symmetry

MDPI

Article

Level-Wise Feature-Guided Cascading Ensembles for Credit Scoring

Yao Zou * and Guanghua Cheng

School of Economics and Management, Huainan Normal University, Huainan 232038, China; chengguanghua@hnnu.edu.cn

* Correspondence: zouyaodhu@mail.dhu.edu.cn

Abstract: Accurate credit scoring models are essential for financial risk management, yet conventional approaches often fail to address the complexities of high-dimensional, heterogeneous credit data, particularly in capturing nonlinear relationships and hierarchical dependencies, ultimately compromising predictive performance. To overcome these limitations, this paper introduces the level-wise feature-guided cascading ensemble (LFGCE) model, a novel framework that integrates hierarchical feature selection with cascading ensemble learning to systematically uncover latent feature hierarchies. The LFGCE framework leverages symmetry principles in its cascading architecture, where each ensemble layer maintains structural symmetry in processing its assigned feature subset while asymmetrically contributing to the final prediction through hierarchical information fusion. The LFGCE model operates through two synergistic mechanisms: (1) a hierarchical feature selection strategy that quantifies feature importance and partitions the feature space into progressively predictive subsets, thereby reducing dimensionality while preserving discriminative information, and (2) a cascading ensemble architecture where each layer specializes in learning risk patterns from its assigned feature subset, while iteratively incorporating outputs from preceding layers to enable cross-level information fusion. This dual process of hierarchical feature refinement and layered ensemble learning allows the LFGCE to extract deep, robust representations of credit risk. Empirical validation on four public credit datasets (Australian Credit, German Credit, Japan Credit, and Taiwan Credit) demonstrates that the LFGCE achieves an average AUC improvement of 0.23% over XGBoost (Python 3.13) and 0.63% over deep neural networks, confirming its superior predictive accuracy.

Keywords: credit scoring; cascading ensemble; hierarchical feature selection; LFGCE

check for updates

Academic Editors: Xiao Xu, Hui Gao and Jun Xie

Received: 16 April 2025
Revised: 2 June 2025
Accepted: 6 June 2025
Published: 10 June 2025

Citation: Zou, Y.; Cheng, G. Level-Wise Feature-Guided Cascading Ensembles for Credit Scoring. *Symmetry* **2025**, *17*, 914. https://doi.org/10.3390/sym17060914

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Credit scoring serves as a fundamental mechanism for evaluating personal credit risk, influencing access to financial products—from credit cards and consumer loans to mortgages—by determining eligibility and interest rates [1]. To enhance risk assessment, financial institutions systematically gather applicant data (e.g., income, debt levels) and employ scoring models to synthesize this information into predictive measures of repayment capacity. Historically, credit scoring relied on statistical methodologies, leveraging extensive historical data to construct standardized risk evaluation frameworks using techniques such as linear discriminant analysis (LDA) [2] and logistic regression (LR) [3,4]. However, the efficacy of these models is constrained by their reliance on assumptions, such as multivariate normality [5]. This limitation has driven the evolution of credit scoring toward more sophisticated computational approaches.

The advancement of information technology has propelled machine learning methods to surpass traditional statistical models in credit scoring, owing to their ability to process complex, high-dimensional data without restrictive distributional assumptions while autonomously identifying nonlinear relationships and latent patterns [6]. These models exhibit superior adaptability and generalization across diverse financial contexts, consistently demonstrating enhanced predictive accuracy and robustness in credit risk assessment [7,8]. Among machine learning approaches, ensemble learning has emerged as a pivotal methodology, integrating multiple classifiers to optimize performance—exemplified by random forest (RF) [9], gradient boosting machine (GBM) [10], and XGBoost [11]—which excel in modeling the intricate nonlinearities inherent in credit data.

Despite these advancements, conventional credit scoring models remain constrained by the challenges of manual feature engineering and parameter optimization. Deep learning has revolutionized this paradigm through end-to-end training mechanisms that autonomously extract hierarchical feature representations and nonlinear mappings from raw data, circumventing human-dependent design. Innovations such as the LSTM-GRU stacking ensemble with MLP meta-learner [12], improved DNN architectures for feature selection [13], and the VAE-DF framework combining variational autoencoders with deep forest classifiers [14] illustrate how deep ensemble models are redefining credit scoring in complex financial ecosystems.

Although neural network algorithms have shown remarkable potential in the field of credit scoring, their inherent limitations cannot be ignored, especially in terms of model interpretability, data dependency, and generalization ability. The asymmetric distribution of credit features across different demographic groups creates symmetry-breaking patterns that conventional neural networks struggle to capture, leading to biased risk assessments. Credit scoring datasets typically exhibit characteristics such as sparse samples and imbalanced categories, which undoubtedly exacerbate the difficulty and complexity of training neural network models. Insufficient data quality may result in a lack of representativeness in the feature representations learned by the model, thereby affecting its predictive performance. In addition, even with relatively abundant data, neural network models still face the risk of overfitting, which weakens their generalization ability. Due to the powerful fitting capability of neural networks, the models may overlearn the specific patterns and noise in the training data during the training process, leading to a significant decline in performance on unseen data. This lack of generalization ability directly affects the predictive accuracy and stability of credit scoring models, limiting their reliability in practical applications.

To address the above limitations, this paper proposes a credit scoring model based on the Cascaded Gradient Boosting Trees algorithm. Our cascaded architecture introduces symmetry-preserving transformations at each level, maintaining balanced information flow while allowing asymmetric feature interactions to capture complex risk patterns. The contributions of this paper are summarized as follows:

(1) We introduce a novel hierarchical feature selection strategy that systematically refines high-dimensional heterogeneous data into a more discriminative and parsimonious feature representation. This structured approach significantly reduces data complexity and enhances the interpretability of key credit risk indicators, laying a robust foundation for subsequent modeling.

(2) Building upon this refined feature subspace, we develop a cascaded gradient boosting tree architecture designed for the deep exploration of complex nonlinear relationships inherent in credit data. This layered structure enables progressive learning and effective information fusion across levels, thereby substantially improving the model's expressive power and predictive accuracy for credit risk.

(3) The inherent ensemble nature of our proposed cascaded framework significantly bolsters model robustness. By integrating multiple weak learners within the cascaded structure, the LFGCE model effectively mitigates the adverse impacts of data noise and outliers and curtails the risk of overfitting, which is critical for ensuring stable and reliable credit scoring in volatile financial environments.

This work makes significant theoretical and methodological advances beyond existing credit scoring approaches through three key innovations that collectively address fundamental limitations in the field. Unlike conventional feature selection methods that treat variables in isolation, our hierarchical feature selection strategy systematically transforms high-dimensional heterogeneous data into a discriminative yet parsimonious representation through a structured refinement process, thereby resolving the critical trade-off between information completeness and model interpretability that has persistently plagued credit risk modeling. The cascaded gradient boosting tree architecture represents a fundamental departure from traditional single-layer ensemble methods by enabling deep exploration of complex nonlinear relationships through progressive learning across hierarchical levels, a design that captures intricate credit risk patterns while maintaining computational efficiency—an achievement unattainable by either conventional gradient boosting machines or deep learning approaches in this domain. Furthermore, the model's inherent ensemble mechanism within this cascaded framework introduces a novel paradigm for robustness enhancement, where the synergistic integration of multiple weak learners not only mitigates data noise and outlier effects but also establishes a self-regulating mechanism against overfitting, thereby addressing two persistent challenges in credit scoring simultaneously through a unified architecture. These innovations collectively represent a substantial advancement over existing methods by providing a comprehensive solution that achieves superior predictive performance while maintaining practical interpretability and operational stability—qualities rarely coexisting in current credit risk assessment systems.

2. Literature Review

Credit risk, also known as default risk, generally refers to the possibility that a borrower may be unable to fulfill the contract due to certain reasons, resulting in the provider of funds bearing economic losses. For the personal loan business of financial companies, the main risk they face is credit risk, which refers to the risk of the borrower not fulfilling or not fully fulfilling the contract, known as personal default risk. The characteristics of default risk include customer dispersion, concealment, periodicity, etc. Credit scoring is the quantification of customer credit risk. The connotation of credit scoring is to quantitatively characterize evaluation indicators such as repayment ability and performance probability by studying users' historical credit data. According to the different properties of the indicators, they are divided into several levels, and corresponding scores are given to each level. The final credit score is obtained by a weighted calculation of the scores of each indicator.

Traditional credit scoring models have extensively used concepts and principles from econometrics, statistics, and operations research, resulting in methods such as discriminant analysis, linear regression, and logistic regression. Myers et al. [15] used discriminant analysis and regression analysis to establish a model based evaluate the credit risk of customers based on the data in their loan application forms. In the following years, Orgler [16] and Fitzpatrick [17] also introduced linear regression methods into credit scoring. Compared with linear regression and discriminant analysis, the logistic regression model is more suitable for binary classification problems. Wiginton [18] applied logistic regression to the field of credit scoring and compared it with discriminant analysis and found that the logistic regression model had better classification performance. The logistic

regression model is a linear classification model that has the characteristics of a simple model structure and easily interpretable results. However, it cannot effectively extract nonlinear information, which is why an increasing number of scholars have begun to consider applying more machine learning algorithms to credit scoring.

The commonly used machine learning models for credit risk assessment include the K-nearest neighbors (K-NN) [19,20], decision trees (DT) [21], neural network (NN) [22,23], support vector machine (SVM) [24,25], and hybrid machine learning model [26,27]. Blanco et al. [28] found that neural network models outperform the other three classic techniques in terms of AUC. Harris [29] proposed a clustered support vector machine (CSVM) for credit scoring. Kang et al. [30] proposed a conditional Wasserstein generative adversarial network with a gradient penalty (CWGAN-GP)-based multi-task learning (MTL) model (CWGAN-GP-MTL) for consumer credit scoring.

To compensate for the shortcomings of individual-classifier learning methods, Kearns and Valiant [31] proposed the principle of equivalence between strong learning and weak learning. To obtain a learning model with strong generalization ability, multiple simple weak learning models can be "improved", which is known as ensemble learning algorithms. The three common forms of ensemble models are bagging [32], boosting [11], and stacking [33]. In existing research, researchers usually improve on the three ensemble models mentioned above to establish effective models. Luo [34] found that the bagging ensemble method can substantially improve individual base learners such as decision trees, multilayer perceptrons, and k-nearest neighbors. Plaia et al. [35] confirmed that boosting outperforms bagging. Liu et al. [36] proposed a multi-grained and multi-layered gradient boosting decision tree (GBDT) for credit scoring and found that multi-grained feature augmentation effectively increased the diversity of prediction and further improved the performance of credit scoring. Rao et al. [37] combined particle swarm optimization (PSO) with an extreme gradient boosting (XGBoost) model to evaluate the credit risk of the loans and found that the proposed model was superior in classification performance and classification effect. Mushava and Murray [38] used new flexible loss functions for binary classification in gradient-boosted decision trees (GBDT) to assess the credit risk of the loans. Liu et al. [39] proposed a heterogeneous deep forest model (Heter-DF) for credit scoring, which featured a scalable cascading framework with multiple heterogeneous tree-based ensembled base learners and a weighted voting mechanism, demonstrating superior performance on various datasets and its effectiveness in both large-scale and small-scale credit scoring tasks. Yin et al. [40] proposed a stacking ensemble machine-learning model to assess credit default risk for P2P lending platforms and found that the proposed model had a minimum error rate and provided more accurate credit default risk prediction.

3. Methodology

3.1. Overview of LFGCE

The proposed model, termed level-wise feature-guided cascading ensembles, constitutes a sophisticated framework for credit scoring endeavors. The architecture, as illustrated in Figure 1, is structured hierarchically with multiple layers, each contributing significantly to the overall learning mechanism.

The input layer receives the raw feature vector $x \in \mathbb{R}^d$, where d denotes the dimensionality of the feature space. These features are concatenated (denoted by $\oplus$) and fed into a set of base learners $\{\mathcal{B}_t^1\}_{t=1}^T$. Each base learner $\mathcal{B}_t^1$ processes the concatenated features and generates an output o_t^1. Mathematically, the output of the t-th base learner in Layer 1 can be expressed as

$$o_t^1 = \mathcal{B}_t^1(x) \tag{1}$$

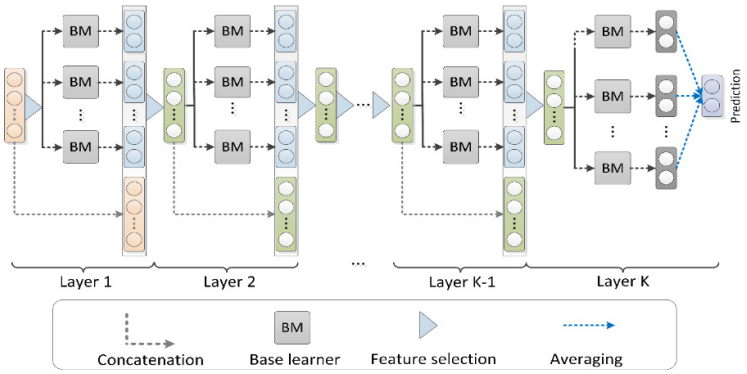


Figure 1. Framework of LFGCE.

The outputs of all base learners in Layer 1 are then aggregated to form the representation for the next layer. Let H_1 denote the aggregated output of Layer 1, which can be written as

$$o^1 = o_1^1 \oplus o_2^1 \oplus \cdots \oplus o_T^1 \tag{2}$$

For layer $l (2 \leq l \leq K-1)$, the output from the previous layer H_{l-1} undergoes feature selection (denoted by $\mathcal{F}_l$) before being concatenated and fed into the base learners $\{\mathcal{B}_t^1\}_{t=1}^T$ of the current layer. The feature selection operation can be modeled as

$$H'_{l-1} = \mathcal{F}_l(H_{l-1}) \tag{3}$$

The output of the t-th base learner in layer l is then given by

$$o_t^l = \mathcal{B}_t^l\left(H'_{l-1} \oplus H'_{l-1}\right) \tag{4}$$

The aggregated output of layer l, H_l, is calculated as

$$H_l = \sum_{t=1}^{T} o_t^l \tag{5}$$

The final layer, Layer L, takes the output from Layer $L-1$, H_{L-1}, and follows the same base learner and feature selection process. Let $H'_{l-1} = \mathcal{F}_l(H_{l-1})$. The output of the L-th base learner in Layer L is $o_t^L = \mathcal{B}_t^l\left(H'_{L-1} \oplus H'_{L-1}\right)$. The final prediction $\mathcal{B}_t^l$ is obtained by averaging the outputs of the base learners in this layer:

$$y = \frac{1}{T}\sum_{t=1}^{T} o_t^L \tag{6}$$

The base learners $\mathcal{B}_t^l$ in each layer are essential components. They can be instantiated as various simple machine learning models, such as decision trees, neural network units, etc. These base learners are responsible for learning local patterns and relationships within the input features at their respective layers. The ensemble of base learners in each layer helps to capture the diversity and complexity of the data. To formalize the learning process of a base learner, assume a loss function L_l for layer l. The goal of training a base learner $\mathcal{B}_t^l$ is to minimize the expected loss of the training data:

$$\min_{\mathcal{B}_t^l} \mathbb{E}_{(x,y)\sim\mathcal{D}}[L_l(y, \mathcal{B}_t^l(x))] \tag{7}$$

where $\mathcal{D}$ represents the training dataset.

The entire model is trained by minimizing a global loss function over all layers. This can be expressed as

$$\min_{\theta} L_{global}(y, \hat{y}) = \min_{\theta} \sum_{l=1}^{K} \lambda_l L_l(y_l, H_l) \tag{8}$$

where θ represents the parameters of all base learners and feature selection mechanisms in the model, λ_l are layer-wise weights that can be used to balance the contributions of different layers, and y_l is the intermediate prediction at layer l. The optimization process can be carried out using gradient-based methods or other optimization algorithms based on the nature of the base learners and the loss functions.

In conclusion, the level-wise feature-guided cascading ensemble model, through its elaborate layer-wise structure, base learners, feature selection, and optimization mechanisms, offers a powerful and flexible approach to handle complex data and achieve high-performance results in machine learning tasks. The mathematical formulations and notations presented herein enhance the theoretical rigor and clarity of the model description, facilitating a deeper understanding and analysis of its workings and potential applications.

3.2. XGBoost as Base Learner

XGBoost, short for extreme gradient boosting, is a prominent machine learning algorithm widely applied in regression and classification tasks. The following presents a detailed description of its algorithmic principles with relevant mathematical formulations.

XGBoost constructs an ensemble of decision trees sequentially. Let us assume we have a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathbb{R}^m$ represents the feature vector of the i-th sample, and y_i is the corresponding target value (either continuous for regression or discrete for classification).

The prediction function of the ensemble model at the t-th iteration is given by

$$\hat{y}_i^{(t)} = \sum_{t=1}^{t} f_k(x_i) \tag{9}$$

where $\hat{y}_i^{(t)}$ is the predicted value for the i-th sample at the t-th iteration, and $f_k(x)$ represents the k-th decision tree in the ensemble. Initially, when $k = 0$, we can set f_0 (for regression, often the mean of the target values; for classification, it could be a uniform distribution-based initial guess).

The goal of the gradient boosting process is to minimize the loss function $L(\hat{y}, y)$ over the training data. In each iteration t, we aim to find a new decision tree $f_t(x)$ that minimizes the following objective function:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} L\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{10}$$

where $\Omega(f_t)$ is a regularization term that penalizes the complexity of the decision tree $f_t(x)$ to prevent overfitting.

To optimize the loss function, we calculate the gradient and the second-order derivative (Hessian) of the loss function to the predicted value $\hat{y}$. The gradient g_i and the Hessian h_i for the -th sample are

$$g_i = \frac{\partial L(y_i, \hat{y}_i)}{\partial \hat{y}_i} \tag{11}$$

$$h_i = \frac{\partial^2 L(y_i, \hat{y}_i)}{\partial \hat{y}_i^2} \tag{12}$$

These gradients and Hessians are then used to guide the construction of the decision tree in the next step.

During the construction of a decision tree, at each node, we need to find the best split to minimize the loss function. Let us assume we are considering splitting a node based on a feature j at a value s. We divide the samples into two subsets $\mathcal{D}_l = \{(x_i, y_i) \mid x_{ij} \leq s\}$ and $\mathcal{D}_r = \{(x_i, y_i) \mid x_{ij} > s\}$.

$$\text{Gain} = \frac{1}{2}\left[\frac{\left(\sum_{i\in\mathcal{D}_l} g_i\right)^2}{\sum_{i\in\mathcal{D}_l} h_i + \lambda} + \frac{\left(\sum_{i\in\mathcal{D}_r} g_i\right)^2}{\sum_{i\in\mathcal{D}_r} h_i + \lambda} - \frac{\left(\sum_{i\in\mathcal{D}} g_i\right)^2}{\sum_{i\in\mathcal{D}} h_i + \lambda}\right] - \gamma \tag{13}$$

where is λ the L2 regularization parameter that penalizes the leaf weights of the tree and γ is the minimum loss reduction required for a split to be considered. XGBoost searches for the feature j and split value s that maximizes the gain across all possible splits.

The regularization term $\Omega(f_t)$ for a decision tree $f_t(x)$ with T leaves and leaf weights $w = \left\{w_j\right\}_{j=1}^{T}$ is defined as

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{14}$$

This regularization term consists of two parts: the first part γT penalizes the number of leaves in the tree, encouraging simpler tree structures, and the second part $\frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2$ penalizes the magnitude of the leaf weights.

Shrinkage is another important technique in XGBoost. The output of each decision tree $f_t(x)$ is multiplied by a learning rate η (a small positive value typically between 0 and 1). So, the updated prediction formula becomes

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i) \tag{15}$$

This makes the learning process more gradual and helps in better generalization.

3.3. Importance-Driven Feature Selection

Given a trained XGBoost model with T decision trees, the importance score $\mathcal{I}_j \in \mathbb{R}^+$ for feature x_j is computed as the normalized total gain across all splits utilizing this feature:

$$\mathcal{I}_j = \frac{1}{Z}\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t} \mathbb{I}(v(s) = x_j) \cdot \Delta\mathcal{L}_s \tag{16}$$

where $\mathcal{S}_t$ denotes the set of split nodes in tree t, $v(s)$ identifies the splitting feature at node s, $\Delta\mathcal{L}_s$ represents the corresponding gain in the loss function, and $Z = \sum_{j=1}^{d}\sum_{t=1}^{T}\sum_{s\in\mathcal{S}_t} \mathbb{I}\big(v(s) = x_j\big) \cdot \Delta\mathcal{L}_s$ serves as the normalization constant, ensuring $\sum_{j=1}^{d} \mathcal{I}_j = 1$. The importance-driven feature selection process then defines a selector function $\phi_k : \mathbb{R}^d \to \mathbb{R}^k$ that projects the original d-dimensional feature space onto the k-most important features through an ordered selection operation:

$$\phi_k(\mathbf{x}) = (x_{\sigma(1)}, \ldots, x_{\sigma(k)}), \sigma(i) = \underset{j \notin \{\sigma(1),\ldots,\sigma(i-1)\}}{\text{argmax}} \mathcal{I}_j \tag{17}$$

where σ generates the importance ranking permutation. This selection mechanism induces a truncated feature representation that preserves the maximal predictive information as measured by the cumulative importance $\sum_{i=1}^{k} \mathcal{I}_{\sigma(i)}$, while satisfying the dimensionality constraint $||\phi_k(\mathbf{x})||_0 = k$. The optimal cardinality k^* can be determined through cross-validation by minimizing the generalization error $\epsilon(k) = \mathbb{E}[L(y, f(\phi_k(\mathbf{x})))]$ over a held-out

validation set, where f denotes the subsequent predictive model utilizing the selected features.

3.4. Training and Inference Procedure of LFGCE

The training of the LFGCE algorithm starts with receiving training credit data $\{X_i, y_i\}_{i=1}^{D}$, cascade level L, maximum depth d of the decision tree, feature selection ratio r, and k-fold cross-validation parameter k, setting initial cross-validation accuracy $\mathcal{A}_0$ to 0 and initializing enhanced feature matrix $\mathbf{X}$. In cascade layer iteration, for each lay $l(1 \leq l \leq L)$, T basic learners are operated in turn. First, the dataset is divided into K parts for K-fold cross-validation. $X_{train} = X \setminus X_{val}$ is used as a training s ubset and X_{val} is used as the validation set. Based on X_{train} and y, the reference tree ensemble model $B_{tl} = Train(X_{train}, y)$ is trained and integrated into the cascade model B_l of layer l. Then, $V_{tl} \leftarrow B_{tl}(X_{train})$ is predicted for X_{train} by B_{tl} and the feature importance I_{tl} is calculated. After completing the training of T basic learners of this layer, calculate the K-fold cross-validation accuracy $\mathcal{A}_l = \frac{1}{K}\sum_{k=1}^{K} Acc\left(\mathcal{B}^l(X_{k,val}), y\right)$ of the l-th cascade layer. If $\mathcal{A}_l > \mathcal{A}_{l-1}$, calculate the overall feature importance score $I^l = \frac{1}{T}\sum_{t=1}^{T} I_t^l$ of the l-layer cascade model. After sorting the feature index according to feature importance scores, select the features according to the ratio r to update the feature matrix $\mathbf{X}_t^{l+1}$ with $\mathbf{X}_t^{l+1} \leftarrow \mathbf{X}_t^{l+1} \bigcup \mathbf{V}_t^l$ and simultaneously update $\mathcal{A}_{l-1}$ to $\mathcal{A}_l$. If $\mathcal{A}_l > \mathcal{A}_{l-1}$, the training is terminated in advance. Finally, a level-wise feature-guided cascading ensemble (LFGCE) model is constructed for subsequent tasks. The training pseudo-code of level-wise feature-guided cascading ensembles is shown in Algorithm 1.

Algorithm 1 Pseudo-code of level-wise feature-guided cascading ensembles

Input: Training data $\{X_i,\ y_i\}_{i=1}^{D}$ with training size D, cascade layers L, number of base learners per layer T, maximum depth d, feature selection ratio r, K for K-fold cross-validation, $\mathcal{A}_0 = 0$ as the initial cross-validation accuracy score.
Output: level-wise feature-guided cascading ensembles
Initialize $\mathbf{X}$ as the original enhanced feature matrix
for l = 1 to L **do**
  for t = 1 to T **do**
    Split the dataset into K parts for K-fold cross-validation as $[X_1, X_2, \ldots, X_{val}, \ldots, X_K]$, $X_{train} = X \backslash X_{val}$ is the subset for training, X_{val} denotes the validation set
  Train a benchmark tree ensemble model $\mathcal{B}_t^l = Train(X_{train}, y)$
    Ensemble base learner as l-th cascade layer $\mathcal{B}^l \leftarrow \mathcal{B}^l \bigcup \mathcal{B}_t^l$
    Get the prediction vector $\mathbf{V}_t^l \leftarrow \mathcal{B}_t^l(X_{train})$
    Compute feature importance I_t^l
  Compute K-fold cross-validation accuracy score for l-th cascade layer $\mathcal{A}_l = \frac{1}{K}\sum_{k=1}^{K} Acc\left(\mathcal{B}^l(X_{k,val}), y\right)$
  if $\mathcal{A}_l > \mathcal{A}_{l-1}$ do
    Get importance scores of l-th layer $I^l = \frac{1}{T}\sum_{t=1}^{T} I_t^l$
    Sort feature index by feature importance scores idx $= \text{argsort}\left(I^l\right)[:: -1]$
  Perform feature selection with a feature selection ratio $\mathbf{X}_t^{l+1} = \mathbf{X}_t^l[:, \text{idx}[0:r]]$
    Update $\mathbf{X}_t^{l+1}$ with $\mathbf{X}_t^{l+1} \leftarrow \mathbf{X}_t^{l+1} \bigcup \mathbf{V}_t^l$
    $\mathcal{A}_{l-1} \leftarrow \mathcal{A}_l$
  else
    break
return cascade ensemble LFGCE

In the practical application of credit scoring, given the trained hierarchical feature-guided cascading forest (LFGCE) model and a test sample $\mathbf{x}$ to be evaluated for credit risk assessment, the prediction process follows these steps: first, initialize the feature vector of the test sample $\mathbf{x}$ as the initial enhanced feature vector $f^{(0)}$. Subsequently, the model enters a layer-by-layer progressive prediction phase. For layer $l \in \{1, 2, \ldots, L\}$, obtain the selected feature subset $S(l)$ from the trained model for that layer. Then, initialize an empty category probability vector set $V'^{(l)}$ to store the prediction results of each random forest at that layer. For each random forest $RF^{(l,t)}(t \in \{1, 2, \ldots, T\})$ at that layer, extract the feature sub-vector $\mathbf{f}_{\text{sub}}^{(l-1)}$ corresponding to $S(l)$ from the current enhanced feature vector $\mathbf{f}^{(l-1)}$, and use $RF^{(l,t)}$ to predict $\mathbf{f}_{\text{sub}}^{(l-1)}$, generating a category probability vector $\mathbf{P}'^{(l,t)}$, which is added to $V'^{(l)}$. After completing the predictions of all random forests at that layer, calculate the average of all category probability vectors in $V'^{(l)}$ as $\overline{\mathbf{P}}'^{(l)} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{P}'^{(l,t)}$ and concatenate $\overline{\mathbf{P}}'^{(l)}$ with $\mathbf{f}^{(l-1)}$ horizontally to update the enhanced feature vector to $\mathbf{f}^{(l)} = [\mathbf{f}^{(l-1)}, \overline{\mathbf{P}}'^{(l)}]$, which serves as the input for the next layer. After completing the iterative predictions for all layers, the model enters the final prediction phase. Obtain the selected feature subset $S(L)$ of the last layer ($L - th$ layer) from the trained model and extract the corresponding feature sub-vector $\mathbf{f}_{\text{sub}}^{(L)}$ from the final enhanced feature vector $\mathbf{f}^{(L)}$. Subsequently, initialize a category probability vector set $V_{\text{final}}^{(L)}$ to store the prediction results of each random forest in the last layer. For each random forest $RF^{(L,t)}(t \in \{1, 2, \ldots, T\})$ in the last layer, use it to predict $\mathbf{f}_{\text{sub}}^{(L)}$, generating a category probability vector $\mathbf{P}_{\text{final}}^{(L,t)}$, which is added to $V_{\text{final}}^{(L)}$. Finally, calculate the average of all category probability vectors in $V_{\text{final}}^{(L)}$ as $\overline{\mathbf{P}}_{\text{final}}^{(L)} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{P}_{\text{final}}^{(L,t)}$. In $\overline{\mathbf{P}}_{\text{final}}^{(L)}$, select the class index corresponding to the element with the highest probability value as the final predicted credit category y_{pred} for the test sample $\mathbf{x}$. The inference pseudo-code of LFGCE is shown in Algorithm 2.

Algorithm 2 Inference pseudo-code of LFGCE

Input: A trained hierarchical feature-guided cascade forest model, comprising L levels. Each level l ($1 \le l \le L$) consists of T base learners $\left\{\mathcal{B}^l\right\}(1 \le \mathrm{t} \le T$ and a test sample $\mathbf{x}$.
Output: Predicted class label $\hat{y}$.

For layer l = 1 to $L - 1$
  //get top feature set
  $\mathbf{x}^l = \{\left(\mathbf{x}^l\right)$ //{ is a feature selection operation
  $\mathrm{V}^l = []$
  for base learner t = 1 **to** T
    $P_t^l = \mathcal{B}_t^l\left(\mathbf{x}^l\right)$
    concatenate predictive probability $\mathrm{V}^l \leftarrow [\mathrm{V}^l\ P_t^l]$
  $\mathbf{x}^{l+1} = [\mathbf{x}^l\ \mathrm{V}^l]$
$\mathbf{x}^L = \{(\mathbf{x}^L)$
$\mathrm{V}^l = []$
for $t = 1$ **to** T
  $P_t^L = \mathcal{B}_t^L(\mathbf{x}^L)$
Predictive probability $P = \frac{1}{T}\sum_{t=1}^{T} P_t^L$
$\hat{y} = \mathrm{argmax}(P)$

4. Experimental Settings

4.1. Credit Scoring Datasets

The process of validating the credit scoring performance of credit scoring models demands rigorous training and testing procedures, grounded in extensive historical credit data repositories. In this research endeavor, four benchmark datasets, meticulously curated for the purpose of credit risk assessment studies, are utilized. These include the Australian, German, Japanese, and Taiwanese credit datasets, each characterized by unique structural attributes that are comprehensively outlined in Table 1. These publicly accessible datasets function as standardized, experimental benchmarks, facilitating comparative evaluations of credit scoring models. The datasets employed in this study are sourced from the UCI repository, accessible via https://archive.ics.uci.edu/ on 1 December 2024.

Table 1. Information details of credit datasets.

Dataset	Samples	Variables	Good/Bad
Australian	690	14	307/383
German	1000	24	700/300
Japanese	690	15	296/357
Taiwan	6000	23	3000/3000

Table 1 employs credit scoring labels where "good" customers denote individuals with unblemished credit histories, while "bad" customers represent those exhibiting at least one default record, as even isolated default incidents substantially compromise creditworthiness assessments. Consequently, this study operationalizes creditworthiness differentiation through the binary indicator of default occurrence. The Australian dataset comprises 690 samples (383 good, 307 bad) featuring 8 continuous and 6 discrete anonymized attributes, with symbolic representations obscuring feature semantics for confidentiality. Similarly structured, the German dataset contains 1000 cases (700 good, 300 bad) characterized by 11 discrete and 13 continuous attributes. Comparatively, the Japanese dataset's 690 samples (296 good, 357 bad) incorporate 15 features per observation, whereas the Taiwan dataset's balanced 6000-sample cohort (3000 per class) exhibits higher dimensionality with 23 attributes per record.

4.2. Evaluation Metrics

The efficacy of precise credit assessment fundamentally depends on the predictive performance and operational reliability of credit scoring models, necessitating rigorous validation through multidimensional evaluation metrics. This study employs a comprehensive suite of statistically robust indicators encompassing accuracy, precision, recall, Brier score, F1 score, and AUC to systematically evaluate the proposed credit scoring algorithm's discriminative power and calibration quality, where each metric provides distinct yet complementary insights into different aspects of model performance. These carefully selected metrics collectively form an analytical framework that quantitatively assesses classification capability, probability estimation accuracy, and risk differentiation effectiveness, thereby ensuring a thorough validation of the model's practical applicability in credit decision-making scenarios.

Credit scoring yield four distinct prediction outcomes contingent upon the alignment between actual default status and model predictions: true positives (TPs) represent correctly identified non-default cases, false positives (FPs) denote erroneous non-default predictions for actual default cases, true negatives (TNs) reflect accurate default identifications, while false negatives (FNs) indicate misclassified default predictions for solvent cases. These classification outcomes, quantified as TP, FP, TN, and FN, respectively, constitute exhaustive

and mutually exclusive categories whose summation equals the total sample size, with optimal model performance characterized by maximized TP/TN values and minimized FP/FN occurrences. This classification framework finds its canonical representation in the confusion matrix (Table 2), which provides a comprehensive visualization of prediction accuracy across all possible outcome combinations.

Table 2. Prediction confusion matrix of the LFGCE model.

		Predicted	
		Bad	Good
Actual	Bad	TP	FP
	Good	FN	TN

(1) Accuracy (Acc)

The accuracy represents the overall prediction accuracy of the model, which is the proportion of correctly predicted samples in all predicted samples. Combined with the confusion matrix in Table 2, the accuracy can be expressed as

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{18}$$

(2) Recall (Rec)

Recall represents the completeness of the model prediction, with the numerator being the number of samples predicted as true positives and the denominator being the sum of true positives and false negatives. The larger the numerator, the more samples are predicted correctly. Its calculation formula is

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{19}$$

(3) Precision (Pre)

Precision indicates the validity of the model's predictions, with the numerator being the number of samples predicted as true positives and the denominator being the sum of true positives and false positives. The larger this ratio, the more individuals who are predicted to have defaulted. Its calculation formula is

$$\text{prec} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{20}$$

(4) F1-score (F1)

Since recall and precision are two contradictory evaluation metrics, pursuing a high recall will inevitably result in a low precision. Simply put, when there is a demand for information retrieval, if all information is retrieved, the proportion of truly desired information will be very low. Conversely, when precision is high, it means that the truly needed information has been accurately retrieved, while recall is low. Usually, both cannot be achieved simultaneously. Therefore, in statistics, there is an indicator that can be used to measure the accuracy of binary classification and takes into account both recall and precision. This indicator is the F1 score. The larger its value, the stronger the predictive ability of the model. The specific formula is as follows:

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{21}$$

(5) Brier Score (BS)

The Brier score represents the error of the model's default prediction, which describes the average error between the predicted result and the true state. The smaller the Brier score, the smaller the prediction error of the model. It can be calculated as

$$\mathrm{BS} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left(\mathrm{p_i} - \mathrm{y_i}\right)^2 \tag{22}$$

where p_i represents the predicted probability of the ith sample, y_i is the true label value of the ith sample, and N is the number of samples.

(6) ROC curve and AUC

AUC is the area under the curve of the receiver operating characteristic (ROC), and the horizontal axis of the ROC curve represents the false positive rate (FPR). The vertical axis represents the true positive rate (TPR). The value of AUC ranges from 0.5 to 1. The closer the AUC is to 1, the better the prediction performance of the model.

The true positive rate (TPR), also known as sensitivity, is the probability that the proposed model predicts no default among all samples that have not defaulted. Its formula is

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{23}$$

The false positive rate (FPR) is the probability that the model predicts non-default among all samples that defaulted, and its formula is represented as follows:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{24}$$

True negative rate (TNR), also known as specificity, represents the probability that the model predicts default among all samples that defaulted. Its formula is

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{FP} + \mathrm{TN}} \tag{25}$$

By plotting the classification results at different threshold values with the TPR on the vertical axis and the FPR on the horizontal axis, and connecting these points, the ROC curve is formed.

The receiver operating characteristic (ROC) curve serves as a robust diagnostic tool for evaluating model performance across diverse disciplines, having established its methodological rigor in biomedical research, psychological assessment, and clinical diagnostics before gaining prominence in machine learning applications. This graphical representation fundamentally illustrates the trade-off between sensitivity (TPR) and specificity (1-FPR), where elevated false positive rates correspond to the increased misclassification of negative instances, while heightened true positive rates reflect superior predictive accuracy for positive cases.

Assuming that the ROC curve is formed by connecting points with coordinates $\{(x_1, y_1), (x_2, y_2)...(x_i, y_i)\}$ in sequence, the AUC value can be calculated as follows:

$$\mathrm{AUC} = \frac{1}{2}\sum_{i=1}^{n}\left(x_{i+1} - x_i\right) \times \left(y_i + x_{i+1}\right) \tag{26}$$

4.3. Implementation Details

For the hyperparameter tuning, we first focus on the key parameters associated with the cascade layers and base learners. The number of cascade layers, denoted as L, is a

pivotal parameter that governs the depth of the cascading architecture. Although the LFGCE framework employs an adaptive structure that dynamically expands following the complexity of credit datasets, the maximum number of cascade layers remains variable. To enhance training efficiency, we establish an initial exploration range for L at 10. A larger L allows for a more complex hierarchical feature-guided structure but may also increase the computational cost and the risk of overfitting. The number of base learners per layer T affects the diversity and strength of each cascade layer. We set T as 4. A higher T can enhance the performance of each layer through increased ensemble strength, but it also demands more computational resources.

The maximum depth d of the base DTs in each cascade layer is an important factor. We set the initial searching space for d as [6,10]. A larger d enables the trees to capture more complex relationships in the data but may lead to overfitting. The node-splitting criteria can be chosen from common ones like Gini impurity or information gain.

The feature selection ratio r plays a vital role in guiding the feature selection process in each cascade layer. We start with an exploration range of r as [0.5, 1], with a step size of 0.1. A smaller r selects fewer features, which can reduce dimensionality and overfitting risk but may also discard useful information. A larger r includes more features, potentially capturing more information but increasing the complexity. For K-fold cross-validation, we set K to 10. All experiments were conducted on a workstation equipped with an Intel® Core™ i7-7700HQ processor (four cores, 2.8 GHz base frequency, up to 3.8 GHz Turbo Boost) and 32 GB DDR4 RAM (2400 MHz), running a 64-bit Windows 10 operating system.

5. Experimental Results

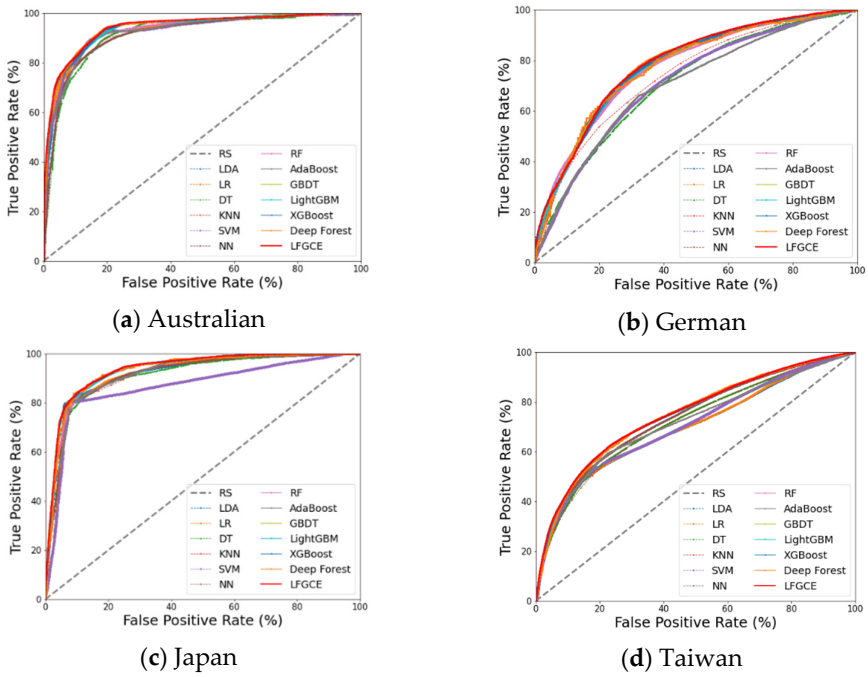
5.1. Performance Comparison and Analysis

For the enhanced visualization of model performance assessment, this paper implemented a comprehensive graphical analysis of credit scoring outcomes. The comparative visualization framework, as depicted in Figure 2, systematically contrasts the receiver operating characteristics of the proposed LFGCE methodology against an extensive spectrum of reference algorithms, spanning conventional statistical approaches (LR, LDA), individual machine learning architectures (SVM, DT, NN), and advanced ensemble techniques (RF, XGBoost). To establish methodological rigor, all performance metrics were derived through an intensive validation protocol employing 50 iterations of stratified 10-fold cross-validation, thereby ensuring statistical robustness and minimizing variance in the evaluation outcomes.

Figure 2 shows the ROC curve of credit scoring models on four different datasets. Figure 2a shows the ROC curve drawn based on the Australian credit score dataset from the UCI machine learning database; Figure 2b shows the ROC curve based on the German credit dataset; Figure 2c shows the ROC curve based on the Japanese credit dataset; and Figure 2d is a ROC curve plotted based on the Taiwan credit dataset.

As shown in Figure 2, experimental results across four benchmark datasets demonstrate the superior classification performance of the proposed LFGCE model, as evidenced by its dominant ROC curve positioning that consistently envelopes competing models' trajectories. The Australian dataset analysis reveals LFGCE's dark red ROC curve achieving near-complete coverage of other models' performance spaces, with deep forest and XGBoost exhibiting proximate yet inferior trajectories, while traditional approaches like LDA and KNN cluster near the random classifier baseline, establishing a clear performance hierarchy where modern ensemble methods significantly outperform conventional algorithms. This pattern persists across geographical datasets, with the German dataset showing LFGCE maintaining classification supremacy while deep forest and LightGBM demonstrate competitive performance, contrasted against traditional models' markedly in-

ferior discriminative capacity evidenced by their depressed ROC trajectories. The Japanese dataset further validates LFGCE's robust discriminative power through its comprehensive coverage of other models' curves, accompanied by stable performances from XGBoost-enhanced deep forest and LightGBM implementations, whereas traditional methods exhibit critical deficiencies in positive sample identification. Taiwan dataset results complete this consistent performance pattern, with LFGCE and deep forest forming a high-performance tier while KNN and DT approach random classification levels. The comprehensive experimental evidence establishes LFGCE as the optimal solution for credit scoring tasks, followed by deep forest and LightGBM, with their superior AUC metrics and ROC curve characteristics confirming the transformative potential of ensemble and deep learning architectures in handling complex financial risk assessment challenges, while relegating traditional models to simpler application scenarios due to their fundamental limitations in discriminative performance.



(**a**) Australian (**b**) German

(**c**) Japan (**d**) Taiwan

Figure 2. ROCs of credit scoring models for credit scoring datasets.

The quantitative evaluation of credit scoring algorithms on the Australian dataset, as presented in Table 3, reveals a clear performance hierarchy where the proposed LFGCE(XGBoost) model demonstrates superior discriminative capability with an AUC of 0.9411, marginally outperforming deep forest and XGBoost, while traditional methods such as LR and RF exhibit intermediate performance, and KNN with DT show limited effectiveness. In addition, LFGCE achieves an optimal Brier score, indicating reliable probability estimation, though deep forest demonstrates marginally superior precision in positive class identification, while traditional models like LDA excel in recall at the expense of other metrics. The comprehensive F1 scores further validate LFGCE's balanced performance, comparable to LDA but significantly surpassing other traditional approaches, establishing modern ensemble methods as fundamentally superior for complex credit scoring tasks requiring both discriminative power and probabilistic calibration, with traditional models retaining niche advantages in specific metrics but demonstrating systematic limitations in overall performance.

Table 3. Performance comparison of credit scoring models for the Australian dataset.

Algorithm	AUC	Acc	Pre	Rec	BS	F1
LDA	0.9269	0.8594	0.7961	**0.9196**	0.1089	**0.8534**
LR	0.9298	0.8649	0.8309	0.8741	0.0992	0.8520
DT	0.9140	0.8437	0.8270	0.8202	0.1112	0.8236
KNN	0.9134	0.8494	0.8640	0.7851	0.1112	0.8227
SVM	0.9262	0.8626	0.8497	0.8395	0.1008	0.8446
NN	0.9148	0.8502	0.8328	0.8298	0.1186	0.8313
RF	0.9338	0.8645	0.8575	0.8341	0.1032	0.8457
AdaBoost	0.9273	0.8555	0.7913	0.9173	0.1516	0.8496
GBDT	0.9392	0.8637	0.8426	0.8530	0.0956	0.8478
LightGBM	0.9371	0.8624	0.8476	0.8421	0.0964	0.8448
XGBoost	0.9394	0.8633	0.8449	0.8487	0.0991	0.8468
Deep Forest	0.9382	**0.8725**	**0.8763**	0.8306	0.1036	0.8528
LFGCE	**0.9411**	0.8687	0.8478	0.8591	**0.0908**	**0.8534**

Note: the best ranking values are in bold.

Table 4 presents a comprehensive performance evaluation of credit scoring models on the German dataset, where LFGCE(XGBoost) demonstrates superior discriminative capability with an AUC of 0.7856, significantly outperforming traditional approaches like DT and KNN that exhibit fundamental limitations in handling complex data patterns. While NN achieves marginally better accuracy and the highest precision, LFGCE maintains competitive performance across all metrics. The probabilistic calibration analysis reveals LFGCE's optimal performance with the lowest Brier score, closely followed by LR and NN, whereas AdaBoost's extreme recall (0.9941) comes at the cost of poor calibration (BS:0.1979), demonstrating the proposed model's unique advantage in maintaining both discriminative power and probabilistic reliability. This comprehensive superiority is further confirmed by LFGCE's leading F1 score, establishing it as the optimal solution that consistently outperforms both traditional models and contemporary alternatives like XGBoost and LightGBM, with its hybrid architecture of ensemble and deep learning techniques proving particularly effective for credit scoring tasks demanding high classification accuracy and robust probability estimation.

Table 4. Performance comparison of credit scoring models for the German dataset.

Algorithm	AUC	Acc	Pre	Rec	BS	F1
LDA	0.7795	0.7585	0.7926	0.8871	0.1653	0.8372
LR	0.7808	0.7601	0.7942	0.8872	0.1646	0.8381
DT	0.7096	0.7232	0.7791	0.8439	0.1928	0.8102
KNN	0.7383	0.7280	0.7341	0.9586	0.1803	0.8315
SVM	0.7112	0.7065	0.7966	0.7798	0.1859	0.7881
NN	0.7799	**0.7659**	**0.8067**	0.8753	0.1650	0.8396
RF	0.7702	0.7437	0.7545	0.9394	0.1713	0.8369
AdaBoost	0.7035	0.7021	0.7031	**0.9941**	0.1979	0.8237
GBDT	0.7792	0.7587	0.7879	0.8967	0.1654	0.8388
LightGBM	0.7776	0.7615	0.7887	0.9007	0.1657	0.8410
XGBoost	0.7811	0.7582	0.7757	0.9208	0.1652	0.8420
Deep Forest	0.7755	0.7447	0.7526	0.9471	0.1709	0.8387
LFGCE	**0.7856**	0.7643	0.7856	0.9126	**0.1621**	**0.8444**

Note: the best ranking values are in bold.

Table 5 presents a comprehensive performance evaluation of credit scoring models on the Japanese credit dataset, revealing distinct characteristics across different algorithmic approaches. Among traditional machine learning models, while LDA demonstrates supe-

rior precision in minimizing false positives of normal users misclassified as defaulters, its compromised recall rate results in potential defaulter omissions that elevate risk exposure. Similarly, SVM maintains high precision but underperforms in holistic metrics including AUC and F1 when compared to ensemble methods. Both DT and KNN exhibit limited predictive capability due to their inherent inability to capture complex credit risk patterns, whereas NN delivers mediocre performance across all metrics. In contrast, ensemble models demonstrate marked improvements, with XGBoost and GBDT showing strong AUC and F1 performance, though still marginally surpassed by the proposed LFGCE model. RF achieves optimal accuracy and F1. The deep forest model attains maximal recall for defaulter identification at the expense of precision-induced misclassification costs, rendering it particularly suitable for risk-averse scenarios. Notably, the LFGCE model establishes state-of-the-art performance through leading AUC and BS, achieving optimal equilibrium between misclassification cost control and risk exposure minimization. XGBoost and deep forest constitute competitive alternatives, collectively demonstrating that modern ensemble methods substantially outperform traditional models across multiple evaluation dimensions. Although conventional approaches like LDA and SVM exhibit isolated strengths in specific metrics, their comprehensive performance remains constrained. The empirical results confirm that LFGCE's integration of ensemble learning and deep learning paradigms yields superior capability in processing complex financial data structures.

Table 5. Performance comparison of credit scoring models for the Japan dataset.

Algorithm	AUC	Acc	Pre	Rec	BS	F1
LDA	0.9127	0.8606	**0.9402**	0.7997	0.1136	0.8643
LR	0.9156	0.8549	0.9175	0.8116	0.1030	0.8613
DT	0.9134	0.8490	0.8698	0.8561	0.1123	0.8629
KNN	0.9111	0.8487	0.8862	0.8345	0.1108	0.8596
SVM	0.8682	0.8566	0.9311	0.8008	0.1175	0.8611
NN	0.9177	0.8487	0.8895	0.8310	0.1048	0.8593
RF	0.9319	**0.8694**	0.8719	0.8964	0.1056	**0.8840**
AdaBoost	0.9210	0.8548	0.9293	0.7993	0.1509	0.8594
GBDT	0.9362	0.8642	0.8898	0.8625	0.0960	0.8759
LightGBM	0.9349	0.8634	0.8829	0.8696	0.0957	0.8762
XGBoost	0.9362	0.8678	0.8959	0.8625	0.0950	0.8789
Deep Forest	0.9324	0.8679	0.8687	**0.8982**	0.1064	0.8832
LFGCE	**0.9374**	0.8673	0.8954	0.8620	**0.0944**	0.8784

Note: the best ranking values are in bold.

Table 6 shows the performance comparison of various credit scoring models on the Taiwan credit dataset. The experimental results show that the LFGCE achieves the highest AUC of 0.7508, slightly outperforming RF and XGBoost, indicating its superior ability to distinguish between defaulting and non-defaulting users. NN and GBDT follow closely in terms of AUC. The AUC of LDA and LR are the lowest, indicating the weak discriminative power of traditional linear models on this dataset. The LFGCE has the highest accuracy, closely followed by LightGBM, XGBoost, and RF. NN also achieves a relatively high accuracy. The lowest accuracy is obtained by LDA and LR, which again highlights the limitations of traditional models on complex datasets. The AdaBoost model has the highest precision, indicating the lowest misjudgment rate in predicting defaulters and effectively controlling the misjudgment cost. The precision of SVM and LFGCE closely follows. KNN and LDA perform poorly in terms of precision. The deep forest has the highest recall. LightGBM and RF also have relatively high recall. SVM and AdaBoost have the lowest recall, implying a higher likelihood of missing defaulters. LFGCE has the lowest BS, indicating higher prediction reliability. The BS of RF and LightGBM are also relatively lower.

The highest BS is obtained by LDA. The F1 of LFGCE is the highest, slightly outperforming LightGBM and RF. The F1 of deep forest and GBDT are also close to optimal. Overall, LFGCE performs the best on the Taiwan credit dataset, leading in AUC, BS, and F1. The comprehensive performance of deep forest is also very similar and can be used as an alternative solution. The performance of traditional models on complex credit data is significantly inferior to modern ensemble learning models and deep learning models.

Table 6. Performance comparison of credit scoring models for the Taiwan dataset.

Algorithm	AUC	Acc	Pre	Rec	BS	F1
LDA	0.6985	0.6512	0.6676	0.6023	0.2183	0.6333
LR	0.6999	0.6486	0.6612	0.6099	0.2179	0.6345
DT	0.7199	0.6666	0.6851	0.6167	0.2152	0.6491
KNN	0.7169	0.6675	0.7163	0.5550	0.2135	0.6254
SVM	0.7058	0.6731	0.7561	0.5114	0.2140	0.6101
NN	0.7377	0.6806	0.7131	0.6044	0.2061	0.6543
RF	0.7502	0.6949	0.7298	0.6192	0.2011	0.6700
AdaBoost	0.7170	0.6728	**0.7599**	0.5053	0.2169	0.6070
GBDT	0.7496	0.6948	0.7297	0.6189	0.2009	0.6698
LightGBM	0.7494	0.6954	0.7292	0.6218	0.2010	**0.6712**
XGBoost	0.7504	0.6945	0.7301	0.6175	0.2006	0.6691
Deep Forest	0.7479	0.6904	0.7189	**0.6257**	0.2026	0.6690
LFGCE	**0.7508**	**0.6957**	0.7356	0.6114	**0.2003**	0.6677

Note: the best ranking values are in bold.

5.2. Significance Test

To test whether the proposed LFGCE model statistically outperforms other loan default algorithms, in this study, the Friedman test [41] was first performed to identify whether there was a significant difference between LFGCE and other models. The Friedman test is a non-parametric test that performs statistical significance tests based on the rankings of loan default predictors. The statistical value of the Friedman test can be computed as

$$\chi^2 = \frac{12D}{J(J+1)}\left[\sum_{j=1}^{J} AvR_j^2 - \frac{J(J+1)^2}{4}\right] \tag{27}$$

where D is the number of credit datasets; J represents the number of classifiers that realize the credit scoring. R_j is the average rank of j-th classifier over credit datasets. Specifically, it refers to the ranking of classifiers on the ACC, AUC, Gmean, and BS, respectively. If the Friedman statistic value is larger than a critical value, the null hypothesis (there is no significant difference among loan default prediction models) is rejected. Next, a post hoc test procedure is followed to detect the performance difference between pairwise comparisons. In this study, the Nemenyi test [42] is performed to determine if the averaged rankings differ by at least a critical difference (CD). The CD is defined as

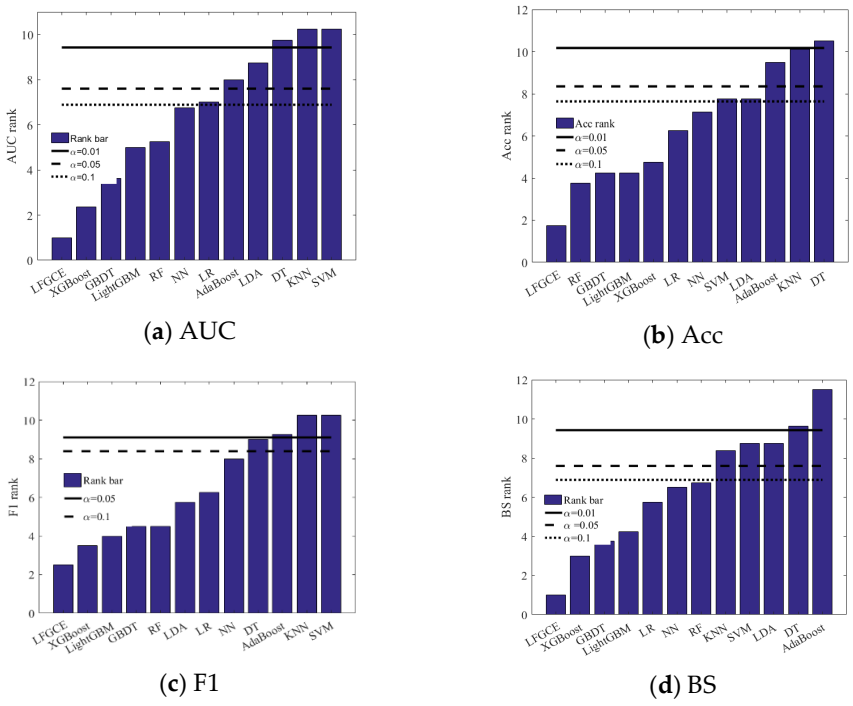
$$CD = q_\alpha \sqrt{\frac{J(J+1)}{12D}} \tag{28}$$

where CD is the critical difference at significance level α, q_α is the critical value at significance level α, which is computed from a studentized range.

To assess the statistical disparities among credit scoring frameworks, we initially calculate $\chi_F = 6.81$ using Equation (25), leading to the rejection of the null hypothesis at a significance threshold of 0.01. Subsequently, we employed the Nemenyi test for pairwise comparisons. Critical values were determined as $q_{0.01}$ = 3.82, $q_{0.05}$ = 3.35, and $q_{0.1}$ = 3.12,

which were then utilized to derive the critical differences (CDs) at various significance levels. Applying Equation (26), we obtain $CD_{0.01}$ = 7.99, $CD_{0.05}$ = 7.01, and $CD_{0.05}$ = 6.52.

Since AUC, ACC, F1, and BS fully cover the core performance requirements of credit scoring models from the perspectives of risk discrimination ability, overall prediction accuracy, adaptability to imbalanced data, and probability prediction accuracy. We conducted significance tests on them to systematically verify the practicality and superiority of the LFGCE algorithm in credit evaluation scenarios from a statistical perspective. Figure 3 shows a comparison of the significance tests of each algorithm on four comprehensive indicators.



(**a**) AUC (**b**) Acc

(**c**) F1 (**d**) BS

Figure 3. Results of significance test.

Figure 3a shows the significance test analysis on the AUC. As can be seen from Figure 3a, the AUC ranking of LFGCE is significantly lower than the three significance level lines. Its rank is below the extremely strict significance line $\alpha = 0.01$, which contrasts with algorithms such as SVM, KNN, and DT above the line, indicating that LFGCE has a significant advantage in distinguishing high and low-risk customers. This means that in credit risk stratification, LFGCE can more accurately identify high-risk customers and reduce bad debt losses for financial institutions. In addition, the ranking of AUC also reflects the relative performance of other models, with some models, such as RF and XGBoost, also performing well, but still unable to reach the level of LFCGE models. This result indicates that the LFCGE model can provide more accurate predictive capabilities when dealing with complex datasets, demonstrating its potential in practical applications.

Figure 3b shows the significance test analysis on the Acc. In terms of the Acc metric, the rank of LFGCE persistently falls below all significance level lines, forming a sharp contrast with algorithms above the threshold line, such as k-NN and DT. Taking $\alpha = 0.05$ as an example, its rank resides below the significance line, demonstrating that LFGCE achieves statistically significant superiority over online algorithms (AdaBoost, KNN, and DT) in overall predictive accuracy for credit data. This enhanced performance effectively reduces the misclassification of both high-risk and low-risk clients, thereby fulfilling financial institutions' requirement for "precise categorization". It is worth noting that the performance of other models is relatively stable, although some models, such as RF and GBDT, also

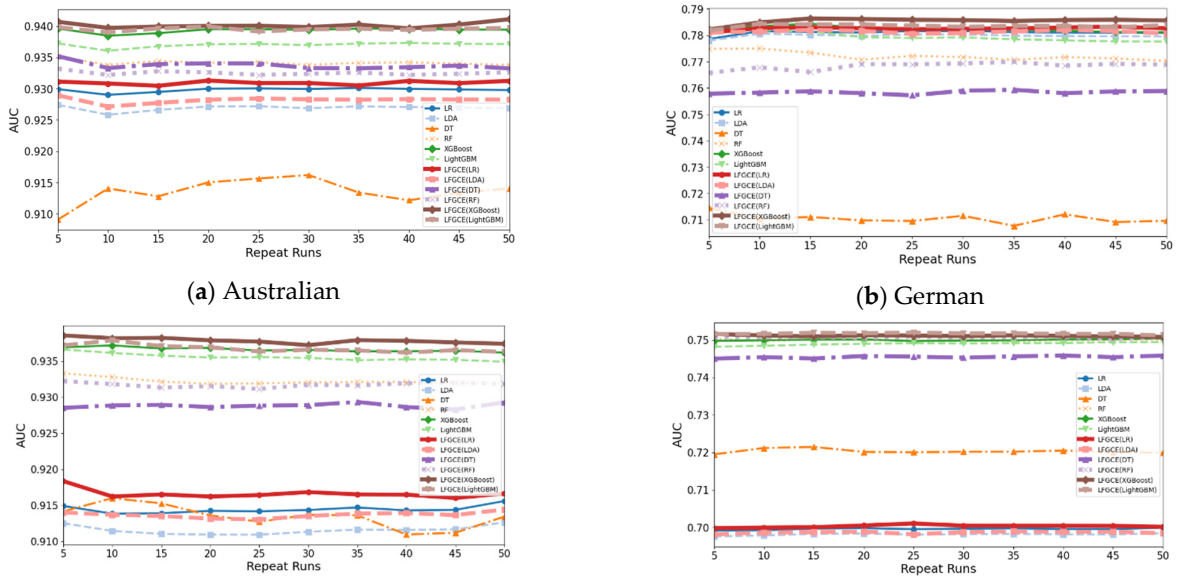
show good accuracy; they still appear insufficient compared to the LFCGE model. This result emphasizes the advantages of the LFCGE model in reducing misclassification and improving the accuracy of classification.

Figure 3c shows the significance test analysis on the F1. As shown in Figure 3c, the F1 rank of LFGCE consistently resides below all significance level lines, forming a striking contrast with algorithms above the significance threshold, such as LR and LDA. Taking $\alpha = 0.01$ as an example, its rank offline advantage indicates that when dealing with credit data imbalance, LFGCE has a significantly better ability to accurately capture high-risk customers (balancing accuracy and recall) than online algorithms, avoiding losses caused by missed or misjudgments. In addition, LFGCE significantly outperforms DT, AdaBoost, KNN, and SVM in balancing positive-negative sample classification at a confidence level of $\alpha = 0.1$. This disparity directly impacts the quality of high-risk client identification, making LFGCE more valuable for business applications.

Figure 3d shows the significance test analysis on the BS. The BS ranking of LFGCE is lower than the significance level lines, and compared with online algorithms such as AdaBoost, it has a significant advantage in probability prediction accuracy. LFGCE outperforms KNN, SVM, LDA, DT, AdaBoost, and other algorithms at a significance level of $\alpha = 0.05$, indicating that its predicted default probabilities align more closely with empirical observations. This alignment provides reliable foundations for critical financial decision-making processes, including interest rate pricing and credit approval. While other models, such as LR and XGBoost, achieve commendable BS performance, they have not yet reached the level demonstrated by the LFCGE model. The lower BS value validates that the LFCGE model provides reliable prediction probabilities in practical applications.

In the context of credit risk evaluation, the stability and reliability of credit scoring algorithms are crucial for accurate assessment outcomes. Given the inherent complexity of financial datasets and the heterogeneous data processing mechanisms across algorithms, single experimental results may exhibit statistical variability and insufficient generalizability. To mitigate these limitations, a systematic validation framework was implemented through multiple repeated experiments with diverse base classifiers (e.g., decision trees, SVM, and neural networks). This methodology enabled a comprehensive performance evaluation under varying data distributions, ultimately identifying algorithms demonstrating statistically robust classification accuracy and consistent stability across different credit cohorts. The empirically validated approach provides enhanced operational reliability for financial institutions' decision-making processes.

To verify the stability of LFGCE(XGBoost) and the effectiveness of the LFGCE, we test the performance of the LFGCE framework with different base learners, with repeated experimental results. Figure 4 represents the AUC comparison of popular credit scoring models such as LR, LDA, DR, XGBoost, and LightGBM, and their advanced versions that include LFGCE(LR), LFGCE(LDA), LFGCE(DT), LFGCE(XGBoost), and LFGCE(LightGBM). As can be seen from Figure 4a, the AUC value of the LFGCE(XGBoost) algorithm, using XGBoost as the base learner, remains stable at around 0.940 with slight fluctuations. In contrast, LFGCE(LightGBM), with LightGBM as the base learner, has a higher AUC value but is slightly lower than LFGCE(XGBoost). Traditional DT and LR algorithms have AUC values around 0.91 and 0.930, respectively. This indicates that LFGCE(XGBoost) has a stronger ability to capture and utilize data features on the Australian dataset, distinguishing high-risk and low-risk customers more precisely and showing superior credit scoring performance over some other algorithms.



(**a**) Australian (**b**) German

(**c**) Japan (**d**) Taiwan

Figure 4. AUC comparison under different repeated experimental times.

As can be seen from Figure 4b, the AUC value of LFGCE(XGBoost) is slightly greater than 0.78, which relatively high level compared to other algorithms. LR and LFGCE(DT) algorithms have AUC values of 0.78 and about 0.76, respectively. This shows that LFGCE(XGBoost) can better adapt to the data distribution of the German dataset, excelling in credit risk differentiation and offering more reliable credit evaluation results for financial institutions.

As can be seen in Figure 4c, both LFGCE(XGBoost) and LFGCE(LightGBM) maintain a value above 0.935, outperforming the other algorithms. Of the two, LFGCE(XGBoost) demonstrates superior curve stability. The AUC value of the LR algorithm falls between 0.91 and 0.92, while that of the LDA algorithm is also relatively low. This indicates that in the credit scoring tasks on the Japan dataset, both LFGCE algorithms with XGBoost and LightGBM as base learners perform exceptionally well, with LFGCE(XGBoost) offering better stability and more precise credit risk assessment than traditional algorithms.

As can be seen from Figure 4d, LFGCE(XGBoost) maintains an AUC value above 0.75 with a stable trend. The DT algorithm's AUC value is approximately 0.72, which is lower than LFGCE(XGBoost). This indicates that on the Taiwan dataset, LFGCE(XGBoost) can process data more effectively, evaluate customer credit risk more accurately, and outperform other base learning algorithms in terms of credit scoring performance.

6. Discussion

Although the LFGCE framework has achieved significant performance improvements in large-scale credit scoring, it also faces several challenges. For instance, predefined hierarchical feature partitioning methods based on expert experience or simple correlation analysis may not be optimal, lacking adaptability to different credit scoring scenarios and data distributions, and making it difficult to effectively capture the diverse credit risk characteristics of borrowers. Additionally, the fixed cascading integration strategy may fail to fully capture the complex dependencies between levels, which limits the model's ability to finely characterize and predict credit risks. Therefore, future research will explore the following directions: (1) developing adaptive hierarchical feature partitioning methods based on the characteristics of credit scoring data, such as feature correlation, information gain, stability metrics, and the economic significance of variables, to dynamically adjust

feature subsets partitioning and avoid information redundancy or loss; (2) exploring more refined cascading ensemble learning strategies such as introducing diversity-enhancing mechanisms (e.g., different base learners or training sample perturbations) to build diverse hierarchical ensemble learner and avoid model homogenization, or incorporating attention mechanisms to learn the importance weights of different subsets of hierarchical features (e.g., demographic, credit history features, consumer behavior features, etc.) for optimized information fusion between layers; and (3) extending the LFGCE model to other areas of financial risk management, such as fraud detection and investment portfolio optimization, to validate its generalization and practical value and further expand its application prospects in financial risk management.

7. Conclusions

Credit scoring plays a crucial role in financial risk management, directly impacting the performance of credit decisions and the risk control level of financial institutions. This study aims to address the challenges faced by traditional credit scoring models in processing high-dimensional, heterogeneous credit data, particularly the difficulty in effectively capturing complex nonlinear relationships and hierarchical structures among features. Therefore, this article proposes the level-wise feature-guided cascaded ensemble (LFGCE) model. The model first uses a hierarchical feature selection strategy to divide the high-dimensional feature space into multiple hierarchical feature subsets based on feature importance, effectively reducing dimensionality and focusing on key information. It then uses a cascaded ensemble learning architecture to build and train ensemble learners layer by layer, using the output of the previous-layer learners as input features for the subsequent layer to achieve information transmission and fusion between layers, thereby gradually enhancing the predictive performance of the model. Experiment results on multiple public credit scoring datasets show that LFGCE outperforms traditional machine learning models and single ensemble models, validating its effectiveness and superiority in credit scoring tasks. Moreover, the hierarchical structure of the LFGCE model enhances its interpretability, revealing the impact of different hierarchical feature subsets on credit risk and providing a more refined perspective for risk assessment in financial institutions.

The LFGCE model offers broad applicability across financial contexts through its adaptive hierarchical feature selection and modular cascaded architecture. Its automated feature processing handles diverse lending scenarios while maintaining interpretability—crucial for regulatory compliance. The framework demonstrates consistent performance across market conditions and borrower profiles, with computational efficiency enabling practical deployment. This combination of adaptability, accuracy and explainability will make it particularly valuable for modern credit risk management, supporting applications from retail lending to complex corporate financing decisions. The model's balanced approach addresses key industry needs for both predictive power and operational transparency.

Author Contributions: Conceptualization, Y.Z. and G.C.; methodology, Y.Z.; software, Y.Z.; validation, Y.Z.; formal analysis, G.C. and Y.Z.; resources, G.C.; data curation, Y.Z.; writing—original draft preparation, Y.Z.; writing—review and editing, G.C. and Y.Z.; supervision, G.C. All authors have read and agreed to the published version of the manuscript.

Funding: The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Data Availability Statement: All the datasets involved are sourced from the UCI repository https://archive.ics.uci.edu/on 1 December 2024.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Dastile, X.; Celik, T.; Potsane, M. Statistical and Machine Learning Models in Credit Scoring: A Systematic Literature Survey. *Appl. Soft Comput.* **2020**, *91*, 106263. [CrossRef]
2. Eisenbeis, R.A. Problems in Applying Discriminant Analysis in Credit Scoring Models. *J. Bank. Financ.* **1978**, *2*, 205–219. [CrossRef]
3. Sohn, S.Y.; Kim, D.H.; Yoon, J.H. Technology Credit Scoring Model with Fuzzy Logistic Regression. *Appl. Soft Comput.* **2016**, *43*, 150–158. [CrossRef]
4. Runchi, Z.; Liguo, X.; Qin, W. An Ensemble Credit Scoring Model Based on Logistic Regression with Heterogeneous Balancing and Weighting Effects. *Expert Syst. Appl.* **2023**, *212*, 118732. [CrossRef]
5. Ogundimu, E.O. On Lasso and Adaptive Lasso for Non-Random Sample in Credit Scoring. *Stat. Model.* **2024**, *24*, 115–138. [CrossRef]
6. Montevechi, A.A.; de Carvalho Miranda, R.; Medeiros, A.L.; Montevechi, J.A.B. Advancing Credit Risk Modelling with Machine Learning: A Comprehensive Review of the State-of-the-Art. *Eng. Appl. Artif. Intell.* **2024**, *137*, 109082. [CrossRef]
7. Gambacorta, L.; Huang, Y.; Qiu, H.; Wang, J. How Do Machine Learning and Non-Traditional Data Affect Credit Scoring? New Evidence from a Chinese Fintech Firm. *J. Financ. Stab.* **2024**, *73*, 101284. [CrossRef]
8. Liu, Y.; Baals, L.J.; Osterrieder, J.; Hadji-Misheva, B. Leveraging Network Topology for Credit Risk Assessment in P2P Lending: A Comparative Study under the Lens of Machine Learning. *Expert Syst. Appl.* **2024**, *252*, 124100. [CrossRef]
9. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
10. Liu, W.; Fan, H.; Xia, M. Step-Wise Multi-Grained Augmented Gradient Boosting Decision Trees for Credit Scoring. *Eng. Appl. Artif. Intell.* **2021**, *97*, 104036. [CrossRef]
11. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
12. Mienye, I.D.; Sun, Y. A Deep Learning Ensemble with Data Resampling for Credit Card Fraud Detection. *IEEE Access* **2023**, *11*, 30628–30638. [CrossRef]
13. Yang, M.; Lim, M.K.; Qu, Y.; Li, X.; Ni, D. Deep Neural Networks with L1 and L2 Regularization for High Dimensional Corporate Credit Risk Prediction. *Expert Syst. Appl.* **2023**, *213*, 118873. [CrossRef]
14. Xiao, J.; Zhong, Y.; Jia, Y.; Wang, Y.; Li, R.; Jiang, X.; Wang, S. A Novel Deep Ensemble Model for Imbalanced Credit Scoring in Internet Finance. *Int. J. Forecast.* **2024**, *40*, 348–372. [CrossRef]
15. Myers, J.H.; Forgy, E.W. The Development of Numerical Credit Evaluation Systems. *J. Am. Stat. Assoc.* **1963**, *58*, 799–806. [CrossRef]
16. Orgler, Y.E. A Credit Scoring Model for Commercial Loans. *J. Money Credit Bank.* **1970**, *2*, 435–445. [CrossRef]
17. Fitzpatrick, D.B. An Analysis of Bank Credit Card Profit. *J. Bank Res.* **1976**, *7*, 199–205.
18. Wiginton, J.C. A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior. *J. Financ. Quant. Anal.* **1980**, *15*, 757–770. [CrossRef]
19. Çetin, A.İ.; Büyüklü, A.H. A New Approach to K-Nearest Neighbors Distance Metrics on Sovereign Country Credit Rating. *Kuwait J. Sci.* **2025**, *52*, 100324. [CrossRef]
20. Su, J.-H. Utility-Maximizing Binary Prediction via the Nearest Neighbor Method and Its Application to Credit Scoring. *J. Bus. Econ. Stat.* **2025**, 1–23. [CrossRef]
21. Sohn, S.Y.; Kim, J.W. Decision Tree-Based Technology Credit Scoring for Start-up Firms: Korean Case. *Expert Syst. Appl.* **2012**, *39*, 4007–4012. [CrossRef]
22. Liu, T.; Yang, L. Financial Risk Early Warning Model for Listed Companies Using BP Neural Network and Rough Set Theory. *IEEE Access* **2024**, *12*, 27456–27464. [CrossRef]
23. Ma, Z.; Hou, W.; Zhang, D. A Credit Risk Assessment Model of Borrowers in P2P Lending Based on BP Neural Network. *PLoS ONE* **2021**, *16*, e0255216. [CrossRef] [PubMed]
24. Benítez-Peña, S.; Blanquero, R.; Carrizosa, E.; Ramírez-Cobo, P. Cost-Sensitive Probabilistic Predictions for Support Vector Machines. *Eur. J. Oper. Res.* **2024**, *314*, 268–279. [CrossRef]
25. Shen, F.; Yang, Z.; Zhao, X.; Lan, D. Reject Inference in Credit Scoring Using a Three-Way Decision and Safe Semi-Supervised Support Vector Machine. *Inf. Sci.* **2022**, *606*, 614–627. [CrossRef]
26. Yang, D.; Xiao, B.; Cao, M.; Shen, H. A New Hybrid Credit Scoring Ensemble Model with Feature Enhancement and Soft Voting Weight Optimization. *Expert Syst. Appl.* **2024**, *238*, 122101. [CrossRef]
27. Lu, Z.; Li, H.; Wu, J. Exploring the Impact of Financial Literacy on Predicting Credit Default among Farmers: An Analysis Using a Hybrid Machine Learning Model. *Borsa Istanb. Rev.* **2024**, *24*, 352–362. [CrossRef]
28. Blanco, A.; Pino-Mejías, R.; Lara, J.; Rayo, S. Credit Scoring Models for the Microfinance Industry Using Neural Networks: Evidence from Peru. *Expert Syst. Appl.* **2013**, *40*, 356–364. [CrossRef]
29. Harris, T. Credit Scoring Using the Clustered Support Vector Machine. *Expert Syst. Appl.* **2015**, *42*, 741–750. [CrossRef]

30. Kang, Y.; Chen, L.; Jia, N.; Wei, W.; Deng, J.; Qian, H. A CWGAN-GP-Based Multi-Task Learning Model for Consumer Credit Scoring. *Expert Syst. Appl.* **2022**, *206*, 117650. [CrossRef]
31. Kearns, M. *Learning Boolean Formulae or Finite Automata Is as Hard as Factoring*; Technical Report TR-14-88; Harvard University Aikem Computation Laboratory: Cambridge, MA, USA, 1988.
32. Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
33. Wolpert, D.H. Stacked Generalization. *Neural Netw.* **1992**, *5*, 241–259. [CrossRef]
34. Luo, C. A Comparison Analysis for Credit Scoring Using Bagging Ensembles. *Expert Syst.* **2022**, *39*, e12297. [CrossRef]
35. Plaia, A.; Buscemi, S.; Fürnkranz, J.; Mencía, E.L. Comparing Boosting and Bagging for Decision Trees of Rankings. *J. Classif.* **2022**, *39*, 78–99. [CrossRef]
36. Liu, W.; Fan, H.; Xia, M. Multi-Grained and Multi-Layered Gradient Boosting Decision Tree for Credit Scoring. *Appl. Intell.* **2022**, *52*, 5325–5341. [CrossRef]
37. Rao, C.; Liu, Y.; Goh, M. Credit Risk Assessment Mechanism of Personal Auto Loan Based on PSO-XGBoost Model. *Complex Intell. Syst.* **2023**, *9*, 1391–1414. [CrossRef]
38. Mushava, J.; Murray, M. Flexible Loss Functions for Binary Classification in Gradient-Boosted Decision Trees: An Application to Credit Scoring. *Expert Syst. Appl.* **2024**, *238*, 121876. [CrossRef]
39. Liu, W.; Fan, H.; Xia, M. Tree-Based Heterogeneous Cascade Ensemble Model for Credit Scoring. *Int. J. Forecast.* **2023**, *39*, 1593–1614. [CrossRef]
40. Yin, W.; Kirkulak-Uludag, B.; Zhu, D.; Zhou, Z. Stacking Ensemble Method for Personal Credit Risk Assessment in Peer-to-Peer Lending. *Appl. Soft Comput.* **2023**, *142*, 110302. [CrossRef]
41. Lessmann, S.; Baesens, B.; Seow, H.-V.; Thomas, L.C. Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring: An Update of Research. *Eur. J. Oper. Res.* **2015**, *247*, 124–136. [CrossRef]
42. Demšar, J. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.