

A Multimetric Evaluation of Ensemble and Neural Models for Imbalanced Credit Scoring on the German Credit Dataset

Rino¹, Aditiya Hermawan², Elfindah Princes³, Satria Abadi⁴, Umarkhon Mukhamedov⁵

^{1,2}Department of Informatics Engineering, Universitas Buddhi Dharma, Tangerang, Indonesia

³University of Manchester, Manchester, United Kingdom

⁴Faculty of Computing and Meta Technology, Sultan Idris Education University, Perak, Malaysia

⁵University of Science and Technology, Tashkent, Uzbekistan

Received:

June 7, 2026

Revised:

July 21, 2026

Accepted:

September 3, 2026

Published:

October 24, 2026

Corresponding Author:

Author Name*:

Rino

Email*:

rino@ubd.ac.id

DOI:

10.63158/journalisi.v8i5.1825

© 2026 Journal of Information Systems and Informatics. This open access article is distributed under a (CC-BY License)



Credit scoring models must balance global discrimination, minority-class detection, probability quality, and computational efficiency, particularly when default cases are less frequent than non-default cases. This study benchmarks Logistic Regression, Random Forest, XGBoost, and a Multilayer Perceptron using the German Credit Dataset, which contains 1,000 borrower records. All models were evaluated using a consistent preprocessing, partitioning, and imbalance-aware training protocol. Performance was assessed through ROC-AUC, PR-AUC, Balanced Accuracy, precision, recall, F1-score, Brier Score, and computational time. The results indicate moderate predictive performance across the evaluated models. XGBoost achieved the highest global discrimination with a ROC-AUC of 0.764, whereas the MLP obtained the highest PR-AUC of 0.378 and the lowest Brier Score of 0.123, indicating the best native probability quality among the evaluated models. No post-hoc probability calibration was applied. However, the MLP produced lower recall at the selected classification threshold. Logistic Regression required the least computation but recorded the lowest ROC-AUC and PR-AUC. Exploratory SHAP analysis was used to identify influential borrower attributes. These benchmark-specific findings demonstrate that model selection depends on the intended decision objective rather than architectural complexity alone.

Keywords: Brier Score, Credit Scoring, Ensemble Learning, Imbalanced Data, PR-AUC, Tabular Credit Risk

1. INTRODUCTION

Credit scoring is a binary classification problem that estimates the probability of borrower default based on heterogeneous tabular financial and demographic attributes. In modern financial institutions, credit scoring systems play a central role in loan approval decisions, credit limit determination, and portfolio risk management. As artificial intelligence becomes increasingly embedded in financial analytics, traditional linear statistical approaches such as logistic regression are progressively being complemented or, in some cases, replaced by machine learning and deep learning models capable of capturing complex nonlinear interactions in structured financial data [1], [2].

Recent literature demonstrates the dominance of ensemble tree-based methods in contemporary tabular credit scoring. Gradient boosting approaches such as XGBoost and LightGBM consistently report superior discriminative performance, particularly when modeling nonlinear feature interactions and heterogeneous financial variables [3], [4]. Random Forest has likewise been widely recognized for its predictive stability and robustness to noise, making it a strong representative of classical machine learning in credit risk modeling [5], [6], [7]. These findings position ensemble learning as a prevailing paradigm in structured credit risk prediction [8].

At the same time, deep learning architectures have gained attention as potential alternatives for modeling higher-order feature interactions. Multilayer Perceptron (MLP) models combined with embedded feature selection have demonstrated promising predictive performance on benchmark datasets such as the German Credit Dataset [9]. More advanced neural architectures incorporating adaptive optimization and regularization techniques have been proposed to enhance minority-class sensitivity in imbalanced credit data [10]. However, empirical evidence also indicates that for small- to medium-sized tabular datasets, deep learning does not consistently outperform well-optimized ensemble boosting methods [11], [12].

This contradiction suggests that algorithmic superiority in credit scoring is contextual rather than universal. Increased model complexity does not automatically translate into improved discriminative ability or generalization stability. In fact, several benchmark studies have reported that deep learning models exhibit greater sensitivity to

hyperparameter configurations and dataset size, potentially reducing robustness in operational financial environments [11], [13]. Consequently, the central challenge in credit scoring is not merely identifying the model with the highest predictive accuracy, but determining the optimal balance among representational capacity, performance stability, and operational feasibility.

The natural asymmetry between non-default and default cases introduces biased learning dynamics when models optimize global accuracy alone. Cost-sensitive learning has been recommended as a more realistic strategy than synthetic resampling techniques because it preserves the original financial data distribution while adjusting the learning objective to reflect asymmetric misclassification costs [14], [12]. Moreover, evaluation under highly imbalanced conditions requires appropriate performance metrics to avoid misleading performance estimates [13].

Credit scoring systems must provide consistent, auditable explanations, particularly in cases involving loan rejection or credit limit adjustments. Explainable machine learning approaches such as SHAP and other interpretability frameworks have been developed to enhance transparency, including in imbalanced credit datasets [13], [16]. However, many studies assess interpretability independently from discriminative performance and stability analysis, resulting in a fragmented understanding of the trade-offs between accuracy and transparency.

Taken together, these observations reveal a significant research gap. Prior studies often emphasize the superiority of either ensemble boosting or deep learning models without integrating evaluation of global discrimination, minority-class sensitivity, probabilistic calibration, computational efficiency, and interpretability within a unified experimental framework. Furthermore, although the German Credit Dataset remains a widely used benchmark in contemporary research, systematic comparative evaluations of Logistic Regression, Random Forest, XGBoost, and MLP under identical preprocessing pipelines and consistent cost-sensitive strategies remain limited [17], [7].

Based on the identified research gap, this study addresses the following research questions: RQ1: Which model achieves the strongest global discrimination in the German

Credit Dataset? RQ2: Which model provides the most favorable minority-class performance when evaluated through PR-AUC, recall, F1-score, and Balanced Accuracy? RQ3: How stable are Logistic Regression, Random Forest, XGBoost, and MLP across repeated runs and controlled ablation settings? RQ4: What trade-offs emerge among discrimination, minority-class performance, probability quality, computational cost, and interpretability?

Unlike prior studies that emphasize isolated performance gains, this research conceptualizes model comparison in imbalanced tabular credit scoring as a multidimensional trade-off analysis. The study integrates ROC-AUC, PR-AUC, Balanced Accuracy, precision, recall, F1-score, Brier Score, repeated-run stability, ablation analysis, and computational efficiency within a controlled experimental framework. Logistic Regression, Random Forest, XGBoost, and Multilayer Perceptron are compared using consistent preprocessing, data partitions, and imbalance-aware training procedures. The novelty of the study lies not in proposing a new algorithm, but in combining these evaluation dimensions within a single benchmark protocol.

In this study, operationally meaningful benefits refer to improvements that support specific analytical objectives, including borrower-risk ranking, identification of high-risk applicants, probability estimation, performance stability, and computational feasibility. The term does not imply verified financial gains or institutional deployment because the study does not use a bank-specific cost matrix or operational loan portfolio.

2. METHODS

This study employed a quantitative experimental design to compare four supervised classification models for imbalanced credit scoring: Logistic Regression (LR), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Multilayer Perceptron (MLP). The methodological objective was to evaluate these models within a unified and consistent pipeline, ensuring that observed performance differences could be attributed to intrinsic model characteristics rather than variations in preprocessing, sampling strategies, or evaluation procedures. The overall research workflow is illustrated in Figure 1, comprising six sequential stages: dataset selection, preprocessing, stratified data partitioning, model development, hyperparameter optimization, and multidimensional performance

evaluation. This structured protocol was designed to guarantee methodological rigor, reproducibility, and fair comparison across all candidate models.

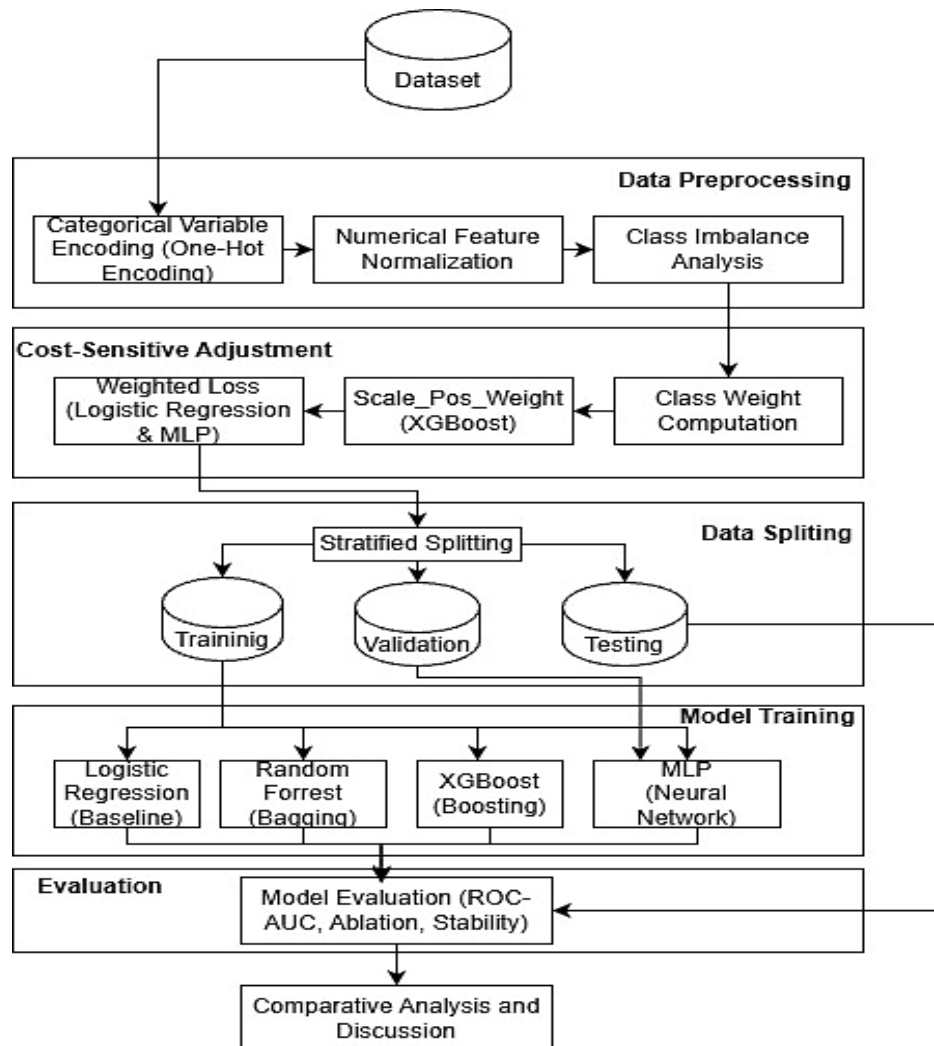


Figure 1: Research Method

2.1. Dataset

This study uses the German Credit Dataset from the UCI Machine Learning Repository, consisting of 1,000 borrower instances and 20 predictive attributes covering numerical and categorical variables, including savings account status, credit duration, loan amount, credit history, employment status, credit purpose, and housing conditions [18]. The binary target classifies applicants into good and bad credit risk groups, while the dataset's mixed tabular structure supports controlled benchmarking of interpretable linear models, ensemble methods, and neural networks. The dataset also reflects moderate class imbalance, with high-risk borrowers forming a smaller proportion than low-risk

borrowers [19], resembling real credit portfolios where default cases are less frequent and classification becomes challenging. To assess robustness, models trained on the German Credit Dataset were also evaluated on an auxiliary dataset with comparable structure and binary labels. This secondary evaluation examined performance degradation under distributional variation, providing insight into model stability, external validity, and deployment relevance.

Tabel 1. German Credit Dataset

Dataset Column	Description
Checking_Account_Status	Status of the applicant's checking account, such as no account, negative balance, low balance, or high balance.
Duration_Months	Requested or current credit duration, measured in months.
Credit_History	Applicant's credit history, including repayment behavior, previous credit records, and the status of existing credit.
Purpose	Intended use of the credit, such as purchasing a vehicle, furniture, education, business, or other needs.
Credit_Amount_DM	Amount of credit requested or granted, measured in Deutsche Mark (DM).
Savings_Account_Bonds	Status of the applicant's savings account or bonds.
Present_Employment_Since	Length of time the applicant has been employed in the current job.
Installment_Rate_Pct	Percentage of the applicant's disposable income allocated to credit installments.
Personal_Status_Sex	Applicant's personal status and sex.
Other_Debtors_Guarantors	Presence of other parties who share responsibility for the credit, such as co-debtors or guarantors.
Present_Residence_Since	Length of time the applicant has lived at the current residence, generally represented as year categories.

Dataset Column	Description
Property	Type of asset or property owned by the applicant, such as real estate, life insurance, a car, or no valuable property.
Age_Years	Applicant's age in years.
Other_Installment_Plans	Existence of other installment plans held by the applicant, for example at a bank, store, or none.
Housing	Applicant's housing status, such as renting, owning a home, or living free of charge.
Existing_Credits_At_Bank	Number of existing credits held by the applicant at the same bank.
Job	Applicant's job category or occupational skill/position level.
People_Liable_For_Maintenance	Number of people financially dependent on the applicant.
Telephone	Availability of a telephone registered in the applicant's name.
Foreign_Worker	Whether the applicant is a foreign worker.
Credit_Risk_Class	Target variable representing the credit-risk class; it generally distinguishes good-credit and bad-credit cases.

2.2. Data Preprocessing

The preprocessing stage addressed mixed measurement scales, heterogeneous feature types, and skewed class distribution in tabular credit data while preserving comparability across model families. Initial data integrity checks identified invalid entries, duplicate records, and extreme outliers that could affect parameter estimation or destabilize gradient-based optimization. Categorical attributes were transformed into numerical form according to semantic structure: nominal variables were encoded using one-hot encoding, whereas ordinal variables used ordinal encoding only when a meaningful ranking existed. This follows contemporary credit risk modeling pipelines that emphasize reliable end-to-end preprocessing under data constraints [20].

Numerical variables were standardized using z-score normalization to stabilize optimization and ensure a controlled benchmarking protocol. Normalization parameters were calculated only from the training subset and then applied to validation and test sets to avoid information leakage. Missing values were handled through simple imputation, using median values for numerical features and the most frequent category for categorical features. The target class distribution was examined because credit scoring involves asymmetric misclassification costs, especially when high-risk borrowers are misclassified [21]. Therefore, preprocessing, imbalance-aware training, and evaluation planning were treated as interconnected stages. This design supports appropriate metric selection under skewed data [13] and improves the reliability of model explanations [22], [23].

Tabel 2. Data Preprocessing For Each Step

Step	Input	Process	Rule/Technique	Output
1	Raw credit dataset	Data integrity checking	Check dimensions, data types, invalid values, duplicates, and extreme observations	Cleaned raw dataset
2	Cleaned dataset	Train-validation-test partition	Stratified split based on target class	Training, validation, and test sets
3	Training data	Missing-value treatment	Median for numerical; mode for categorical	Complete dataset without missing values
4	Categorical features	Categorical encoding	One-hot for nominal; ordinal encoding only for ordered categories	Numerical feature representation
5	Numerical features	Feature standardization	Z-score using training-set mean and SD	Standardized numerical attributes
6	Target variable	Class-distribution analysis	Count and percentage of Good/Bad classes	Imbalance profile
7	Processed features	Final transformation	Apply training-fitted transformations to validation/test data	Model-ready matrices

2.3. Data Partitioning

The dataset was divided into training, validation, and testing subsets using stratified sampling to preserve the original distribution of good and bad credit borrowers. This strategy is crucial in imbalanced classification because slight changes in minority-class prevalence may alter task difficulty, produce unstable estimates, and bias model comparisons [14]. Therefore, data partitioning was treated as part of the evaluation-control design, since validation strategies directly affect the credibility of generalization claims [25].

The data were allocated into 64% training, 16% validation, and 20% testing, corresponding to 640, 160, and 200 instances, respectively. The training subset was used for model fitting and parameter estimation, while classical machine learning models applied stratified cross-validation within the training data for robust hyperparameter selection. The validation subset was mainly used for models requiring training monitoring, particularly MLP with early stopping. The test subset was strictly held out for final evaluation and was never used during training or tuning to prevent information leakage and inflated estimate [16]. Overall, this stratified protocol strengthens empirical validity and supports fair comparison across models [17], [18]

2.4. Model Selection

This study adopts four representative algorithmic paradigms—Logistic Regression (LR), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Multilayer Perceptron (MLP)—to examine trade-offs among structural simplicity, nonlinear expressiveness, and generalization stability in imbalanced credit scoring. These models represent different inductive biases, ranging from linear statistical assumptions to high-capacity nonlinear learning. Recent benchmarking studies show that algorithmic superiority in structured financial data is highly context-dependent, especially in tabular datasets with moderate feature interactions and limited sample sizes [11], [19]. Therefore, model comparison in this study is framed not as identifying one universally superior algorithm, but as assessing how different paradigms perform within a unified imbalance-aware evaluation framework.

Logistic Regression is used as the linear baseline because of its long-standing role in credit risk modeling and its compatibility with regulatory expectations. It applies a linear

decision boundary, which generally produces low variance but may introduce bias when nonlinear relationships exist. Despite this limitation, LR remains competitive in several credit scoring benchmarks and is still widely used in operational environments because of its robustness and transparency [10], [11]. Its probabilistic outputs and coefficient-based explanations also make it useful for decision-support systems where accountability and explainability are essential [23].

Random Forest is selected as a bagging-based ensemble model that improves prediction by combining multiple decision trees trained on bootstrapped samples. This mechanism reduces variance and supports robustness in noisy or moderately complex tabular datasets. RF can also capture nonlinear feature interactions without extensive feature engineering, making it suitable for credit scoring tasks with implicit interaction effects. However, although RF is generally robust, it may face limitations under severe class imbalance and therefore requires careful tuning of class weights or sampling-related strategies to preserve minority-class sensitivity.

XGBoost represents the boosting-based ensemble approach and is widely recognized for strong performance in structured tabular data through sequential error correction and regularized optimization [22], [12]. Unlike bagging, boosting focuses iteratively on difficult-to-classify cases, enabling sharper decision boundaries and stronger discrimination. Its regularization mechanisms help balance bias and variance, especially in datasets with moderate complexity and limited observations. Including both RF and XGBoost allows the study to compare bagging and boosting strategies, offering insight into how different ensemble mechanisms influence robustness under imbalanced conditions.

MLP is included as the neural network representative to evaluate whether high-capacity nonlinear modeling improves credit scoring performance. Through hidden layers and nonlinear activation functions, MLP can approximate complex decision boundaries beyond linear or tree-based models. However, this flexibility increases sensitivity to feature scaling, sample size, and hyperparameter configuration, which may affect stability. Prior studies show that deep learning does not consistently outperform ensemble methods in moderate-sized tabular financial datasets [11], [19]. Thus, MLP is examined critically to determine whether greater model capacity improves discrimination, calibration, and robustness under realistic imbalance conditions.

2.5. Imbalance-Aware Training Strategy

Given that the bad-credit class constitutes the minority class and that credit scoring inherently involves asymmetric misclassification costs—where failing to identify high-risk borrowers is substantially more consequential than misclassifying low-risk applicants—this study adopts a cost-sensitive learning strategy rather than synthetic resampling. Cost-sensitive approaches are widely recognized as more appropriate in financial risk modeling because they allow the optimization process to reflect decision-theoretic priorities without altering the original data distribution, thereby preserving real-world class prevalence and avoiding potential sampling bias [17], [14]. This design choice is also consistent with recent methodological discussions emphasizing that imbalance handling should be integrated into the learning objective rather than treated as a preprocessing artifact, particularly in regulated domains where model transparency and data fidelity are critical [18], [25]. Importantly, this study frames imbalance handling as part of the model optimization design rather than as an external data manipulation step, thereby maintaining consistency between the learning objective and real-world decision constraints, which in turn strengthens the validity of model comparisons under imbalanced conditions.

To ensure methodological coherence across different algorithmic paradigms, imbalance-aware adaptations were implemented in accordance with each model's native optimization mechanism. Logistic Regression and Random Forest incorporated class weighting to penalize minority-class misclassification more heavily, while XGBoost utilized the `scale_pos_weight` parameter to rebalance gradient contributions and improve sensitivity to rare events. The Multilayer Perceptron (MLP), in contrast, employed focal loss to dynamically emphasize hard-to-classify minority instances, thereby enhancing learning under skewed distributions. The focal-loss function is defined as shown in Equation (1):

$$FL(pt) = -\alpha_t(1 - pt)\gamma \log(pt) \quad (1)$$

where p_t denotes the predicted probability assigned to the true class, α_t is the class-balancing factor, and γ is the focusing parameter. The parameter γ controls the extent to which well-classified observations are down-weighted, whereas α_t controls the relative contribution of the minority class. In this study, the focal-loss parameters were set to $\alpha = [value]$ and $\gamma = [value]$. It is important to note that these strategies are not

strictly equivalent in terms of loss formulation; rather, they represent model-specific implementations of imbalance-aware learning tailored to each algorithmic family.

For XGBoost, `scale_pos_weight` was initialized using the ratio of negative to positive observations in the dataset used in this experiment. Good credit was coded as the negative class and bad credit as the positive class. The dataset contained 845 good-credit and 155 bad-credit observations; therefore:

$$\text{scale_pos_weight} = \frac{845}{155} = 5.4516 \quad (2)$$

This value increased the contribution of errors involving bad-credit observations during gradient optimization. It should be noted that the ratio differs from the commonly reported 700:300 distribution of the original UCI German Credit Dataset because the dataset version used in this experiment contained 845 good-credit and 155 bad-credit records after the applied data preparation and label construction procedure.

Accordingly, the objective of this study is not to enforce identical optimization functions across models, but to compare each model under its most appropriate and practically relevant imbalance-aware configuration. This design choice reflects a best-practice approach that balances methodological fairness with practical relevance, while ensuring that each model operates under conditions that are consistent with its intended use. Consequently, the proposed framework maintains comparability across models while aligning optimization strategies with the asymmetric risk structure inherent in credit decision-making, ultimately supporting more reliable identification of high-risk borrowers [22], [16].

2.6. Hyperparameter Tuning

Hyperparameter optimization was conducted using validation data to identify model configurations that best capture minority-class patterns. For classical machine learning models, hyperparameter tuning was performed using stratified k-fold cross-validation within the training data ($k = 5$) to ensure class balance across folds and reduce variance in performance estimation, while validation monitoring was applied for models requiring iterative training. The best configuration for each model was selected based on validation

PR-AUC, as this metric is more informative than ROC-AUC in imbalanced classification settings where accurate detection of minority-class instances is critical.

A predefined hyperparameter search space was established for each model to ensure a systematic and reproducible optimization process. Randomized search was performed using five-fold stratified cross-validation within the training subset. Logistic Regression was evaluated using 20 randomly sampled candidate configurations, whereas Random Forest and XGBoost were each evaluated using 15 candidate configurations. For each model, the configuration achieving the highest mean cross-validated PR-AUC was selected as the final hyperparameter setting. The search spaces and numbers of randomized-search iterations are presented in Table 3.

Table 3. Hyperparameter Search Spaces for the Evaluated Models

Model	Hyperparameter	Search Range or Candidate Values	Final Selected Value
Logistic Regression	penalty	['l1', 'l2']	l2
	C	Log-uniform distribution from 0.001 to 10	0.0429
	solver	['liblinear', 'saga']	liblinear
	class_weight	[None, 'balanced']	'balanced'
Random Forest	n_estimators	Integer values from 100 to 500	446
	max_depth	[3, 5, 7, 10, 15, None]	5
	min_samples_split	Integer values from 2 to 20	4
	min_samples_leaf	Integer values from 1 to 15	10
	max_features	['sqrt', 'log2', 0.5, 0.75]	0.75
	bootstrap	[True, False]	True
	class_weight	[None, 'balanced', 'balanced_subsample']	'balanced'
	n_estimators	Integer values from 200 to 1,000	817
XGBoost	learning_rate	Log-uniform distribution from 0.001 to 0.1	0.0026
	max_depth	Integer values from 2 to 8	5
	subsample	Continuous values from 0.60 to 1.00	0.6645

Model	Hyperparameter	Search Range or Candidate Values	Final Selected Value
MLP	colsample_bytree	Continuous values from 0.60 to 1.00	0.8782
	min_child_weight	Integer values from 1 to 10	4
	gamma	Continuous values from 0.00 to 0.50	0.2787
	reg_alpha	Log-uniform distribution from 0.0001 to 1.0	0.1540
	reg_lambda	Log-uniform distribution from 0.01 to 10	1.8039
	scale_pos_weight	Continuous values from 1.0 to 6.0	5.4516*
	Number of hidden layers	[1, 2, 3]	2
	Neurons per hidden layer	[(64, 32), (128, 64), (256, 128), (128, 128, 64), (256, 128, 64)]	(256, 128)
	Activation function	ReLU	ReLU
	Dropout rate	[0.1, 0.2, 0.3, 0.4, 0.5]	0.30
	Batch size	[16, 32, 64]	16
	Learning rate	[0.0001, 0.0003, 0.001, 0.003]	0.001
	Weight decay	[1e-5, 3e-5, 1e-4, 3e-4, 1e-3]	1e-4
	Focal-loss alpha, (α)	[0.15, 0.25, 0.35, 0.50]	0.15
	Focal-loss gamma, (γ)	[1.0, 2.0, 3.0]	1.0
MLP	Maximum training epochs	Fixed maximum during randomized search	250
	Early-stopping patience	Fixed during randomized search	18
	Learning-rate reduction patience	Fixed during randomized search	8
	Probability-output method	Sigmoid output or post-hoc calibration	Direct sigmoid output

For the final MLP configuration, focal loss was parameterized using $\alpha=0.15$ and $\gamma=1.0$. These values were selected from 25 randomized-search trials based on the highest validation PR-AUC. The selected architecture consisted of two hidden layers containing 256 and 128 neurons, respectively, with ReLU activation, a dropout rate of 0.30, an L2 regularization coefficient of 1×10^{-4} , a learning rate of 0.001, and a batch size of 16. The network was trained for a maximum of 250 epochs, while early stopping terminated training when validation PR-AUC did not improve for 18 consecutive epochs. A learning-rate reduction mechanism with a patience of eight epochs was also applied. Predicted probabilities were obtained directly from the sigmoid output layer, and no post-hoc probability-calibration method was used. Therefore, the reported Brier Score evaluates the quality of the model's native probability estimates rather than the performance of an explicitly calibrated model.

For Logistic Regression, Random Forest, and XGBoost, hyperparameter optimization was conducted using randomized search with five-fold stratified cross-validation. The optimization criterion was the mean validation PR-AUC because it provides a more informative assessment of minority-class performance under class imbalance. For the MLP, candidate configurations were evaluated using the validation subset, while early stopping was applied to prevent overfitting. The final hyperparameters selected through this procedure are reported in Table 4.

Table 4: Final optimized hyperparameters for each model

Model	Hyperparameter	Final Value
Logistic Regression	penalty	l2
	C	0.0429
	solver	liblinear
	class_weight	balanced
Random Forest	n_estimators	446
	max_depth	5
	min_samples_split	4
	min_samples_leaf	10
	bootstrap	True
	class_weight	balanced
XGBoost	n_estimators	817

Model	Hyperparameter	Final Value
MLP	learning_rate	0.0026
	max_depth	5
	scale_pos_weight	5.4516
	objective	binary:logistic
	eval_metric	logloss
	hidden layers	2
	neurons per layer	(256, 128)
	activation	ReLU
	dropout rate	0.3
	batch size	16
	learning rate	0.001
	weight decay (L2)	1e-4
	loss function	Focal loss

A predefined search space was specified for each model, and hyperparameter search was conducted using randomized search within this space. The tuning space included the following candidate parameters: Logistic Regression (penalty type, inverse regularization strength (C), and solver); Random Forest (number of trees, maximum tree depth, minimum samples per split, minimum samples per leaf, and maximum features); XGBoost (number of estimators, learning rate, maximum depth, subsample ratio, column subsample ratio, minimum child weight, and regularization parameters); and Multilayer Perceptron (number of hidden layers, neurons per layer, activation function, dropout rate, batch size, learning rate, weight decay, and number of training epochs).

The final optimized hyperparameters selected for each model are summarized in Table 1. These configurations correspond to the best-performing models identified during the tuning stage and were subsequently used for the final training and evaluation experiments.

2.7. Training Procedure

Following preprocessing and stratified partitioning, all models were trained only on the training subset, while hyperparameter optimization and model selection used the validation subset [13]. This separation reduces optimistic bias and supports reproducibility

when comparing models with different representational capacities [19]. Because credit data are imbalanced, training incorporated imbalance-aware mechanisms to reduce majority-class bias and reflect asymmetric misclassification costs in credit risk decisions [22], [13]. Logistic Regression was trained using penalized likelihood optimization, where regularization controls the bias–variance trade-off and limits overfitting in heterogeneous financial data. Random Forest and XGBoost were configured as nonlinear ensemble learners: RF controlled variance through tree aggregation, depth, and feature subsampling, while XGBoost used sequential boosting, shrinkage, and regularization to improve discrimination under cost-sensitive conditions.

For the classical machine-learning models, five-fold stratified cross-validation was performed exclusively within the 640-instance training subset. The external validation subset was not merged with the training data during hyperparameter search or final model fitting. For the MLP, the same 640-instance training subset was used for gradient-based optimization, while the 160-instance validation subset was used only to monitor training and trigger early stopping. Therefore, differences in model performance cannot be attributed to unequal training-set sizes.

The Multilayer Perceptron used multilayer nonlinear transformations to increase representational flexibility, but required careful hyperparameter control because neural models are sensitive to scaling, sample size, and regularization [11], [19]. Weight penalties, dropout, and early stopping were applied to improve stability and prevent overfitting. All models followed a unified validation protocol to ensure fair comparison across linear, ensemble, and neural paradigms. The experiments were implemented in Python 3.12.13 using scikit-learn 1.6.1, XGBoost 3.3.0, PyTorch 2.11.0+cpu, NumPy 2.0.2, pandas 2.2.2, and SHAP 0.52.0. All experiments were executed in Google Colab using an Intel Xeon CPU environment. Random seeds were fixed at 42 for data partitioning, model initialization, and stochastic training components.

2.8. Evaluation Metrics

Model evaluation followed a multidimensional protocol integrating discriminative assessment, stability analysis, and ablation-style sensitivity testing to ensure reliable conclusions under imbalanced credit-risk conditions. Rather than relying on single-point accuracy estimates, evaluation was framed as an approximation of out-of-sample

behavior, emphasizing metric interpretability and robustness across repeated experimental settings, particularly given the risk of misleading summaries in highly skewed data [15]. ROC-AUC was used as the primary discrimination metric to quantify threshold-independent separation between high- and low-risk borrowers; however, its interpretation was complemented by considerations of imbalance to avoid overreliance on aggregate ranking measures. Robustness was assessed through repeated runs with controlled perturbations and summarized via ROC-AUC dispersion to capture sensitivity to sampling and initialization. Additionally, ablation experiments systematically modified key components such as cost-sensitive weighting and regularization to attribute performance shifts to structural design choices, recognizing the central role of asymmetric cost modeling in imbalanced credit scoring [22]. This comprehensive evaluation design aligns with evidence that performance on structured/tabular data can vary significantly across algorithms and experimental configurations.

Threshold-independent metrics, including ROC-AUC and PR-AUC, were calculated directly from predicted probabilities. Threshold-dependent metrics, including Accuracy, Precision, Recall, F1-score, and Balanced Accuracy, were calculated after converting predicted probabilities into class labels. Logistic Regression, Random Forest, and XGBoost used a decision threshold of 0.50. For the MLP, the threshold was selected exclusively from the validation set and was set to 0.32 because this value produced the most favorable combination of F1-score and Balanced Accuracy. The test set was not used for threshold selection

2.9. Robustness and Ablation Analysis

Robustness was evaluated through a two-stage analytical framework designed to capture both stochastic variability and structural sensitivity of the models. First, repeated-run evaluation was conducted by varying random seeds to quantify the variability of performance metrics across independent training instances. For each model, the mean and standard deviation of ROC-AUC, PR-AUC, Balanced Accuracy, and Brier Score were computed, enabling a distributional assessment of model performance rather than reliance on single-point estimates. Lower dispersion in these metrics was interpreted as an indication of greater stability and robustness, consistent with methodological recommendations emphasizing the importance of variability-aware evaluation in machine learning experiments [25], [16].

Second, an ablation study was performed to investigate the contribution of key design components to overall model performance and stability. For ensemble-based models, ablation involved removing imbalance-aware mechanisms such as class weighting and relaxing regularization constraints to assess their impact on predictive consistency. In the case of the Multilayer Perceptron (MLP), ablation was conducted by replacing the focal loss function with standard binary cross-entropy and disabling regularization techniques, including dropout and weight decay. The performance of each ablated configuration was evaluated using the same experimental protocol and metrics, allowing for a controlled comparison with the full model setup. This analysis was intended to disentangle the relative influence of model architecture and training strategies, thereby clarifying whether observed performance gains were driven primarily by structural model capacity or by imbalance-aware optimization and regularization design choices [22], [14].

3. RESULTS AND DISCUSSION

3.1. Performance Evaluation

This section presents the empirical findings of the comparative evaluation, focusing on the predictive performance, stability, and calibration characteristics of the four modeling paradigms under a unified experimental framework. The analysis is structured to first examine quantitative performance outcomes, followed by deeper interpretation of model behavior across multiple evaluation dimensions relevant to imbalanced credit scoring. By integrating discrimination, classification, and calibration metrics, the results aim to provide a comprehensive assessment of model suitability rather than relying on a single performance indicator.

3.1.1. Quantitative Performance Comparison

Table 4 reports the out-of-sample predictive performance of the four evaluated models Logistic Regression, Random Forest, XGBoost, and Multilayer Perceptron (MLP) measured on the hold-out test set. The evaluation framework incorporates threshold-independent discrimination metrics (ROC-AUC and PR-AUC), threshold-dependent classification metrics (Accuracy, Balanced Accuracy, Precision, Recall, and F1-score), and a probabilistic calibration metric (Brier Score). All performance indicators were computed using identical test partitions to ensure methodological consistency and comparability across modeling paradigms.

For threshold-independent discrimination, XGBoost achieved the highest ROC-AUC (0.7643), followed by Random Forest (0.7404), MLP (0.7173), and Logistic Regression (0.7007). The differences in ROC-AUC reflect variations in ranking performance across all possible decision thresholds for the positive and negative classes. In contrast, PR-AUC values indicate that MLP attained the highest minority-class precision-recall trade-off (0.3785), while Logistic Regression recorded the lowest PR-AUC (0.2886).

Table 5: Test Set Performance Comparison

Model	Accuracy	ROC-AUC	PR-AUC	Balanced Accuracy	Precision	Recall	F1-score	Brier Score
Logistic Regression	0.72	0.7	0.2	0.689	0.3	0.6	0.41	0.2
	0	00	89		08	45	7	01
Random Forest	0.71	0.7	0.3	0.699	0.3	0.6	0.4	0.17
	5	40	66		09	78	24	9
XGBoost	0.74	0.7	0.3	0.701	0.3	0.6	0.4	0.1
	0	64	52		28	45	35	62
MLP	0.8	0.71	0.37	0.671	0.3	0.4	0.4	0.1
	00	7	8		85	84	29	23

Balanced Accuracy values range between 0.6709 and 0.7013, with XGBoost producing the highest score among the evaluated models. Accuracy values vary from 0.7150 to 0.8000, with MLP yielding the highest overall classification accuracy on the test set. The Brier Score, which quantifies the mean squared difference between predicted probabilities and observed outcomes, is lowest for MLP (0.1233), followed by XGBoost (0.1623), Random Forest (0.1791), and Logistic Regression (0.2010), indicating differences in probabilistic calibration quality across models.

Table 6. Class-Wise Classification Performance

Model	Class	Precision	Recall	F1-score	Support
Logistic Regression	Good credit	0.9185	0.7337	0.8158	169
Logistic Regression	Bad credit	0.3077	0.6452	0.4167	31
Random Forest	Good credit	0.9242	0.7219	0.8106	169
Random Forest	Bad credit	0.3088	0.6774	0.4242	31

Model	Class	Precision	Recall	F1-score	Support
XGBoost	Good credit	0.9209	0.7574	0.8312	169
XGBoost	Bad credit	0.3279	0.6452	0.4348	31
MLP	Good credit	0.9006	0.8580	0.8788	169
MLP	Bad credit	0.3846	0.4839	0.4286	31

3.1.2. ROC Curve Analysis

The discrimination performance of all evaluated models across the complete range of classification thresholds is illustrated using Receiver Operating Characteristic (ROC) curves. ROC analysis provides a threshold-independent visualization of the trade-off between the true positive rate (sensitivity) and the false positive rate ($1 - \text{specificity}$), thereby enabling ranking performance assessment without reliance on a fixed decision cutoff. This representation facilitates direct comparison of separability among models under identical test-set conditions.

The ROC curves for the four models are presented in Figure 2, where panel (a) corresponds to Logistic Regression, panel (b) to Random Forest, panel (c) to XGBoost, and panel (d) to Multilayer Perceptron (MLP). Logistic Regression achieves a ROC-AUC of 0.7007; Random Forest yields 0.7404; XGBoost attains the highest ROC-AUC of 0.7643; and MLP records 0.7173. The relative positioning of the curves across the ROC space visually reflects these quantitative differences, with XGBoost exhibiting the greatest area under the curve among the evaluated approaches.

Across the full threshold spectrum, the XGBoost curve remains closest to the upper-left corner of the ROC space, indicating the highest overall discrimination. Random Forest demonstrates intermediate performance, while MLP lies between Random Forest and Logistic Regression for most threshold intervals. Logistic Regression consistently occupies the lowest position in the ROC space, corresponding to its lower ROC-AUC value reported in Table 2. The graphical ordering of the curves is numerically consistent with the discrimination metrics presented in the quantitative comparison section.

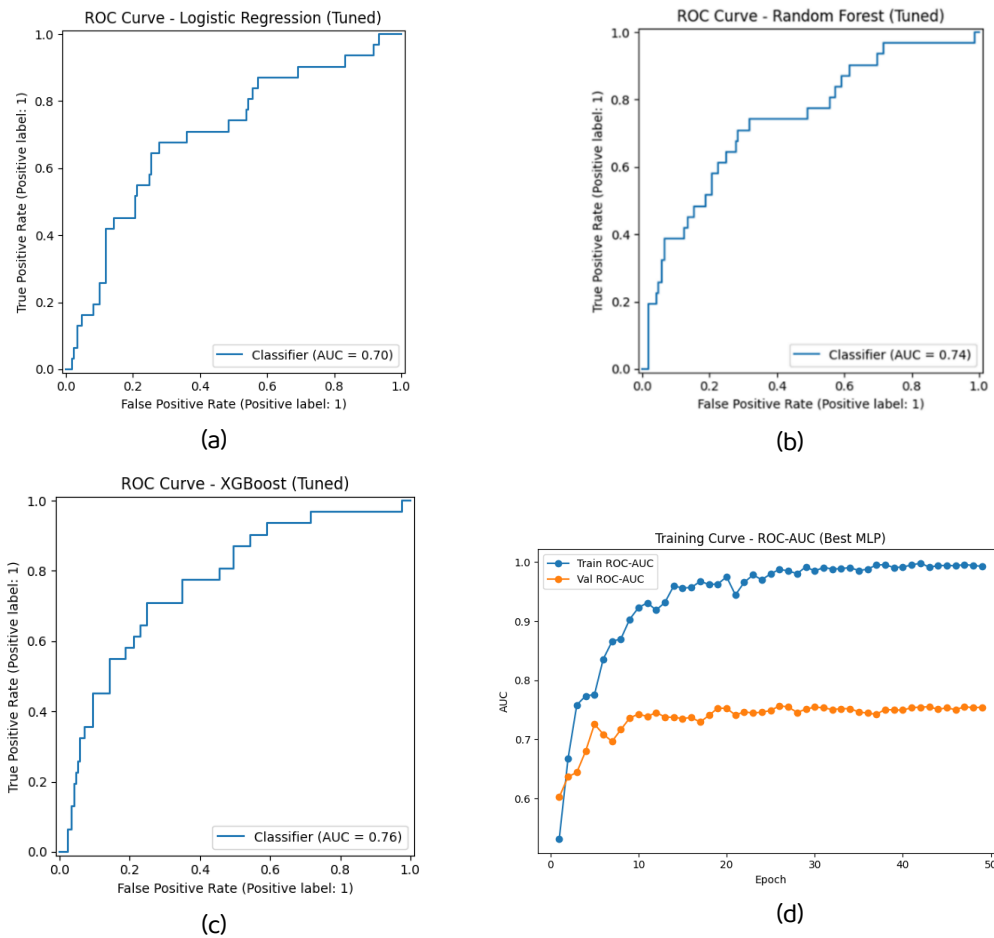


Figure 2: ROC Curve of (a) Logistic Regression, (b) Random Forest, (c) XGBoost, (d) MLP

3.1.3. Precision Recall Analysis

Performance on the minority class (bad credit) was further evaluated using Precision–Recall (PR) analysis, which provides complementary insight to ROC-based evaluation under imbalanced conditions. Unlike ROC curves, PR curves are more sensitive to changes in class distribution and therefore offer a more informative assessment of model effectiveness in identifying rare but critical instances. By explicitly capturing the trade-off between positive predictive value (precision) and sensitivity (recall), PR analysis directly reflects the quality of minority-class predictions. The overlaid PR curves for all evaluated models are presented in Figure 3, enabling a consistent and visually interpretable comparison under identical test-set conditions.

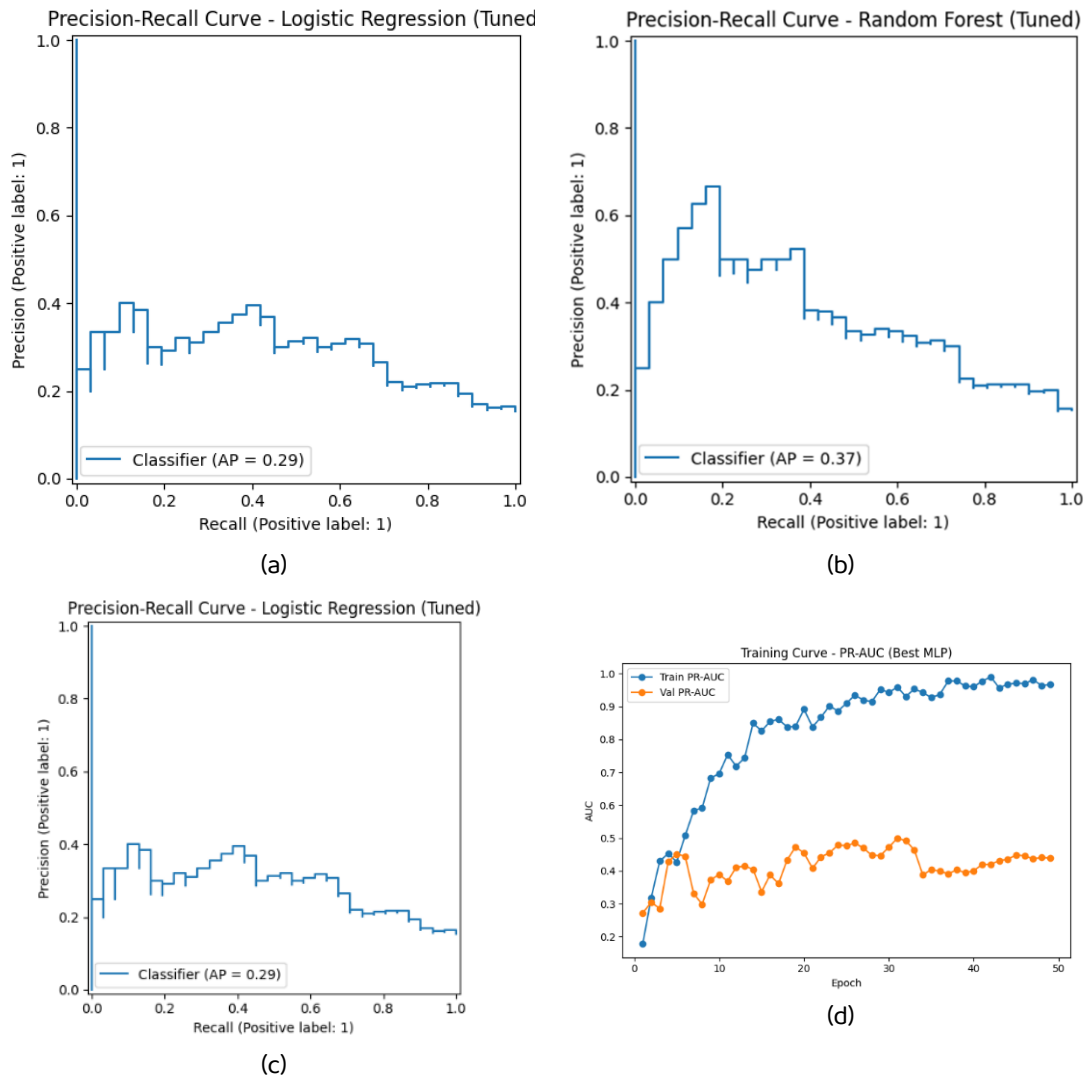


Figure 3: Precision–Recall Curves of (a) Logistic Regression, (b) Random Forest, (c) XGBoost, (d) MLP

The PR-AUC values obtained on the test set were 0.2886 for Logistic Regression, 0.3657 for Random Forest, 0.3519 for XGBoost, and 0.3785 for MLP. These results indicate meaningful variation in minority-class detection performance across different modeling paradigms, highlighting the limitations of relying solely on global discrimination metrics such as ROC-AUC. Notably, the Multilayer Perceptron (MLP) achieves the highest PR-AUC, suggesting superior capability in capturing minority-class patterns under imbalance-aware training conditions. In contrast, Logistic Regression records the lowest PR-AUC, reflecting the limitations of linear decision boundaries in modeling complex relationships associated with rare default events.

Across the recall spectrum, the PR curves exhibit distinct precision-decay patterns as recall increases. The MLP curve maintains relatively higher precision at lower recall levels than the other models, whereas Random Forest and XGBoost exhibit competitive precision retention across intermediate recall levels. Logistic Regression shows a more pronounced decline in precision as recall increases. The graphical ordering of the curves corresponds to the PR-AUC values reported in Table 2, providing a consistent visual and quantitative representation of minority-class discrimination performance.

3.1.4. Stability and Model Complexity Analysis

Beyond discrimination metrics, additional analyses were conducted to examine performance stability and computational complexity across models under identical experimental conditions. The comparative predictive results reported in Table 3 indicate that XGBoost achieved the highest ROC-AUC (0.7643), whereas MLP attained the highest PR-AUC (0.3785). Random Forest demonstrated intermediate discrimination performance, whereas Logistic Regression yielded the lowest ROC-AUC (0.7007). These numerical differences reflect differences in how the models handle nonlinear interactions and the imbalanced class distribution observed in the dataset.

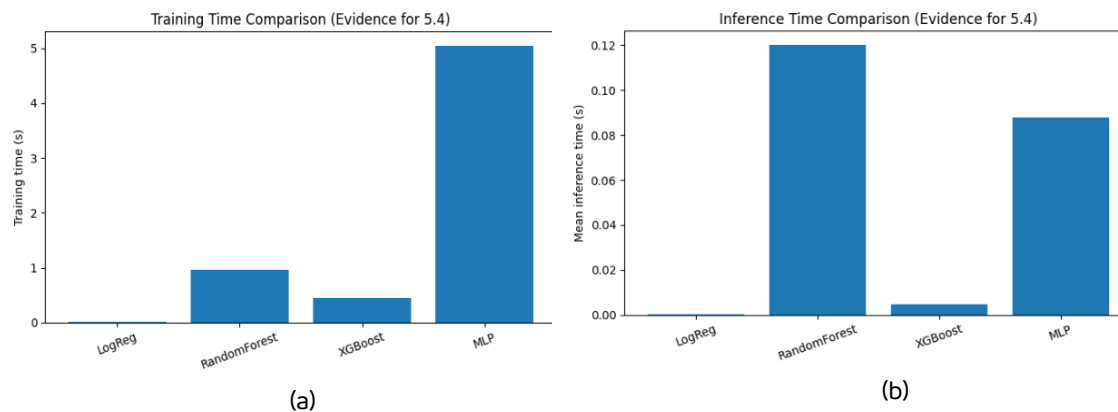
Stability was evaluated through repeated training procedures with controlled variation in initialization and data partitioning. Across repeated runs, ensemble-based approaches (Random Forest and XGBoost) exhibited lower variance in ROC-AUC than MLP, indicating more consistent convergence under stochastic conditions. Logistic Regression showed stable performance across repetitions due to its convex optimization structure. The consistency patterns observed across runs reinforce the robustness comparison among model families under the same evaluation protocol.

Computational performance metrics are summarized in Table 7, which reports the training time and mean inference time for each algorithm. Logistic Regression recorded the shortest training time (0.005928 seconds) and the lowest inference latency (0.000209 seconds), reflecting its low computational complexity. XGBoost completed training in 0.452253 seconds, while Random Forest required 0.958425 seconds. MLP had the highest training time at 5.046373 seconds, consistent with iterative epoch-based optimization and hyperparameter tuning.

Table 7: Training and Inference Time Comparison

Model	Data Split	Train Size	Training Time (s)	Mean Inference Time (s)
Logistic Regression	(64:16):20	800	0.005928	0.000209
XGBoost	(64:16):20	800	0.452253	0.004733
Random Forest	(64:16):20	800	0.958425	0.120279
MLP	(64:16):20	640	5.046373	0.087921

The comparative visualization of training duration is shown in Figure 4(a), highlighting the relative computational differences across models. The inference latency comparison is shown in Figure 4(b), with Logistic Regression having the lowest runtime overhead, followed by XGBoost and Random Forest. In contrast, MLP has a higher inference cost. The scalability trend with increasing dataset size is depicted in Figure 5, where all models exhibit monotonic growth in training time. The growth rate is steepest for MLP and most gradual for Logistic Regression, while ensemble methods maintain intermediate scaling behavior.

**Figure 4:** Comparison Across Models (a) Training Time, (b) Inference Time

Overall, the empirical results demonstrate measurable trade-offs between predictive performance, stability, and computational cost. Ensemble methods provide strong discrimination performance with moderate computational overhead, whereas neural networks incur higher training complexity. Linear models achieve superior computational efficiency but comparatively lower discrimination metrics. These findings provide

quantitative evidence supporting comparative evaluation of modeling paradigms under structured, imbalanced credit-risk conditions.

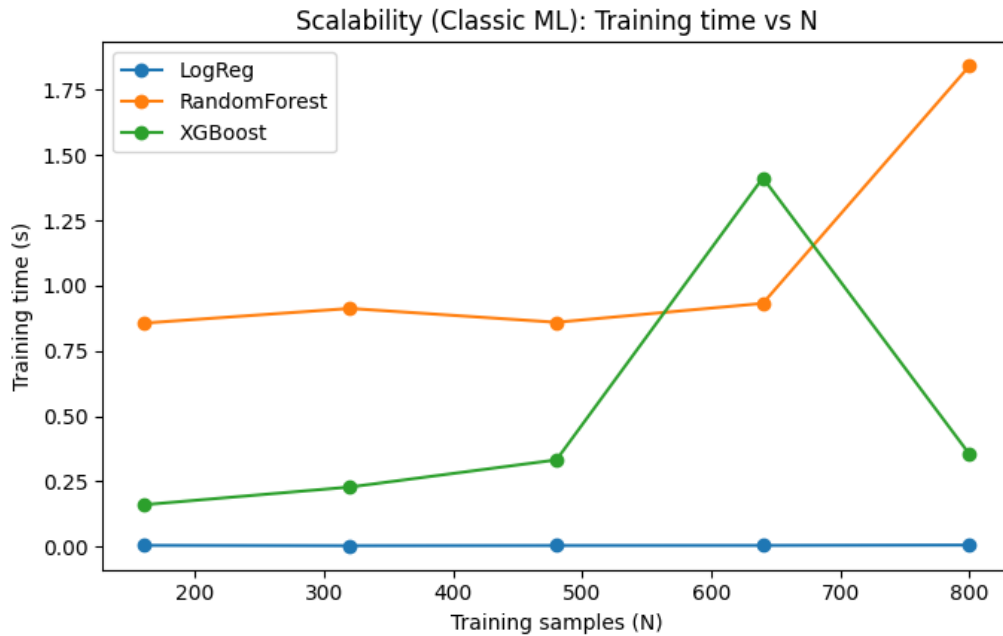


Figure 5: Training Time Scalability with Increasing Dataset Size

3.1.5. Model Robustness Evaluation (Ablation Study, Parameter Sensitivity, and Explainability)

The ablation experiments demonstrate that the effects of imbalance-aware mechanisms and regularization were model- and metric-dependent rather than uniformly beneficial. The default Random Forest configuration achieved the highest ROC-AUC among the evaluated ablation settings, while XGBoost with reduced regularization slightly outperformed the tuned `scale_pos_weight` configuration. For the MLP, selected focal-loss configurations achieved relatively high PR-AUC values, although the effects of dropout, L2 regularization, and class weighting were not monotonic across configurations. Therefore, the ablation results do not support the conclusion that removing imbalance-aware mechanisms consistently reduces predictive performance. Instead, their contribution depends on model architecture, hyperparameter configuration, and the evaluation metric considered.

Table 8. Ablation Results Based on ROC-AUC

Rank	Configuration	ROC-AUC
1	RF_Default	0.772

Rank	Configuration	ROC-AUC
2	XGB_LessReg	0.766
3	XGB_Tuned_SPW	0.760
4	XGB_No_SPW	0.755
5	RF_Tuned_CW	0.755
6	RF_Tuned_NoCW	0.749
7	XGB_Default	0.746
8	MLP_Focal_Drop0.4_L2_No_CW_Yes	0.735
9	MLP_BCE_Drop0.0_L2_No_CW_No	0.735
10	MLP_Focal_Drop0.4_L2_Yes_CW_No	0.721
11	MLP_Focal_Drop0.0_L2_Yes_CW_Yes	0.718
12	MLP_Focal_Drop0.4_L2_Yes_CW_Yes	0.718

Table 8 presents the ROC-AUC results of the ablation experiments. The default Random Forest configuration achieved the highest ROC-AUC of approximately 0.772, followed by XGBoost with reduced regularization at approximately 0.766. The tuned XGBoost configuration using `scale_pos_weight` achieved a ROC-AUC of approximately 0.760, slightly exceeding the configurations without `scale_pos_weight` and the class-weighted Random Forest. The MLP configurations produced lower ROC-AUC values, ranging from approximately 0.718 to 0.735. These results indicate that the contribution of imbalance-aware mechanisms and regularization was model-dependent. For the MLP, neither the combined use of focal loss, dropout, L2 regularization, and class weighting nor the removal of individual components consistently improved global discrimination.

Table 9. Ablation Results Based on PR-AUC

Rank	Configuration	PR-AUC
1	MLP_Focal_Drop0.4_L2_No_CW_Yes	≈ 0.407
2	MLP_Focal_Drop0.0_L2_Yes_CW_Yes	≈ 0.395
3	XGB_LessReg	≈ 0.381
4	RF_Default	≈ 0.369
5	RF_Tuned_CW	≈ 0.351
6	XGB_No_SPW	≈ 0.351
7	MLP_Focal_Drop0.4_L2_Yes_CW_Yes	≈ 0.348

Rank	Configuration	PR-AUC
8	XGB_Tuned_SPW	≈ 0.346
9	MLP_BCE_Drop0.0_L2_No_CW_No	≈ 0.341
10	XGB_Default	≈ 0.338
11	RF_Tuned_NoCW	≈ 0.337
12	MLP_Focal_Drop0.4_L2_Yes_CW_No	≈ 0.334

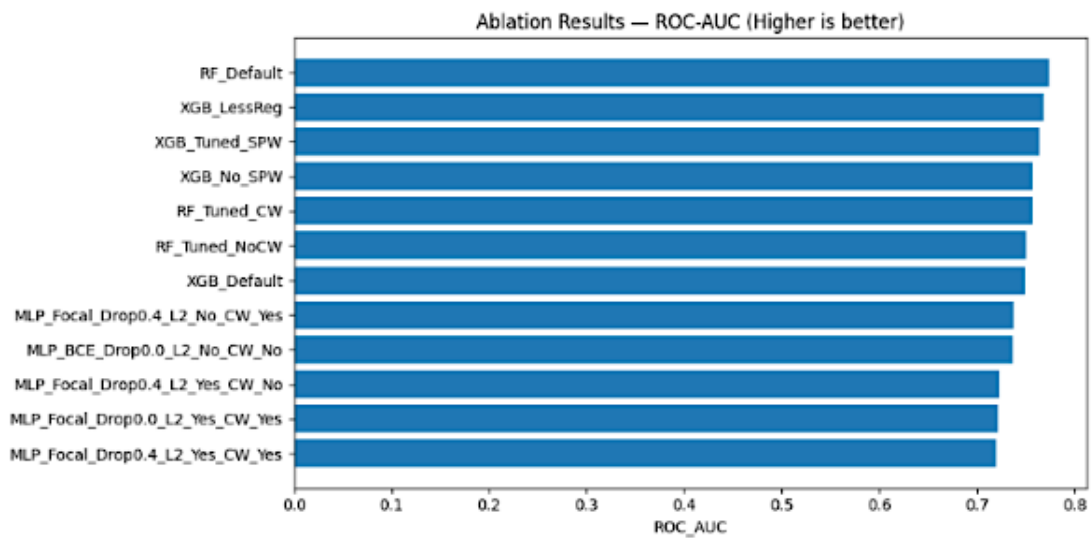
Table 9 presents the PR-AUC results of the ablation analysis. The MLP configuration using focal loss, a dropout rate of 0.4, no L2 regularization, and class weighting achieved the highest PR-AUC of approximately 0.407. The second-best result was obtained by the MLP configuration using focal loss, no dropout, L2 regularization, and class weighting, with a PR-AUC of approximately 0.395. These findings indicate that focal loss combined with class weighting contributed positively to the aggregate precision–recall trade-off, although the effect of dropout and L2 regularization was not uniform across configurations.

Among the ensemble models, XGBoost with reduced regularization achieved the highest PR-AUC of approximately 0.381, followed by the default Random Forest at approximately 0.369. In contrast, tuned scale_pos_weight did not produce the strongest XGBoost PR-AUC, suggesting that increasing minority-class gradient weighting did not automatically improve precision–recall performance. The results further demonstrate that the relative ranking of model configurations differed between ROC-AUC and PR-AUC, reinforcing the importance of evaluating both global discrimination and minority-focused performance.

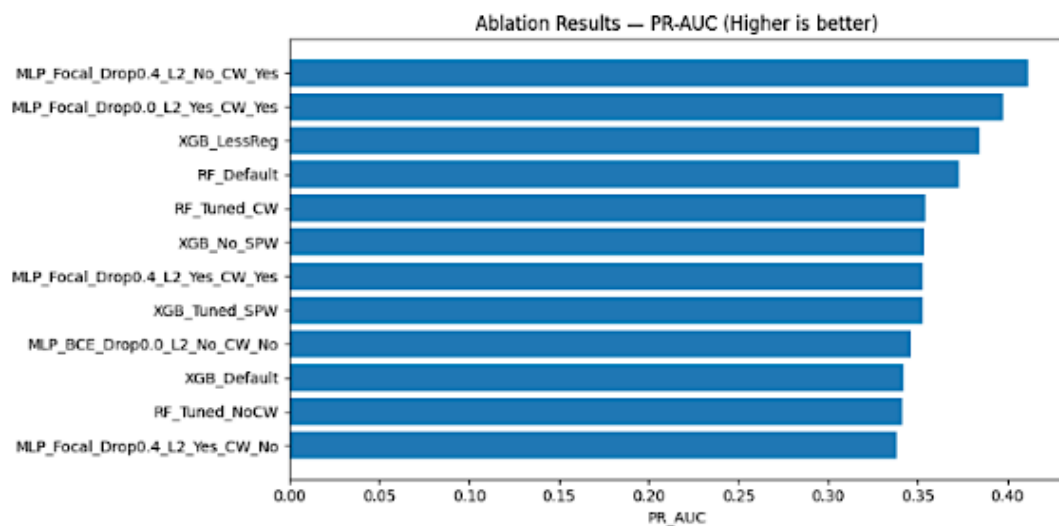
Table 10. Comparative Ablation Results

Configuration	ROC-AUC	PR-AUC
RF_Default	≈ 0.772	≈ 0.369
XGB_LessReg	≈ 0.766	≈ 0.381
XGB_Tuned_SPW	≈ 0.760	≈ 0.346
XGB_No_SPW	≈ 0.755	≈ 0.351
RF_Tuned_CW	≈ 0.755	≈ 0.351
RF_Tuned_NoCW	≈ 0.749	≈ 0.337
XGB_Default	≈ 0.746	≈ 0.338

Configuration	ROC-AUC	PR-AUC
MLP_Focal_Drop0.4_L2_No_CW_Yes	≈ 0.735	≈ 0.407
MLP_BCE_Drop0.0_L2_No_CW_No	≈ 0.735	≈ 0.341
MLP_Focal_Drop0.4_L2_Yes_CW_No	≈ 0.721	≈ 0.334
MLP_Focal_Drop0.0_L2_Yes_CW_Yes	≈ 0.718	≈ 0.395
MLP_Focal_Drop0.4_L2_Yes_CW_Yes	≈ 0.718	≈ 0.348



(a)



(b)

Figure 6: Comparison in Ablation Experiments (a) ROC-AUC, (b) PR-AUC

Hyperparameter sensitivity analysis revealed distinct behavioral patterns across model families. Logistic Regression exhibited relatively stable performance across configurations, although it remained sensitive to the magnitude of regularization strength. XGBoost performance varied primarily with adjustments to learning rate, number of estimators, and subsampling parameters, reflecting the impact of boosting dynamics on convergence behavior. In contrast, MLP showed the highest sensitivity to regularization intensity and focal loss gamma parameters, which directly influence the balance between representational capacity and generalization. These findings indicate that model robustness is not uniform across paradigms, but depends on how each model responds to parameter perturbations under controlled experimental conditions.

Cross-dataset evaluation was conducted to assess generalization robustness under distributional shift conditions. Although strong discrimination was observed in the in-domain test setting, all models exhibited performance degradation when evaluated on data drawn from alternative distributions. As shown in Figure 7, ROC-AUC values declined substantially under cross-domain conditions, approaching near-random discrimination levels. These findings indicate that increased model complexity does not necessarily translate into improved transferability across heterogeneous credit populations. This result highlights the sensitivity of tabular credit models to distributional changes, suggesting that generalization performance is constrained by data similarity rather than model capacity alone.

Table 11. Cross-Dataset Evaluation Based on ROC-AUC

Evaluation Scenario	Logistic Regression	Random Forest	XGBoost	MLP
In-domain: German Credit	≈ 0.66	≈ 0.75	≈ 0.76	≈ 0.70
In-domain: Australian Credit	≈ 0.92	≈ 0.93	≈ 0.93	≈ 0.93
Cross-domain: Train German → Test Australian	≈ 0.50	≈ 0.50	≈ 0.50	≈ 0.34
Cross-domain: Train Australian → Test German	≈ 0.50	≈ 0.50	≈ 0.50	≈ 0.54

Table 11 presents the ROC-AUC results for in-domain and cross-dataset evaluation. In the German Credit Dataset, XGBoost achieved the highest in-domain ROC-AUC of approximately 0.76, followed closely by Random Forest at approximately 0.75. In the Australian Credit Dataset, all models achieved substantially stronger in-domain discrimination, with ROC-AUC values of approximately 0.92–0.93. These differences indicate that the Australian dataset may provide a more separable classification structure under the applied preprocessing and modeling protocol.

Cross-dataset performance declined substantially for all models. When trained on the German Credit Dataset and tested on the Australian Credit Dataset, Logistic Regression, Random Forest, and XGBoost produced ROC-AUC values close to 0.50, indicating discrimination near random ranking, while the MLP declined further to approximately 0.34. When the training and testing direction was reversed, the three classical and ensemble models again achieved approximately 0.50, whereas the MLP reached approximately 0.54. This severe degradation suggests that the models learned dataset-specific relationships that did not transfer reliably across the two credit datasets.

Table 12. Cross-Dataset Performance Comparison

Training dataset	Testing dataset	Model	ROC-AUC	PR-AUC
German Credit	Australian Credit	Logistic Regression	≈0.50	≈0.39–0.45
German Credit	Australian Credit	Random Forest	≈0.50	≈0.39–0.45
German Credit	Australian Credit	XGBoost	≈0.50	≈0.39–0.45
German Credit	Australian Credit	MLP	≈0.50	≈0.39–0.45
Australian Credit	German Credit	Logistic Regression	≈0.50	≈0.15–0.18
Australian Credit	German Credit	Random Forest	≈0.50	≈0.15–0.18
Australian Credit	German Credit	XGBoost	≈0.50	≈0.15–0.18
Australian Credit	German Credit	MLP	≈0.54	≈0.15–0.18

The Australian Credit Dataset was used as the auxiliary dataset for cross-dataset evaluation. Models were evaluated in two transfer directions: training on the German Credit Dataset and testing on the Australian Credit Dataset, and training on the Australian Credit Dataset and testing on the German Credit Dataset. The cross-domain results showed substantial performance degradation compared with the corresponding in-domain evaluations. When trained on the German dataset and tested on the Australian

dataset, all models obtained ROC-AUC values close to 0.50, indicating discrimination performance near random ranking. A similar pattern was observed in the reverse direction, although the MLP achieved a slightly higher ROC-AUC of approximately 0.54. PR-AUC also declined markedly, reaching approximately 0.39–0.45 for German-to-Australian transfer and approximately 0.15–0.18 for Australian-to-German transfer. These findings demonstrate that the models were highly sensitive to differences in feature distributions, class prevalence, and borrower characteristics across the two credit domains.

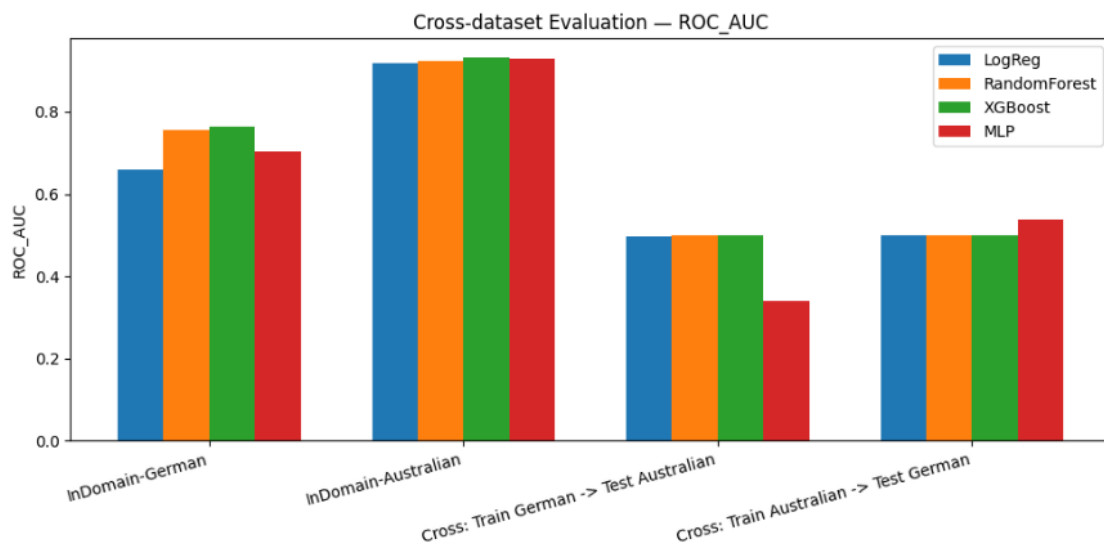


Figure 7: Cross-Dataset Evaluation Performance

To complement predictive evaluation, an exploratory explainability analysis was performed using SHAP-based global and local attribution methods. Global SHAP importance analysis identified consistent contributions from key financial and demographic variables across ensemble and neural network models. Local-level explanations further revealed instance-specific attribution patterns, highlighting variation in decision boundaries among models. However, this analysis is intended to provide qualitative insight into model behavior rather than a comprehensive comparative attribution study. Although neural networks exhibited more abstract attribution distributions than tree-based models, these observations should be interpreted cautiously, as detailed cross-model attribution consistency and interaction effects were not exhaustively evaluated. Future work is therefore required to systematically investigate explainability across model families in imbalanced credit scoring contexts.

3.2. Discussion

3.2.1. Comparative Model Behavior

Random Forest achieved the highest recall for the bad-credit class at 0.6774, but its precision remained relatively low at 0.3088. This result indicates that Random Forest adopted a comparatively inclusive decision boundary when identifying high-risk applicants. The model successfully detected 21 of the 31 bad-credit cases in the test set, but it also incorrectly classified 47 good-credit applicants as bad credit. Therefore, the improvement in true-positive detection was accompanied by an increase in false-positive predictions. This behavior is consistent with an imbalance-aware ensemble model that places greater emphasis on minimizing false negatives. In credit-risk screening, such a trade-off may be acceptable when failing to identify a potentially high-risk borrower is considered more consequential than referring a low-risk borrower for additional assessment. However, its lower precision limits its suitability for fully automated rejection because a substantial proportion of applicants predicted as high risk may actually belong to the good-credit class.

XGBoost achieved the highest ROC-AUC of 0.7643, indicating that it provided the strongest overall ranking of bad-credit and good-credit applicants across the complete range of possible decision thresholds. However, its PR-AUC of 0.3519 was lower than the MLP value of 0.3785 and the Random Forest value of 0.3657. This difference arises because ROC-AUC and PR-AUC evaluate distinct aspects of predictive performance. ROC-AUC assesses the model's ability to rank positive observations above negative observations by considering the relationship between the true-positive rate and false-positive rate across all thresholds. By contrast, PR-AUC focuses specifically on the trade-off between precision and recall for the positive bad-credit class and is more sensitive to false-positive predictions and minority-class prevalence. XGBoost therefore achieved the strongest global separation between the two credit classes, but it did not maintain the most favorable precision-recall relationship for bad-credit applicants across all relevant thresholds. Accordingly, XGBoost should be characterized as the strongest global risk-ranking model rather than the universally superior minority-class classifier.

The MLP obtained the lowest Brier Score of 0.1233, although its recall of 0.4839 and Balanced Accuracy of 0.6709 were the lowest among the evaluated models. These outcomes are not contradictory because the metrics evaluate different dimensions of

model performance. The Brier Score measures the average squared difference between predicted probabilities and observed binary outcomes across all test observations. Recall and Balanced Accuracy, in contrast, are threshold-dependent metrics calculated after continuous probabilities have been converted into class labels. The MLP may therefore generate comparatively accurate probability estimates overall while assigning a substantial number of bad-credit applicants probabilities below the selected decision threshold. Its overall accuracy of 0.8000, combined with the lowest Balanced Accuracy, further suggests that the model performed particularly well on the majority good-credit class but less effectively on the minority bad-credit class. Consequently, the MLP should not be described as providing the strongest threshold-specific minority-class detection. A more precise interpretation is that the MLP achieved the highest PR-AUC and the best raw probability quality, but lower bad-credit sensitivity at the selected threshold.

The Discussion now clarifies that the Brier Score evaluates probability quality, whereas recall and Balanced Accuracy depend on the classification threshold. We also explain that the MLP's high overall accuracy but lower Balanced Accuracy indicates stronger performance on the majority class. The previous claim regarding superior minority-class detection has therefore been moderated.

3.2.2. Threshold-Dependent Model Ranking

The ranking of the models based on threshold-independent metrics differs from their ranking based on threshold-dependent metrics. XGBoost achieved the highest ROC-AUC, whereas MLP achieved the highest PR-AUC. However, XGBoost obtained the highest Balanced Accuracy and F1-score at the selected decision threshold, while Random Forest produced the highest recall. These differences demonstrate that model ranking depends on both the evaluation metric and the classification threshold. The current Precision, Recall, F1-score, and Balanced Accuracy values were calculated using a fixed threshold of 0.50. No threshold optimization was performed using the test set because such optimization would introduce optimistic bias. Therefore, the reported ranking should be interpreted as specific to the selected threshold rather than universally optimal. Validation-based threshold optimization may improve minority-class sensitivity and could alter the ranking, particularly for the MLP, whose relatively high PR-AUC but low recall suggests that several bad-credit observations may have received probabilities below the current cutoff.

3.2.3. Practical Decision Scenarios

The results indicate that no single model is optimal for all credit-decision objectives. Model selection should therefore be aligned with the intended decision scenario rather than determined solely by the highest overall accuracy or a single discrimination metric. For risk-ranking applications, XGBoost is the most appropriate candidate because it achieved the highest ROC-AUC. This indicates that XGBoost provides the strongest overall ordering of applicants from lower to higher predicted credit risk. Such ranking capability may support portfolio prioritization, risk-tier assignment, risk-based pricing, and the sequencing of applications for additional assessment.

For default-screening applications, Random Forest may be preferable because it produced the highest recall for bad-credit applicants. Its lower precision means that more good-credit applicants may be incorrectly flagged as risky, but this trade-off may be acceptable in an initial screening stage where the principal objective is to minimize undetected high-risk borrowers. Applications flagged by Random Forest should therefore undergo secondary manual, rule-based, or document-based verification before an adverse lending decision is issued.

For probability-estimation applications, the MLP produced the lowest Brier Score and therefore the best raw probability quality among the evaluated models. This characteristic may be useful for applications requiring continuous risk estimates, such as credit-limit determination, expected-loss estimation, or borrower risk segmentation. Nevertheless, the MLP probabilities should not be treated as formally calibrated unless post-hoc methods such as Platt scaling, isotonic regression, or temperature scaling are applied and validated. Its lower recall also indicates that probability quality alone does not guarantee effective threshold-specific detection of bad-credit applicants.

For low-cost and transparent deployment, Logistic Regression remains a practical alternative. Although it produced the lowest ROC-AUC and PR-AUC, it required the shortest training and inference times and offers comparatively straightforward coefficient-based interpretation. Logistic Regression may therefore remain suitable for resource-constrained environments, baseline risk assessment, rapid model auditing, or institutional settings where simplicity, transparency, and implementation cost are prioritized over marginal predictive improvements.

Overall, the findings demonstrate that performance in imbalanced credit scoring is inherently multidimensional. XGBoost provides the strongest global discrimination, Random Forest prioritizes bad-credit recall, the MLP achieves the highest PR-AUC and best raw probability quality but weaker threshold-specific sensitivity, and Logistic Regression offers the greatest computational simplicity and transparency. The relative ranking of these models therefore depends on whether the intended objective is global risk ordering, detection of potentially high-risk borrowers, probability estimation, or low-cost deployment. Consequently, the results do not support the existence of a universally superior model. Instead, model selection should reflect the relative consequences of false-negative and false-positive decisions, the required quality of probability estimates, threshold-selection policies, computational constraints, and institutional governance requirements.

4. CONCLUSION

This study provides a multimetric comparison of Logistic Regression, Random Forest, XGBoost, and Multilayer Perceptron for imbalanced credit scoring using the German Credit Dataset. The results indicate that model superiority depends on the evaluation objective. XGBoost achieved the strongest global discrimination, Random Forest produced the highest recall for bad-credit applicants, the MLP obtained the highest PR-AUC and lowest Brier Score but lower threshold-specific recall and Balanced Accuracy, and Logistic Regression offered the greatest computational efficiency and implementation simplicity. These findings should be regarded as preliminary and benchmark-specific because they were derived from a relatively small and historically established dataset rather than a contemporary institutional lending portfolio. Therefore, the results demonstrate comparative model behavior under controlled experimental conditions but should not be interpreted as direct evidence of operational performance across banks, fintech platforms, or broader credit markets.

The cost-sensitive component of this study was implemented through model-specific imbalance-aware mechanisms, including class weighting, `scale_pos_weight`, and focal loss, rather than through an explicit institution-specific business-cost matrix. Consequently, the study does not quantify the actual monetary consequences of false approvals, false rejections, or missed defaults. Future research should validate the models using real bank

loan portfolios and examine reject inference to reduce selection bias caused by the absence of repayment outcomes for rejected applicants. Further investigations should assess algorithmic fairness across protected or socio-demographic groups, model stability under temporal and population drift, and regulatory explainability suitable for adverse-action decisions and model-risk governance. Future work should also compare post-hoc probability-calibration methods, including Platt scaling, isotonic regression, and temperature scaling, and evaluate validation-based threshold optimization according to different operational objectives, such as maximizing recall, Balanced Accuracy, expected utility, or institution-specific decision costs.

REFERENCES

- [1] L. I. Souadda, A. R. Halitim, B. Benilles, J. M. Oliveira, and P. Ramos, "Optimizing Credit Risk Prediction for Peer-to-Peer Lending Using Machine Learning," *Forecasting*, vol. 7, no. 3, p. 35, Jun. 2025, doi: 10.3390/Forecast7030035.
- [2] V. Chang, Q. A. Xu, S. H. Akinloye, V. Benson, and K. Hall, "Prediction of bank credit worthiness through credit risk analysis: an explainable machine learning study," *Ann. Oper. Res.*, vol. 354, no. 1, pp. 247–271, Nov. 2025, doi: 10.1007/s10479-024-06134-x.
- [3] M. Song, H. Ma, Y. Zhu, and M. Zhang, "Credit Risk Prediction Based on Improved ADASYN Sampling and Optimized LightGBM," *Journal of Social Computing*, vol. 5, no. 3, pp. 232–241, 2024, doi: 10.23919/JSC.2024.0019.
- [4] Y. Fang, "Application of a Grey Wolf Optimization-Enhanced Convolutional Neural Network and Bidirectional Gated Recurrent Unit Model for Credit Scoring Prediction," *PLoS One*, vol. 20, no. 5, p. e0322225, 2025, doi: 10.1371/journal.pone.0322225.
- [5] A. E. Khandani, A. J. Kim, and A. W. Lo, "Consumer Credit-Risk Models via Machine-Learning Algorithms," *J. Bank. Financ.*, vol. 34, no. 11, pp. 2767–2787, 2010, doi: 10.1016/j.jbankfin.2010.06.001.
- [6] S. Kokate and M. S. R. Chetty, "Credit Risk Assessment of Loan Defaulters in Commercial Banks Using Voting Classifier Ensemble Learner Machine Learning Model," *International Journal of Safety and Security Engineering*, vol. 11, no. 5, pp. 565–572, 2021, doi: 10.18280/ijssse.110508.

- [7] M. M. Seliem, M. Amin, M. M. A. El Nasr, E. A. Elnaggar, H. A. M. Khalifa, and M. A. A. Arab, "Performance of Machine Learning Algorithms for Credit Risk Prediction with Feature Selection," *Stat. Optim. Inf. Comput.*, vol. 14, no. 1, pp. 311–328, doi: 10.19139/soic-2310-5070-2392.
- [8] C. X. Pham, H. N. Trinh, and L. Q. Tran, "A Robust Approach to Credit Scoring with Deep Learning and Embedded Methods," *Engineering, Technology & Applied Science Research*, vol. 15, no. 6, pp. 29284–29291, 2025, doi: 10.48084/etasr.12649.
- [9] A. Shmuel, O. Glickman, and T. Lazebnik, "A Comprehensive Benchmark of Machine and Deep Learning Models on Structured Data for Regression and Classification," *Neurocomputing*, vol. 655, p. 131337, 2025, doi: 10.1016/j.neucom.2025.131337.
- [10] B. R. Gunnarsson, S. V. Broucke, B. Baesens, M. Óskarsdóttir, and W. Lemahieu, "Deep learning for credit scoring: Do or don't?," *Eur. J. Oper. Res.*, vol. 295, no. 1, pp. 292–305, doi: 10.1016/j.ejor.2021.03.006.
- [11] I. M. B. Vera, J. M. Restrepo, and D. F. Manotas-Duque, "Data Mining for the Adjustment of Credit Scoring Models in Solidarity Economy Entities: A Methodology for Addressing Class Imbalances," *Risks*, vol. 13, no. 2, p. 20, 2025, doi: 10.3390/risks13020020.
- [12] J. T. Hancock, T. M. Khoshgoftaar, and J. M. Johnson, "Evaluating Classifier Performance with Highly Imbalanced Big Data," *J. Big Data*, vol. 10, no. 1, p. 42, 2023, doi: 10.1186/s40537-023-00724-5.
- [13] P. Thölke, "Class Imbalance Should Not Throw You Off Balance: Choosing the Right Classifiers and Performance Metrics for Brain Decoding with Imbalanced Data," *Neuroimage*, vol. 277, p. 120253, 2023, doi: 10.1016/j.neuroimage.2023.120253.
- [14] J. T. Hancock, T. M. Khoshgoftaar, and J. M. Johnson, "Evaluating classifier performance with highly imbalanced Big Data," *J. Big Data*, vol. 10, no. 1, p. 42, doi: 10.1186/s40537-023-00724-5.
- [15] F. Shen, X. Zhao, G. Kou, and F. E. Alsaadi, "A New Deep Learning Ensemble Credit Risk Evaluation Model with an Improved Synthetic Minority Oversampling Technique," *Appl. Soft Comput.*, vol. 98, p. 106852, 2021, doi: 10.1016/j.asoc.2020.106852.
- [16] L. O. Hjelkrem and P. E. D. Lange, "Explaining Deep Learning Models for Credit Scoring with SHAP: A Case Study Using Open Banking Data," *Journal of Risk and Financial Management*, vol. 16, no. 4, p. 221, 2023, doi: 10.3390/jrfm16040221.

- [17] C. X. Pham, H. N. Trinh, and L. Q. Tran, "A Robust Approach to Credit Scoring with Deep Learning and Embedded Methods," *Eng. Technol. Appl. Sci. Res*, vol. 15, no. 6, pp. 29284–29291, doi: 10.48084/etasr.12649.
- [18] J. Xiao, S. Li, Y. Tian, J. Huang, X. Jiang, and S. Wang, "Example Dependent Cost Sensitive Learning Based Selective Deep Ensemble Model for Customer Credit Scoring," *Sci. Rep.*, vol. 15, no. 1, p. 6000, 2025, doi: 10.1038/s41598-025-89880-7.
- [19] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, 2nd ed. in Springer Texts in Statistics. New York, NY: Springer US, 2021. doi: 10.1007/978-1-0716-1418-1.
- [20] M. Sivakumar, S. Parthasarathy, and T. Padmapriya, "Trade-off between training and testing ratio in machine learning for medical image processing," *PeerJ Comput. Sci*, vol. 10, p. 2245, 2024. doi: 10.7717/peerj-cs.2245
- [21] Y. Chen, R. Calabrese, and B. Martin-Barragan, "Interpretable machine learning for imbalanced credit scoring datasets," *Eur. J. Oper. Res*, vol. 312, no. 1, pp. 357–372, 2024. doi: 10.1016/j.ejor.2023.06.036.
- [22] P. Thölke, "Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data," *Neuroimage*, vol. 277, p. 120253, 2023. doi: 10.1016/j.neuroimage.2023.120253.
- [23] W. Liu, H. Fan, M. Xia, and M. Xia, "A focal-aware cost-sensitive boosted tree for imbalanced credit scoring," *Expert Syst. Appl*, vol. 208, p. 118158, 2022. doi: 10.1016/j.eswa.2022.118158.
- [24] A. Aghbalou, A. Sabourin, and F. Portier, "On the Bias of K-Fold Cross Validation with Stable Learners," in *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 206, pp. 3775–3794, 2023