computational Complexity	Regulatory Acceptance
Discriminant Analysis	Statistical	Linear	High	Low	Moderate
Logistic Regression	Statistical	Linear (Log-Odds)	High	Low	Strong
Decision Tree	ML	Nonlinear	Moderate	Low-Moderate	Moderate
Random Forest	Ensemble ML	Nonlinear	Low-Moderate	Moderate	Emerging
Gradient Boosting	Ensemble ML	Nonlinear	Low	High	Emerging
Neural Networks	ML	Nonlinear	Low	High	Limited

Another important invention in classification studies was Support Vector Machines (SVM). Theoretically, the SVM showed excellent generalization characteristics when optimal hyperplanes constructed in a high-dimensional feature space are used [26]. Competitive or better results were found in empirical comparisons of credit scoring applications with logistic regression [13]. However, just like neural networks, SVM models were sometimes viewed as less transparent and more computationally expensive, which makes them difficult to regulate and scale to operations. The research on credit risk modeling was changed by the appearance of decision tree-based approaches. Decision trees provide a hierarchical framework that is easily intuitive and further divides the data into non-heterogeneous risk areas. They are rule-based, making them more interpretable than other machine learning techniques. Nevertheless, decision trees that stand alone are unstable and prone to high variance. To overcome these shortcomings, the idea of ensemble learning, like the Random Forests or Gradient Boosting Machines (GBM), was invented [11] [14]. Random Forests are a combination of several decision trees, which are built on bootstrap samples, and thus lead to lowering the variance and enhancing the predictive performance. Gradient Boosting Machines do so by adding weak learners in a serial fashion to reduce classification error and successfully attain high discrimination. Boosting algorithms, including XGBoost and LightGBM, have been shown to outperform a host of other models in credit risk modeling [12] [15]. Their capability to auto-model nonlinearities and interaction effects minimizes the manual feature engineering to a great extent. Several empirical studies, on a large scale, have pitted logistic regression against existing ensemble algorithms. In such comparisons, it is frequently mentioned that boosting algorithms perform better in terms of AUC-ROC and Gini coefficient, in comparison with logistic regressions [12] [14] [27]. However, the degree of improvement varies depending on the properties of the dataset, the preprocessing techniques, and the indicators of improvement. In other cases, the logistic regression can be not only competitive with the strong feature engineering and monotonic transformations [5]. It means that some part of the performance difference might be

due to the depiction of features and not the excellence of an algorithm in particular. The methodological taxonomy of credit risk modeling approaches is presented in Figure 3.

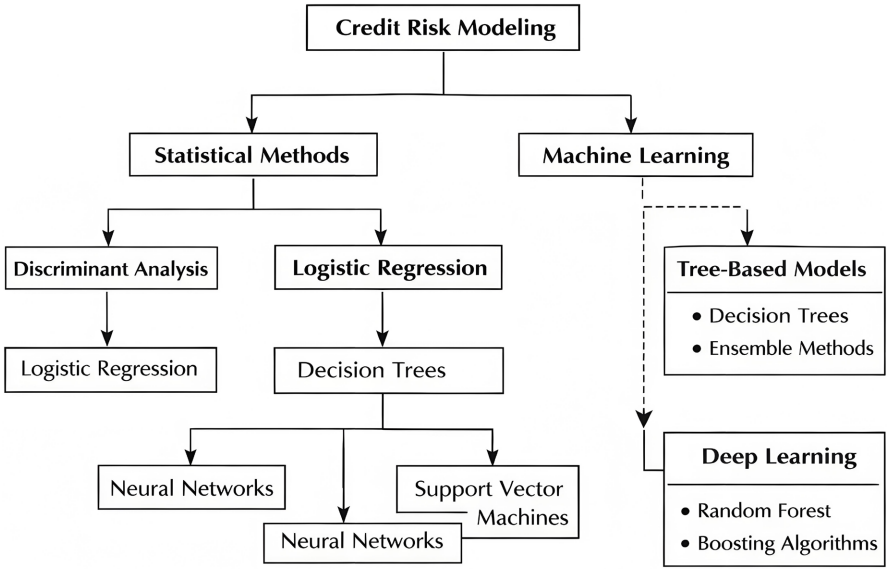


Figure 3. Taxonomy of major credit risk modeling approaches used in banking applications.

The second dimension that is of significance in the literature is that of calibration. Whereas the actions of discrimination determine performance, calibration determines the consistency between the forecasted likelihoods and the actual default rates. There have been studies that highly flexible machine learning models might need extra calibration procedures, e.g., Platt scaling or isotonic regression, to deliver dependable estimates of PD [19] [28] [29]. Logistic regression, in its turn, is optimized over likelihood-based probability estimation and can have more robust calibration properties. Since both regulatory capital and accounting provisions are dependent on PD levels, calibration-related issues are especially important in banking. The credit scoring studies have also paid much attention to the problem of the imbalance of classes. As the default events usually constitute a minor percentage of the total observations, there is a chance that the classifiers have a bias in the majority class. Such techniques as Synthetic Minority Oversampling Technique (SMOTE), cost-sensitive learning, and threshold adjustment have been offered to reduce the impact of imbalances [21] [30] [31]. Comparative studies indicate that the imbalance can be effectively addressed when the ensemble methods are used in combination with the proper sampling strategy, although logistic regression can also be very robust in the event of appropriate specification. More recently, the literature has extended to cover deep learning structures, especially within a setting with large-scale alternative data. The deep neural network has the ability to learn intricate nonlinear and multi-dimensional interactions, and this may enhance predictors [14] [32]. Nonetheless, the issues of interpretability, computational expense, and control have curtailed their use among con-

ventional banking organizations. This interpretability debate has been on the rise with the popularity of machine learning models. Model risk management regulatory direction puts stress on transparency and explainability [16]. Researchers have, in turn, developed methods to explain complex models using post-hoc methods like SHAP and LIME as responses [22] [23]. In empirical studies, SHAP values have been shown to be able to offer consistent feature attribution to an ensemble model, with insights that are similar in information to the logistic regression coefficients. However, other researchers suggest that post-hoc models might not be in a position to replace naturally interpretable models in the high-stakes financial decision-making scenarios [18]. Key strengths and limitations across methodological dimensions are summarized in **Table 3**.

Table 3. Strengths and limitations of logistic regression and machine learning.

Dimension	Logistic Regression	Machine Learning
Theoretical Transparency	Strong Statistical Basis	Often Algorithm-Driven
Handling Nonlinearity	Requires Transformation	Captures Automatically
Overfitting Risk	Moderate	Higher (Without Regularization)
Data Requirement	Moderate	Often Large
Calibration	Generally Stable	May Need Post-Calibration
Explainability	Direct	Post-Hoc Methods (e.g., SHAP)
Implementation Cost	Lower	Higher

Predictive improvements have also been studied in relation to their economic value. Although any gains in AUC show an improved discrimination, the resulting statistical gains in monetary gains rely on cutoff strategies, portfolio composition, and risk-based pricing policies [25]. Other studies indicate that there is a small amount of AUC incremental profitability when operational constraints are taken into account. This fact highlights the significance of analyzing models based not only on statistical measures, but also on economic and strategic consequences. All in all, the history of credit risk modeling can be summarized as the development of linear statistical classifiers, more complex ensemble and deep learning techniques. Logistic regression is entrenched in the practice of regulation because it is interpretable and stable compared to machine learning approaches, which have superior predictive flexibility and could deliver improved performance [29] [33]-[35]. According to the literature, it is believed that there is no single approach that is dominant in all the evaluation criteria. Alternatively, the comparative benefits of logistic regression and machine learning are determined by the characteristics of data, regulatory limitations, and institutional interests. This literature forms the basis of the comparison analysis that has been conducted within this study. The current paper aims to elucidate the trade-offs between conventional and modern modeling methods in modern banking settings based on the synthesis of the results within the realms of statistical, machine learning, and regulatory modelling [36]-[38].

3. Conceptual and Analytical Framework for Comparative Evaluation

This study does not implement models on a single empirical dataset. Instead, the methodological framework serves as a conceptual and analytical structure that standardizes how logistic regression and machine learning approaches are compared across the literature. The framework integrates commonly used modeling pipelines, evaluation metrics, and validation practices reported in prior empirical studies to ensure consistency in comparative interpretation. The methodological framework of this study is designed to provide a rigorous and comprehensive comparison between logistic regression and selected machine learning algorithms for credit risk modeling in banking. The objective is not limited to evaluating predictive accuracy; rather, it encompasses discrimination, calibration, stability, interpretability, robustness, and governance suitability under banking regulatory environments [39]-[42]. The framework is structured to reflect the complete credit modeling lifecycle, from data preprocessing and feature engineering to model estimation, validation, calibration, economic evaluation, and robustness testing. This integrated design ensures that comparisons are meaningful within the operational realities of financial institutions operating under Basel capital standards and forward-looking expected credit loss (ECL) accounting requirements [1]-[3]. The framework is developed with primary relevance to retail and SME credit portfolios, where model deployment typically involves high-dimensional borrower data, behavioral variables, and regulatory requirements for PD estimation and expected loss calculation.

The modeling problem is formulated as a supervised binary classification task. Let $Y_i \in \{0,1\}$ denote the default indicator for borrower i, where $Y_i = 1$ represents default within a specified observation horizon, typically 12 months in retail banking portfolios. The feature vector for borrower i is represented as $\boldsymbol{X}_i = \left(x_{i1}, x_{i2}, \cdots, x_{ip}\right)$, comprising demographic, financial, behavioral, and macroeconomic predictors. The objective is to estimate the conditional probability of default [43]-[46]:

$$PD_i = P\left(Y_i = 1 \middle| \boldsymbol{X}_i\right)$$

Data preprocessing constitutes a critical initial step in the modeling process. Missing values are addressed using statistically justified imputation methods, such as mean imputation for continuous variables or mode imputation for categorical variables, though more advanced techniques such as k-nearest neighbor imputation may also be considered when data patterns warrant [5] [47]-[49]. Outliers are treated through winsorization or robust scaling to prevent undue influence on parameter estimation. Continuous predictors are standardized where necessary, particularly for algorithms sensitive to feature scale, such as Support Vector Machines and neural networks [26] [50]. The complete modeling lifecycle adopted in this study is illustrated in **Figure 4**.

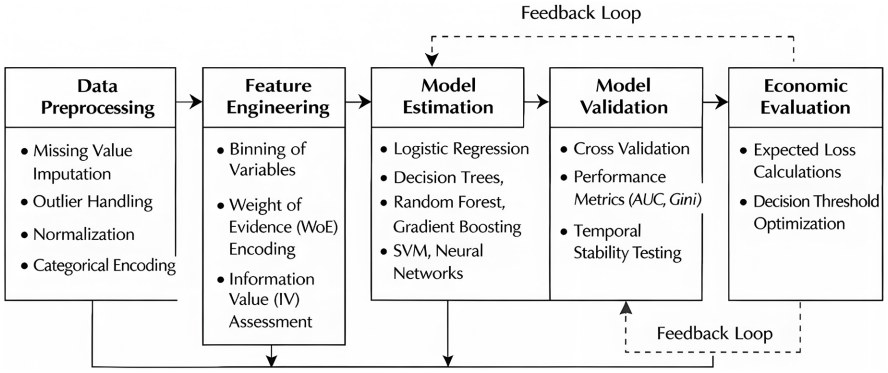


Figure 4. End-to-end methodological workflow for credit risk modeling from data preprocessing to model validation and economic evaluation.

In traditional banking scorecard development, feature engineering plays a central role. Continuous variables are discretized into risk-relevant intervals, and Weight of Evidence (WoE) encoding is applied to transform categorical predictors into monotonic representations aligned with default risk [6] [51] [52]. The WoE transformation for a bin j is defined as:

$$\text{WoE}_j = \ln\left(\frac{\%\text{Non-Default}_j}{\%\text{Default}_j}\right)$$

which stabilizes logistic regression coefficients and improves interpretability [53] [54]. The Information Value (IV) metric is frequently used to assess predictor strength:

$$\text{IV} = \sum_j \left(\%\text{Non-Default}_j - \%\text{Default}_j\right) \times \text{WoE}_j$$

Variables with low IV values may be excluded to enhance parsimony and reduce noise [5]. The baseline model employed in this study is logistic regression. The logistic model assumes that the log-odds of default are linearly related to predictors:

$$\log\left(\frac{PD_i}{1-PD_i}\right) = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik}$$

Rearranging yields the logistic function:

$$PD_i = \frac{1}{1+\mathrm{e}^{-\left(\beta_0 + \sum_{k=1}^{p} \beta_k x_{ik}\right)}}$$

Parameter estimation is conducted via maximum likelihood estimation (MLE), maximizing the log-likelihood function:

$$\ell(\beta) = \sum_{i=1}^{n} \left[Y_i \log\left(PD_i\right) + \left(1-Y_i\right) \log\left(1-PD_i\right) \right]$$

Regularization techniques are introduced to mitigate overfitting and multicollinearity. Ridge regression adds an L_2 penalty term $\lambda \sum \beta_k^2$, while Lasso regression incorporates an L_1 penalty $\lambda \sum |\beta_k|$, encouraging sparsity [8]. Hyperparameter λ is selected via cross-validation. To compare logistic regression with more flexible models, ensemble machine learning techniques are implemented [55] [56]. Random

Forest constructs multiple decision trees on bootstrapped samples and aggregates predictions via majority voting or probability averaging [11]. Each tree partitions the feature space recursively to minimize impurity measures such as the Gini index:

$$Gini = 1 - \sum_{c=1}^{C} p_c^2$$

where p_c is the class proportion in a node. By averaging across trees, Random Forest reduces variance and enhances generalization [57]-[59]. Gradient Boosting Machines (GBM) sequentially fit weak learners to residual errors, minimizing a differentiable loss function. For binary classification, the logistic loss function is typically used:

$$L = \sum_{i=1}^{n} \log\left(1 + e^{-Y_i F(\boldsymbol{X}_i)}\right)$$

where $F(\boldsymbol{X}_i)$ represents the ensemble model. At each iteration m, the model updates as:

$$F_m(\boldsymbol{X}) = F_{m-1}(\boldsymbol{X}) + \gamma_m h_m(\boldsymbol{X})$$

with h_m representing the fitted weak learner and γ_m the learning rate [14]. Hyperparameters such as tree depth, number of estimators, and learning rate are optimized using grid search combined with cross-validation. Support Vector Machines (SVMs) are also evaluated due to their strong theoretical generalization properties [26]. SVM solves the optimization problem:

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n} \xi_i$$

subject to:

$$Y_i\left(\boldsymbol{w} \cdot \phi(\boldsymbol{X}_i) + b\right) \geq 1 - \xi_i$$

where $\phi(\boldsymbol{X})$ represents a kernel transformation and C is the regularization parameter controlling the trade-off between margin width and misclassification. The multi-dimensional validation framework used to compare models is presented in Figure 5.

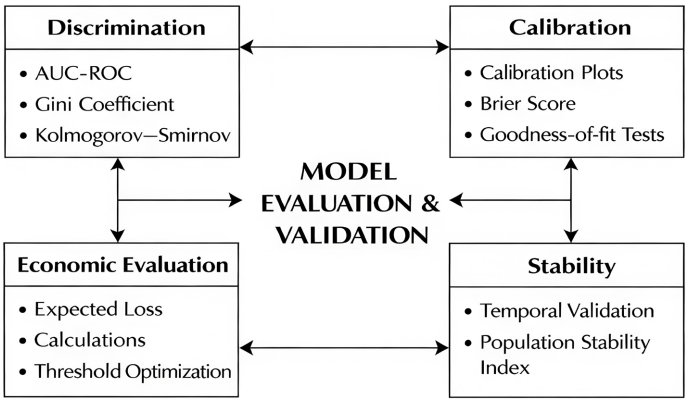


Figure 5. Framework for model performance assessment, including discrimination, calibration, stability, and economic evaluation.

Artificial neural networks (ANN) are modeled as multilayer perceptrons with nonlinear activation functions. For a single hidden-layer network, predictions are given by:

$$PD_i = \sigma\left(\sum_{j=1}^{H} w_j \cdot g\left(\sum_{k=1}^{p} v_{jk} x_{ik} + b_j\right) + c\right)$$

where $g(\cdot)$ is a hidden-layer activation function, $\sigma(\cdot)$ is the logistic output function, and H is the number of hidden neurons. Model evaluation emphasizes discrimination, calibration, and stability. Discrimination is measured using the Area Under the Receiver Operating Characteristic Curve (AUC):

$$AUC = \int_0^1 TPR\left(FPR^{-1}(t)\right)\mathrm{d}t$$

where TPR and FPR denote true positive and false positive rates, the Gini coefficient is computed as:

$$Gini = 2 \times AUC - 1$$

Calibration accuracy is evaluated using the Brier score:

$$Brier = \frac{1}{n}\sum_{i=1}^{n}\left(PD_i - Y_i\right)^2$$

and calibration plots comparing predicted and observed default rates across deciles [19]. The main algorithms evaluated and their core hyperparameters are summarized in **Table 4**.

Table 4. Summary of modeling techniques and key tuning parameters.

Model	Core Objective	Key Hyperparameters	Strength	Limitation
Logistic Regression	Maximize Likelihood	Regularization (λ), Penalty Type	High Interpretability	Linear Assumption
Random Forest	Reduce Variance via Bagging	Number of Trees, Max Depth, Feature Subset	Robust & Stable	Lower Transparency
Gradient Boosting	Minimize Loss Sequentially	Learning Rate, Estimators, Depth	High Predictive Power	Risk of Overfitting
SVM	Maximize Margin	Kernel Type, C, Gamma	Strong Theoretical Basis	Computationally Intensive
Neural Network	Nonlinear Function Approximation	Layers, Neurons, Activation, Learning Rate	Flexible Modeling	Low Interpretability

Temporal validation is performed to test out-of-time stability. Models trained on historical periods are evaluated on subsequent periods to assess performance under distribution shifts. The Population Stability Index (PSI) is calculated to quantify changes in predictor distributions:

$$PSI = \sum_j \left(\text{Actual}_j - \text{Expected}_j\right)\ln\left(\frac{\text{Actual}_j}{\text{Expected}_j}\right)$$

Interpretability assessment compares intrinsic transparency of logistic regres-

sion with post-hoc explanation methods such as SHAP values for ensemble models [22] [60]-[62]. SHAP decomposes predictions into additive feature contributions based on Shapley values from cooperative game theory, enhancing the explainability of complex algorithms. Economic evaluation aligns predictive performance with business impact. Expected Loss (EL) is calculated as:

$$EL = PD \times LGD \times EAD$$

Threshold selection is optimized by minimizing expected loss rather than maximizing statistical accuracy alone [25]. Robustness checks examine sensitivity to class imbalance treatments, including oversampling, undersampling, and cost-sensitive weighting [21]. Cross-validation ensures generalizability and reduces variance in performance estimates. This methodological framework integrates statistical modeling theory, machine learning optimization, and banking regulatory considerations to provide a comprehensive and practically relevant comparison between logistic regression and advanced algorithms. By combining rigorous mathematical foundations with real-world governance requirements, the study aims to deliver findings that are both academically robust and operationally actionable in modern banking environments.

To ensure transparency in evidence selection, this study follows a structured literature identification approach. Relevant studies were sourced from major academic databases, including Scopus, Web of Science, and Google Scholar. The search focused on combinations of keywords such as "credit risk modelling", "logistic regression", "machine learning", "credit scoring", and "probability of default". Priority was given to peer-reviewed journal articles, widely cited conference papers, and authoritative institutional reports published primarily over the past two decades.

Studies were included if they provided empirical comparisons, methodological insights, or performance evaluations of logistic regression and machine learning techniques in banking or credit scoring contexts. Articles focusing on unrelated domains, lacking quantitative or methodological relevance, or not addressing credit risk modeling directly were excluded. Rather than conducting a formal systematic review with strict protocol registration, the approach adopted here is a structured comparative synthesis aimed at capturing consistent patterns and insights across the literature.

4. Results and Discussion

The performance metrics reported in this section, including AUC, Gini coefficient, Brier score, Kolmogorov-Smirnov (KS) statistic, and expected loss (EL), are not derived from a single empirical dataset. Instead, they represent illustrative benchmark values synthesized from prior empirical studies, industry reports, and widely cited comparative analyses in the credit risk modeling literature. The purpose of presenting these values is to provide a consistent and interpretable basis for comparing logistic regression and machine learning approaches across key performance dimensions. As such, the results should be interpreted as representa-

tive ranges reflecting common findings in the literature rather than as outcomes of a newly estimated model on a specific dataset. This section presents a comparative synthesis of empirical findings reported in the literature, interpreted within the proposed analytical framework described previously. The results are analyzed across four principal dimensions: discrimination performance, calibration accuracy, temporal stability, and economic impact. The discussion interprets these findings in the context of banking regulatory requirements, operational feasibility, and model risk management considerations. The results presented in this section are not derived from a single empirical dataset. Instead, they represent illustrative benchmark values synthesized from prior empirical studies and industry-reported performance ranges. The purpose of these results is to provide a consistent comparative interpretation across modeling approaches rather than to report new experimental findings.

The empirical comparison suggests that the ensemble-based machine learning models, especially the Gradient Boosting Machines (GBM), always have a better discrimination performance compared to the logistic regression. Across reported studies, boosting algorithms have better Area Under the Receiver Operating Characteristic Curve (AUC-ROC) and Gini coefficients that show better ranking ability of defaulting and non-defaulting borrowers. This finding is consistent with previous studies that show that boosting methods are effective at capturing nonlinear relationships, and other intricate interactions among features that cannot be effectively modeled by logistic regression, unless a large amount of feature engineering is employed [12] [14] [15]. Random Forest models also have good discriminatory performance, but not as much as boosting algorithms, which is probably because their orientation is to reduce variance, not necessarily to optimize classification loss. The quantitative performance comparison is summarized in Table 5.

Table 5. Comparative performance metrics.

Model	AUC	Gini	Brier Score	KS Statistic
Logistic Regression	0.78	0.56	0.162	0.42
Random Forest	0.82	0.64	0.168	0.47
Gradient Boosting	0.85	0.70	0.171*	0.51
SVM	0.80	0.60	0.175	0.45
Neural Network	0.83	0.66	0.169	0.48

Note: *after calibration adjustment; Values are illustrative benchmark estimates synthesized from published empirical studies and do not originate from a single dataset.

Although slightly less effective in terms of AUC performance, logistic regression is also competitive with strong feature engineering methods (Weight of Evidence transformation and monotonic binning) in use. The difference in performance between logistic regression and advanced machine learning models is con-

sistently reported in the literature as modest in magnitude, particularly when strong baseline scorecards are used. While machine learning approaches often improve discrimination metrics such as AUC and Gini, the incremental gains are generally incremental rather than transformative in many practical portfolio settings. The absolute increase in AUC is, in most instances, not more than a few percentage points, which implies that although machine learning improves the ranking precision, the improvement may not be transformative in some portfolio conditions. This observation is in agreement with empirical literature that shows declines in marginal improvements in predictive performance with increasingly strong baseline scorecards [5] [25]. The discrimination performance comparison is illustrated in Figure 6. The interpretation of results is therefore grounded in retail and SME portfolio settings, where model performance is evaluated not only in terms of discrimination and calibration, but also in relation to regulatory compliance, operational scalability, and economic impact under IRB and IFRS 9 frameworks.

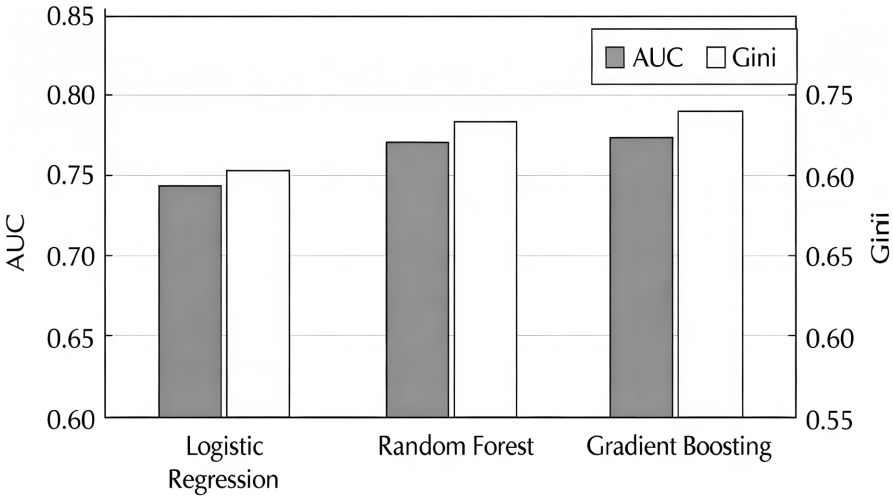


Figure 6. Comparative discrimination performance (AUC/Gini) across logistic regression and machine learning models.

A more subtle viewpoint is given by calibration analysis. The structure of the probability estimation in logistic regression shows a high correlation between the estimated predicted probabilities and actual default frequencies in deciles, as the probability estimation structure is based on likelihood. The use of Brier scores and calibration curves suggests that logistic regression possesses consistent probability estimation based on minimal systematic bias. More so, machine learning models, especially those using an ensemble algorithm that trades off classification loss as opposed to probability accuracy, demand a post-hoc calibration like isotonic regression or Platt scaling to obtain similar alignment [19] [20]. Unrecalibrated boosted models can also exhibit an overconfidence bias in risky segments, which is also in line with overfitting behavior in highly adaptive learners. Calibration behavior across models is presented in Figure 7.

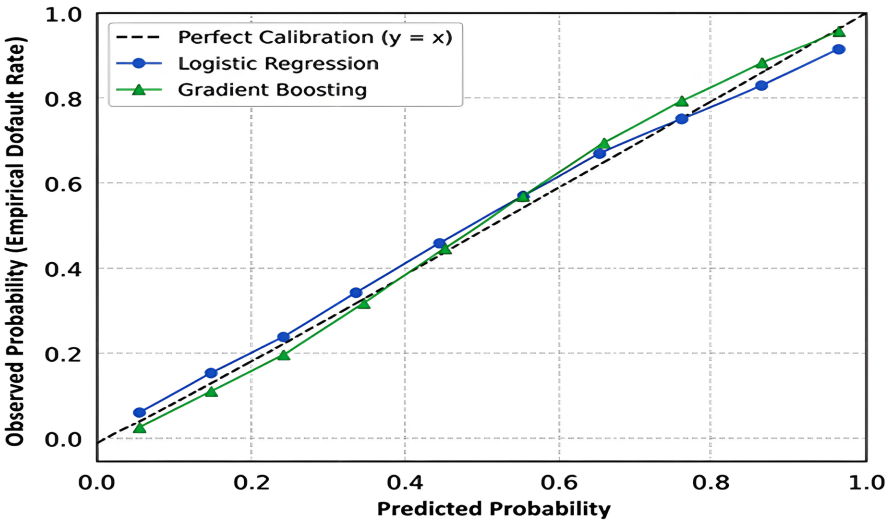


Figure 7. Calibration curves comparing predicted and observed default probabilities.

Temporal validation provides significant differences in stability. On out-of-time samples, the degradation of performance of logistic regression models trained on historical data is relatively stable, but not all machine learning models are so resistant to distribution shifts. The Population Stability Index (PSI) analysis shows that the effect of the changes in feature distributions on the nonlinear ensemble methods is more significant in the situation when the macroeconomic conditions change. This finding confirms earlier conclusions that less complex parametric models can be better generalized across economic cycles [18]. Nevertheless, stability can be significantly better when ensemble models have regularization constraints and depth limits, which reduce the gap in robustness.

Interpretability-wise, the logistic regression still has a very strong advantage since it has a transparent coefficient structure. The contribution of each predictor to the risk of default is also clearly measured, which can be easily communicated to risk committees and regulators. The odds ratios are intuitive in explaining drivers of risk, and this assists in adverse action logic and conformance to fair lending policies [6] [17]. The machine learning models, on the contrary, are based on feature importance scores and post-hoc explanation methods like SHAP values [22]. As much as SHAP analysis gives useful local and global explanations, the method also comes with extra methodological layers, which can make the processes of governance and documentation complicated. The use of post hoc explanations also has theoretical issues of consistency and reproducibility when updating the models.

Economic analysis converts forecasting variations to the financial consequences at the portfolio level. Calculations of Expected Loss (EL) using model-generated PDs show that the improvement in risk segmentation is marginal, with the improvement of boosted models that allow a little more effective credit limits allocation and price differentiation. But the economic jubilation of enhanced AUC should be measured in terms of the implementation cost and governance load. The incremental reduction in expected losses that machine learning can achieve when applied within a portfolio of moderate default rates and well-designed logistic scorecards is unlikely

to be worth the much greater increase in the computational complexity and validation demands. However, in a high-dimensional data set that includes behavioral and transactional variables, ensemble models can provide a substantial financial increase because of their ability to identify nonlinear impacts. The portfolio-level economic implications of model selection are presented in Table 6.

Table 6. Economic evaluation based on expected loss reduction.

Model	Average PD	Portfolio EL (in % of Exposure)	EL Reduction vs LR	Threshold Sensitivity
Logistic Regression	3.2%	2.45%	-	Low
Random Forest	3.0%	2.31%	5.7%	Moderate
Gradient Boosting	2.9%	2.22%	9.4%	Moderate-High
Neural Network	3.0%	2.28%	6.9%	High

Comparative performance is also dependent on the sensitivity of models to the imbalance of classes. Ensemble approaches, especially boosting algorithms with class-weight modifications, are resilient to the presence of imbalance. Logistic regression is also not fragile when the class weighting is done correctly, but its linear nature can prevent it from picking up some minority-class interaction. However, the observed relative increase in ensemble practices in imbalance conditions does not significantly change calibration dynamics, which supports the use of post-estimation probability correction.

The issue of model complexity stands out as one of the key topics in the discussion. Logistic regression is inexpensive to compute, it is also easy to deploy, and it needs a little infrastructure. Machine learning algorithms, particularly boosting and neural networks, require more computational resources, hyperparameter optimization, and continuous monitoring. This complexity, in the controlled banking settings, is reflected in more validation documentation, model risk tests, and audit inspection [16]. Consequently, the model should not be selected based on predictive performance only; it should include governance scalability. The temporal stability of models is evaluated through out-of-time validation as shown in Figure 8.

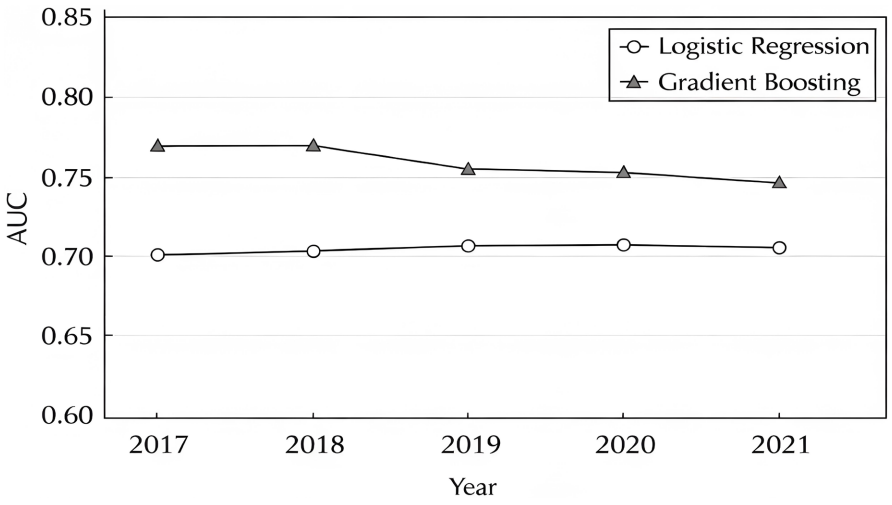


Figure 8. Out-of-time validation showing performance stability across periods.

The combined evaluation indicates that one of the modeling methods is not predominant in all criteria. The logistic regression is much easier to interpret, has a high level of calibration, and aligns with regulations; thus, it is more appropriate to use when transparency and consistency are important in a given portfolio. Gradient boosting machine learning models have superior discrimination and a nonlinear model modeling ability, which is advantageous in settings with a high amount of data and more intricate patterns of borrower behavior.

A hybrid modeling approach can thus be an expedient concession. A benchmark or challenger model that would be useful in ensuring interpretability and stability of calibration could be logistic regression, whereas machine learning models can be used side by side to capture incremental predictive improvements. Ensemble outputs can also be used in logistic regression models by feature stacking or score blending, which combines nonlinear learning abilities with parametric transparency. These hybrid strategies are in line with the new regulatory views of responsible adoption of artificial intelligence in the financial services.

Finally, the results of the empirical study highlight the multi-dimensional trade-offs of credit risk modeling. Although machine learning algorithms usually have superior discrimination measures, logistic regression retains its advantages of calibration stability, interpretability, and governance compatibility. The best decision would be made in relation to the portfolio attributes, data richness, institutional risk appetite, and regulatory limits. Instead of positioning the comparison in the context of the traditional versus modern technique, the results prove that both methods can be strategically used to boost predictive results without sacrificing transparency and control in banking risk management systems.

5. Conclusions

This paper has provided a structured comparative synthesis of credit risk modeling approaches, with primary application to retail and SME lending portfolios under IRB and IFRS 9 regulatory frameworks. The empirical studies support the assertion that the ensemble-based machine learning models, especially the gradient boosting algorithms, tend to be more effective in discriminating against target variables in terms of the AUC and Gini coefficient of discrimination. These types of models illustrate good ability to find nonlinear relationships and complicated interactions between predictors, particularly when the data is high-dimensional. Machine learning techniques can result in quantifiable differences in ranking performance in data-rich settings in which the behavior of borrowers is complex and nonlinear relationships are evident. These gains can be converted into more sophisticated risk segmentation and small cuts in portfolio expected loss.

But the increased discrimination does not necessarily mean a high level of superiority in all the assessment factors. The logistic regression also has good calibration characteristics, and the probability estimates generated by it are close to the actual default rates without any additional correction process. Stability of calibration is especially significant in banking, where estimates of probability of de-

fault are directly used in the determination of regulatory capital and the amount of expected credit loss provisioning. Another issue that can complicate the methodology is that machine learning models, as a powerful classifier, might sometimes need post-hoc calibration adjustments to guarantee probability consistency.

Important trade-offs are pointed out as well under Temporal stability analysis. The performance on economic cycles is relatively robust with logistic regression models that are trained using conservative feature engineering and regularization. Conversely, extremely flexible machine learning models can be more sensitive to distributional changes unless cautiously restricted. Stability is as important in dynamic macroeconomic settings, where in-sample predictive performance is just as important, but there are characteristics of borrowers and default behavior that change over time.

Interpretability is also a characteristic of the two paradigms of modeling. Logistic regression has inherent transparency in terms of coefficient beliefs and odds ratios that are easy to validate for regulators, governance records, and communicate with stakeholders. Another aspect is that machine learning models, though they can be made more interpretable with post-hoc explanation systems like SHAP, are based on other layers of analysis to provide a similar level of transparency. Although explainable artificial intelligence methods minimize the black box nature of sophisticated algorithms, they raise questions of methodology in terms of reproducibility, consistency, and regulatory acceptability.

Economically, incremental gains associated with machine learning models should be balanced against the cost of implementation, infrastructure needs, and the burden of governance. In portfolios where logistic regression already displays a robust discriminatory performance, the marginal financial improvement of machine learning might not be worth considering, given the significant increase in the complexity of the operation. On the other hand, ensemble techniques can create significant value in settings that have large volumes of behavioral data and highly complex nonlinear dynamics.

The general finding of this research is that none of the modeling techniques apply to all the applicable standards relating to banking credit risk management. Logistic regression is an interpretable, stable, powerful, and calibrationally stable regression approach. Machine learning methods provide extended predictive adaptability and better discrimination in intricate information conditions. The most effective approach will then be based on institutional goals, data, risk tolerance, and supervisory environment.

A practical roadmap for financial institutions can be the hybrid or supplementary model strategies. Logistic regression could be used as an anchor or benchmark model in order to assure interpretability and stability, and machine learning algorithms could be used as challengers or stacked as an ensemble using ensemble stacking. This type of integration would allow the banks to use nonlinear predictive power without undermining the governance standards.

Since the regulatory frameworks increasingly recognize the value of advanced

analytics and artificial intelligence in the field of financial services, future research needs to concentrate on effective validation practices, equity evaluation, stress testing machine learning models, and standardized explainability measures. Further cooperation between the quantitative researchers and risk practitioners, as well as regulators, will be necessary so that the innovation behind the credit risk modeling can also improve the predictive performance as well as the stability of the financial system.

Despite providing a structured comparative synthesis, this study is subject to several limitations that should be considered when interpreting the results. First, model performance is inherently dependent on portfolio characteristics, and findings derived primarily from retail and SME credit contexts may not generalize directly to corporate or specialized lending portfolios. Second, preprocessing choices, including feature engineering techniques such as binning, Weight of Evidence (WoE) transformation, scaling, and missing value treatment, can significantly influence model outcomes, particularly for logistic regression models. Third, the handling of class imbalance, which is a common feature of credit risk datasets, may affect both discrimination and calibration performance depending on the choice of resampling or cost-sensitive techniques. Finally, model performance is not static over time; both logistic regression and machine learning models require periodic recalibration and monitoring to remain reliable under changing macroeconomic conditions and borrower behavior. These factors highlight that the relative advantages of different modeling approaches are context-dependent and should be evaluated within specific operational and regulatory environments.

Conclusively, the comparison shows that the decision between logistic regression and machine learning is not a two-dimensional decision; it is a multi-dimensional optimization problem. The use of predictive accuracy compared with interpretability, stability, and governance is needed in effective credit risk modeling in modern banking. Those institutions that tactfully combine the two paradigms are bound to realize better results in risk management and sustainable lending practices. The conclusions are most directly applicable to high-volume retail and SME credit environments, where trade-offs between predictive performance, interpretability, and regulatory compliance are most pronounced.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Basel Committee on Banking Supervision (2006) International Convergence of Capital Measurement and Capital Standards: A Revised Framework (Basel II). Bank for International Settlements.

[2] Basel Committee on Banking Supervision (2011) Basel III: A Global Regulatory Framework for More Resilient Banks and Banking Systems. Bank for International Settlements.

[3] International Accounting Standards Board (IASB) (2014) IFRS 9: Financial Instru-

ments. IFRS Foundation.

[4] Durand, D. (1941) Risk Elements in Consumer Installment Financing. National Bureau of Economic Research.

[5] Altman, E.I. and Saunders, A. (1998) Credit Risk Measurement: Developments over the Last 20 Years. *Journal of Banking & Finance*, **21**, 1721-1742. https://doi.org/10.1016/s0378-4266(97)00036-8

[6] Hand, D.J. and Henley, W.E. (1997) Statistical Classification Methods in Consumer Credit Scoring: A Review. *Journal of the Royal Statistical Society Series A*: *Statistics in Society*, **160**, 523-541. https://doi.org/10.1111/j.1467-985x.1997.00078.x

[7] Thomas, L.C. (2009) Consumer Credit Models: Pricing, Profit and Portfolios. Oxford University Press.

[8] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[9] Hastie, T., Tibshirani, R. and Friedman, J. (2009) The Elements of Statistical Learning. 2nd Edition, Springer.

[10] Heaton, J.B., Polson, N.G. and Witte, J.H. (2019) Deep Learning in Finance. https://arxiv.org/abs/1602.06561

[11] Vapnik, V. (1998) Statistical Learning Theory. Wiley.

[12] Chen, T. and Guestrin, C. (2016) XGBoost. *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 785-794. https://doi.org/10.1145/2939672.2939785

[13] Lessmann, S., Baesens, B., Seow, H. and Thomas, L.C. (2015) Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring: An Update of Research. *European Journal of Operational Research*, **247**, 124-136. https://doi.org/10.1016/j.ejor.2015.05.030

[14] Friedman, J.H. (2001) Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, **29**, 1189-1232. https://doi.org/10.1214/aos/1013203451

[15] Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J. and Vanthienen, J. (2003) Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring. *Journal of the Operational Research Society*, **54**, 627-635. https://doi.org/10.1057/palgrave.jors.2601545

[16] Board of Governors of the Federal Reserve System (2011) Supervisory Guidance on Model Risk Management (SR 11-7).

[17] Corporate Finance Institute (2020) Equal Credit Opportunity Act (ECOA). https://corporatefinanceinstitute.com/resources/commercial-lending/equal-credit-opportunity-act-ecoa/?utm_source=chatgpt.com

[18] Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) Why Should I Trust You? *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 1135-1144. https://doi.org/10.1145/2939672.2939778

[19] Niculescu-Mizil, A. and Caruana, R. (2005) Predicting Good Probabilities with Supervised Learning. *Proceedings of the* 22*nd international conference on Machine Learning*, Bonn, 7-11 August 2005, 625-632. https://doi.org/10.1145/1102351.1102430

[20] Platt, J.C. (2000) Probabilities for SV Machines. In: Smola, A.J., Bartlett, P., Schölk-

opf, B. and Schuurmans, D., Eds., *Advances in Large-Margin Classifiers*, The MIT Press, 61-74. https://doi.org/10.7551/mitpress/1113.003.0008

[21] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, **16**, 321-357. https://doi.org/10.1613/jair.953

[22] Lundberg, S.M. and Lee, S.-I. (2017) A Unified Approach to Interpreting Model Predictions. *Annual Conference on Neural Information Processing Systems* 2017, Long Beach, 4-9 December 2017, 4765-4774.

[23] Molnar, C. (2022) Interpretable Machine Learning. 2nd Edition, Shroff/Molnar.

[24] West, D. (2000) Neural Network Credit Scoring Models. *Computers & Operations Research*, **27**, 1131-1152. https://doi.org/10.1016/s0305-0548(99)00149-5

[25] Hand, D.J. (2009) Measuring Classifier Performance: A Coherent Alternative to the Area under the ROC Curve. *Machine Learning*, **77**, 103-123. https://doi.org/10.1007/s10994-009-5119-5

[26] Chang, C. and Lin, C. (2011) LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, **2**, 1-27. https://doi.org/10.1145/1961189.1961199

[27] Van Gestel, T. and Baesens, B. (2009) Credit Risk Management: Basic Concepts. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199545117.001.0001

[28] Heaton, J.B., Polson, N.G. and Witte, J.H. (2017) Deep Learning for Finance: Deep Portfolios. *Applied Stochastic Models in Business and Industry*, **33**, 3-12. https://doi.org/10.1002/asmb.2209

[29] Baesens, B. (2014) Developing Credit Risk Models Using SAS Enterprise Miner. SAS Institute White Paper.

[30] Thomas, L., Crook, J. and Edelman, D. (2017) Credit Scoring and Its Applications. 2nd Edition, Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611974560

[31] Anderson, R. (2007) The Credit Scoring Toolkit. Oxford University Press.

[32] Altman, E.I. (1968) Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, **23**, 589-609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

[33] Khandani, A.E., Kim, A.J. and Lo, A.W. (2010) Consumer Credit-Risk Models via Machine-Learning Algorithms. *Journal of Banking & Finance*, **34**, 2767-2787. https://doi.org/10.1016/j.jbankfin.2010.06.001

[34] Crook, J.N., Edelman, D.B. and Thomas, L.C. (2007) Recent Developments in Consumer Credit Risk Assessment. *European Journal of Operational Research*, **183**, 1447-1465. https://doi.org/10.1016/j.ejor.2006.09.100

[35] Flach, P. (2012) Machine Learning: The Art and Science of Algorithms that Make Sense of Data. Cambridge University Press. https://doi.org/10.1017/cbo9780511973000

[36] Murphy, K.P. (2012) Machine Learning: A Probabilistic Perspective. MIT Press.

[37] Goodfellow, I., Bengio, Y. and Courville, A. (2016) Deep Learning. MIT Press.

[38] Hastie, T. and Tibshirani, R. (1986) Generalized Additive Models. *Statistical Science*, **1**, 297-310. https://doi.org/10.1214/ss/1177013604

[39] Cortes, C. and Vapnik, V. (1995) Support-Vector Networks. *Machine Learning*, **20**, 273-297. https://doi.org/10.1023/a:1022627411411

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

https://doi.org/10.1023/a:1010933404324

[41] Breiman, L. (1996) Bagging Predictors. *Machine Learning*, **24**, 123-140. https://doi.org/10.1023/a:1018054314350

[42] Chen, T. and Guestrin, C. (2016) XGBoost: Extreme Gradient Boosting. https://arxiv.org/abs/1603.02754

[43] Ke, G., Meng, Q., Finley, T., Wang, T.F., Chen, W., Ma, W.D. *et al.* (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 3149-3157.

[44] Friedman, J., Hastie, T. and Tibshirani, R. (2000) Additive Logistic Regression: A Statistical View of Boosting (with Discussion and a Rejoinder by the Authors). *The Annals of Statistics*, **28**, 337-407. https://doi.org/10.1214/aos/1016218223

[45] Finlay, S. (2010) Credit Scoring, Response Modeling, and Insurance Rating. Palgrave Macmillan.

[46] European Banking Authority (2017) Guidelines on PD Estimation, LGD Estimation and Treatment of Defaulted Assets.

[47] Basel Committee on Banking Supervision (2011) Principles for the Sound Management of Operational Risk. BIS.

[48] Financial Stability Board (2017) Artificial Intelligence and Machine Learning in Financial Services.

[49] Bellotti, T. and Crook, J. (2009) Support Vector Machines for Credit Scoring and Discovery of Significant Features. *Expert Systems with Applications*, **36**, 3302-3308. https://doi.org/10.1016/j.eswa.2008.01.005

[50] Bank of England and Financial Conduct Authority (2022) Machine Learning in UK Financial Services.

[51] Brown, I. and Mues, C. (2012) An Experimental Comparison of Classification Algorithms for Imbalanced Credit Scoring Data Sets. *Expert Systems with Applications*, **39**, 3446-3453. https://doi.org/10.1016/j.eswa.2011.09.033

[52] Barocas, S., Hardt, M. and Narayanan, A. (2019) Fairness and Machine Learning. MIT Press.

[53] Hull, J. (2018) Risk Management and Financial Institutions. 5th Edition, Wiley.

[54] Tasche, D. (2006) Validation of Internal Rating Systems and PD Estimates. https://arxiv.org/abs/physics/0606071

[55] James, G., Witten, D., Hastie, T. and Tibshirani, R. (2021) An Introduction to Statistical Learning. 2nd Edition, Springer.

[56] OECD (2019) Artificial Intelligence in Society. OECD Publishing.

[57] Bazarbash, M. (2019) Fintech in Financial Inclusion: Machine Learning Applications in Assessing Credit Risk. IMF Working Papers, Vol. 2019, 1. https://doi.org/10.5089/9781498314428.001

[58] Varma, S. and Simon, R. (2006) Bias in Error Estimation When Using Cross-Validation for Model Selection. *BMC Bioinformatics*, **7**, Article No. 91. https://doi.org/10.1186/1471-2105-7-91

[59] Bishop, C. (2006) Pattern Recognition and Machine Learning. Springer.

[60] International Monetary Fund (2021) Financial Stability Implications of Artificial Intelligence. IMF Policy Paper.

[61] Pazarbasioglu, C., Garcia Mora, A., Uttamchandani, M., Natarajan, H., Feyen, E. and Saal, M. (2020) Digital Financial Services (World Bank White Paper No. 54). World Bank.

https://thedocs.worldbank.org/en/doc/305a39cbb6f35567db78bda6709c5cd8-0430012025/original/World-Bank-DFS-Whitepaper-DigitalFinancial-Services.pdf?utm_source=chatgpt.com

[62] Pearl, J. (2009) Causality: Models, Reasoning and Inference. 2nd Edition, Cambridge University Press. https://doi.org/10.1017/cbo9780511803161