ISSN 2090-3359 (Print)
ISSN 2090-3367 (Online)

# Advances in Decision Sciences

*Volume 30*
*Issue 3*
*September 2026*



亞洲大學 ASIA UNIVERSITY
SCIENTIFIC & BUSINESS WORLD

Published by Asia University, Taiwan and Scientific and Business World

# Enhancement of Digital Credit Scoring in P2P Lending using a Robust Hybrid Voting–Stacking Ensemble Framework

**Mohamed Galal**
Faculty of Computer and Information Sciences,
Ain Shams University, Cairo, Egypt
***Corresponding author* Email:** mhdgalal@yahoo.com
ORCID: 0000-0003-2314-7219

**Sherine Rady**
Faculty of Computer and Information Sciences,
Ain Shams University, Cairo, Egypt
**Email:** srady@cis.asu.edu.eg
ORCID: 0000-0003-4991-966X

**Mostafa Aref**
Faculty of Computer and Information Sciences,
Ain Shams University, Cairo, Egypt
**Email:** mostafa.aref@cis.asu.edu.eg
ORCID: 0000-0002-1278-0070

Received: October 4, 2024; First Revision: August 15, 2025;

Last Revision: April 26, 2026; Accepted: May 9, 2026;

**Published: May 14, 2026**

**Abstract**

**Purpose -** This study evaluates a hybrid ensemble framework designed to enhance credit-risk prediction accuracy in peer-to-peer (P2P) digital lending platforms and to overcome the limitations of existing models.

**Design/methodology/approach –** The proposed Hybrid Optimized Ensemble Learning (HOEL) framework integrates seven base classifiers through a three-level optimization process that involves K-fold cross-validation, hyperparameter tuning, an intermediate weighted-voting ensemble layer, and a final stacking meta-learner.

**Findings -** Empirical results demonstrate that the HOEL framework outperforms individual classifiers, achieving an ROC-AUC of 97.62% and an accuracy of 96.15%. The ensemble's layered design improves predictive stability and interpretability, confirming its robustness across small and high-dimensional datasets.

**Research limitations -** Although computationally intensive, the framework's performance can be further optimized using cloud-based or parallel processing. Future studies could incorporate additional behavioral features to improve model generalization.

**Practical implications -** The framework can be integrated into a trusted P2P digital lending platform to improve loan-approval decisions for customers with limited credit history, thereby reducing lending risks.

**Originality/value -** HOEL's novelty resides in its three-level optimization architecture, which embeds an intermediate weighted-voting layer that stabilizes ensemble predictions before they are passed to a stacking meta-learner. This combination has not previously been applied to imbalanced P2P credit risk scoring and delivers measurable gains in both ROC–AUC and accuracy over conventional ensemble approaches. The voting layer reduces hyperparameter sensitivity and strengthens stacking stability, making the framework particularly effective for small and high-dimensional financial datasets. This study thereby advances the Decision Sciences literature on quantitative risk modeling by providing a replicable, data-driven credit-assessment framework that supports more informed financial decision-making under uncertainty.

**Keywords:** Ensemble modeling, Machine learning, P2P lending, Credit Scoring, Stacking, Fintech.

**JEL Classifications:** G1, G2, C53

## 1. Introduction

The rapid evolution of financial technology (fintech) has significantly transformed how individuals and businesses access credit. Among the most prominent innovations is peer-to-peer (P2P) online lending, which enables borrowers to obtain loans directly from individual or institutional investors through digital platforms, bypassing traditional banks. This emerging model has expanded global access to financing, particularly for underserved populations such as small and medium enterprises (SMEs) and individuals lacking formal credit histories. However, a persistent challenge in this domain is the accurate assessment of borrower creditworthiness, which is critical to minimizing default risk and ensuring the long-term sustainability of P2P lending platforms (Chang et al., 2022).

Since the establishment of Zopa in the United Kingdom in 2005—the first P2P lending network—numerous platforms have emerged worldwide, including over 1,000 in China and industry leaders such as Lending Club in the United States. Lending Club's 2014 listing on the New York Stock Exchange underscored the sector's rapid growth. According to Verified Market Research (2024), the global P2P lending market was valued at USD 109.13 billion in 2023 and is projected to reach USD 559.73 billion by 2030. This exponential growth has been driven by the ability of P2P platforms to bridge capital access gaps, attract diverse investor profiles, and offer streamlined digital lending services (Jayaram, 2024).

Despite this progress, significant challenges remain. The decentralized and less regulated nature of P2P platforms introduces substantial risks, particularly in assessing credit risk for non-traditional borrowers. Complex borrower profiles, limited financial histories, and the lack of standardized assessment tools contribute to information asymmetry and credit rationing—making it difficult to match capital supply with appropriate demand (Abbasi et al., 2021). Credit risk, in particular, remains the foremost concern for investors and platform sustainability.

To address these challenges, fintech firms are increasingly adopting digital credit scoring models that leverage advanced analytics and machine learning (ML) to improve risk prediction. One promising direction involves hybrid modeling that combines the predictive power of multiple techniques to achieve superior accuracy. Ensemble methods like bagging and boosting aggregate predictions from base classifiers such as decision trees or gradient boosting machines, thereby mitigating overfitting and model bias. Stacking, a more sophisticated variant, introduces a meta-model that learns to combine predictions from base learners, further refining classification performance (Bone-Winkel & Reichenbach, 2024).

These techniques offer substantial benefits for P2P lending. Ensemble and stacking models can capture non-linear relationships across a wide feature space, incorporate diverse data sources, and produce more generalizable predictions. They also allow for automated handling of imbalanced data and reduce the dependency on any single classifier's limitations (Nguyen et al., 2024; Shih et al., 2022). In addition, these models provide deeper insights into the interactions among borrower attributes, aiding in decision-making, reducing adverse selection, and improving the accuracy of default prediction (Yeh et al., 2024).

This study contributes to the field of FinTech and digital lending by proposing an integrated optimized ensemble-stacking framework, HOEL (Hybrid Optimized Ensemble Learning). The proposed framework enhances decision-making processes in digital credit risk classification—an essential area where data-driven methods can inform financial decisions under uncertainty.

Despite prior applications of ensemble models in credit scoring, many existing approaches remain limited in handling imbalanced datasets and ensuring predictive stability. This study proposes the Hybrid Optimized Ensemble Learning (HOEL) framework, which is based on a three-level optimization architecture with an intermediate voting layer that stabilizes predictions prior to stacking, improving ROC–AUC and accuracy for imbalanced P2P credit-risk data.

This study is aligned with recent developments in decision sciences that emphasize the use of artificial intelligence and machine-learning techniques to improve decision-making under risk and uncertainty. Prior research highlights the growing role of AI in economic and financial decision-making (Hamori & Kume, 2018) and the importance of rigorous quantitative modeling in risk-based frameworks (Chang et al., 2018). The proposed HOEL framework contributes to this direction by enhancing credit-risk prediction through a robust hybrid ensemble learning approach. While recent advances in quantitative risk modeling have explored optimization-based frameworks — including large-scale portfolio optimization (Hui et al., 2024) and mean-generalized variance approaches applied to financial markets (Li et al., 2025) — the present study addresses a complementary but distinct problem: improving credit-risk classification accuracy in P2P lending through hybrid ensemble learning.

The remainder of this paper is structured as follows: Section 2 reviews literature review and related work; Section 3 presents the theoretical background; Section 4 presents the proposed methodology; Section 5 discusses the data and variables; Section 6 discusses diagnostic checks, results, and experimental validation; and Section 7 provides the conclusion, contributions, limitations, and directions for future research.

## 2. Literature Review

The emerging trend of fintech P2P online lending, revolutionizing the consumer credit market, allows individuals to lend and borrow money through an online platform without traditional financial institutions. Peer-to-peer online lending services are blooming in many parts of the world and have experienced rapid growth in lending volume in recent years (Noriega et al., 2023). P2P lending platforms, such as Zopa and Prosper Marketplace, emerged in 2005 and have evolved significantly since then. Some platforms have shifted from peer-to-peer lending to non-bank marketplaces or fintech lenders, while others have ceased operations. Prosper, founded in 2005, is the most prominent platform still in operation. It allowed borrowers to post online loan applications for three-year, unsecured loans of up to USD 25,000. The listings included hard financial information and soft information provided by the borrower, such as their reserve interest rate and personal information (Duarte et al., 2023). The importance and circumstances of

consumer credit and online lending practice vary by regions, but they are similar as to market potential, required credit underwriting capacity, and fierce competition among lenders (Trinh, 2024).

The expansion of online marketplaces, such as Lending Club, has been associated with positive changes in borrowers' financial positions, indicating that the increased outreach of peer-to-peer lending may be beneficial to individuals' credit positions. Moreover, P2P lending platforms have played a crucial role in promoting credit to small and medium-sized enterprises (SMEs), with the P2P business lending sector experiencing substantial growth in recent years (Coakley & Huang, 2020). The specialization of P2P platforms in providing services to particular types of borrowers, such as SMEs, and the use of non-standard information to facilitate the matching of investors and project owners have marked this evolution. As a result, P2P lending has contributed to wider access to the financial system for borrowers, thereby shaping the landscape of consumer credit markets (Najaf et al., 2021).

Meshref (2020) presented a new insight into Kaggle Bank Marketing data, analyzing loan approval prediction models. The best classification results were 84%, with Boosting algorithms outperforming other studies by 25%. The white-box model, designed as a white-box model, was informative and beneficial for high-stakes loan decisions. The findings are expected to benefit the machine learning community and bank marketing decision-makers, and the analysis approach could open new directions. Perera and Premaratne (2023) proposed a methodology for evaluating loan credit risks using data from a leading Sri Lankan finance institution. The study also employed a novel approach, voting-based ensemble learning, which involved multiple learners trained to forecast credit risk, resulting in better predictive accuracy. The Stacking Ensemble Classification outperformed other ML techniques with the highest training and test accuracy of 0.99 and 0.78, respectively, with a lower MSE of 0.21.

Wei et al. (2023) explored the use of machine-learning algorithms in credit default risk assessment for P2P lending platforms in China on 126,090 loan deals. It proposed a stacking ensemble-learning model, using a dataset from RenRen Dai, a large Chinese P2P platform. The model uses the Minimum Redundancy Maximum Relevance (MRMR) method for feature selection, with 12 features selected from 16 variables. The results show that the stacking ensemble model predicts credit default risk more accurately than individual classifiers. Kokate and Chetty (2021) presented a credit score model that predicts loan applicant status based on credit history. It uses Decision Tree and Gradient Boosting classifiers, achieving 80% accuracy. The model can be used in commercial banking to avoid future bankruptcy. The accuracy of the predicted model of the decision tree after applying the voting classifier is 0.78, whereas the accuracy of the gradient and ensemble learners is 0.81. The model works significantly well after the voting classifier. Uddin et al. (2023) developed a machine learning (ML)-based loan prediction system to identify suitable loan applicants. The system uses various ML models and deep neural networks. The system outperforms other models and achieves an impressive accuracy of 87.26%. The system has the potential to streamline loan approval processes, benefiting both financial institutions and loan applicants.

Abedin et al. (2022) proposed an ensemble approach called the weighted synthetic minority oversampling technique (WSMOTE) ensemble for small business credit risk assessment, addressing imbalanced default and nondefault classes. The ensemble classifier hybridizes WSMOTE and Bagging with sampling composite mixtures, ensuring robustness and variability. The study uses 3111 records from a Chinese commercial bank and found that the random forest classifier improved minority class accuracy by 15.16%. This study fills a knowledge gap in small business credit risk prediction. Yang (2024) introduced a comprehensive approach to improve bad loan prediction in P2P lending, sourcing Lending Club data. The methodology, which includes data cleaning, feature engineering, feature selection, and balancing the dataset and machine learning models, achieves an accuracy rate of over 92% and a recall rate of above 87%. This research advances academic understanding of loan prediction and improves decision-making and strategic planning in P2P lending platforms. Lenka et al. (2024) used machine learning algorithms to assess applicants' financial credibility before lending. However, high-dimensional and imbalanced datasets can degrade these models. A novel multiple-optimized ensemble learning (MOEL) model is proposed to build a reliable credit scoring model. The model's effectiveness was evaluated using six credit-scoring datasets. The empirical results demonstrate that MOEL achieves the best value of F1_score and G-mean with a mean ranking of 1.5 and 1.333, respectively.

Aleksandrova (2021) evaluated machine learning algorithms for peer-to-peer lending credit scoring, using a dataset from Lending Club. Results show ensemble classifiers outperform single ones, with Stacked Ensemble and XGBoost leading the way. The AUC score for stacked ensemble models was 0.7076. Zhao et al. (2024) presented a new framework for credit risk prediction, integrating benchmark and real private datasets. The research introduced a novel hybrid resampling framework, strategic hybrid SMOTE with double edited nearest neighbors, which demonstrates superior performance in handling extremely imbalanced datasets and substantially enhances the predictive capabilities of ensemble learning classifiers. Munsarif et al. (2022) proposed a stacking ensemble learning model using embedded techniques like gradient boosted trees, random forest, and LGBM to predict credit risk in peer-to-peer lending. The model achieves an average accuracy of 94.54% and a 69.10s execution time. The random forest meta-learner and Stacking ensemble model is the best classification model, while the LGBM meta-learner and stacking ensemble model has the fastest execution time.

Some researchers used ensemble modeling to combine several models, aiming at producing better accuracy in the prediction than that of a single one. This methodology, based on a divide-and-conquer strategy, is developed by constructing multiple classifiers and combining them via voting. Even complicated models often fail to accurately represent the data, either due to underfitting or overfitting (Wattanakitrungroj et al., 2024). Ensemble models could generate low bias and variance, meaning they are data-driven to achieve better generalization (Zhang & Sun, 2023). This study applied SVM and an ensemble method to achieve an accuracy of 94.05%. Hence, ensemble models have the potential to obtain more stable and comprehensive prediction performance compared to standalone models. The idea of a stacking model extends to an ensemble model. While the ensemble model combines multiple base models via, for example, simple voting, the stacking model embraces a higher-level model using the output of the base models as its input. As a result, predictions from base models create a new training set for related

high-level models. Stacking can theoretically represent any mapping from the base learner to the outcome by adapting higher-level functions like averaging or using a meta-model.

Despite the few introduced ensemble models in P2P digital lending, it appears that there is still a gap in enhancing the prediction accuracy, model automation, and generalization. It can be concluded that there is still a need for empowering the credit scoring classifiers and regarding the imbalance in datasets, as well as the proper feature engineering, which can be investigated with exploratory data analysis.

## 3. Theoretical Framework

This section reviews the theoretical foundations of the three core ensemble methods—bagging, boosting, and stacking—that underpin modern machine learning ensembles. As Dietterich (2000) demonstrated, ensembles of classifiers often achieve higher accuracy than individual models, making them central to advances in predictive modeling.

Breiman (1996) introduced bagging (bootstrap aggregating), which generates multiple versions of a predictor by training each on a different bootstrap-resampled subset of the data and then aggregating their outputs. By averaging over many such bootstrapped models, bagging reduces variance and can significantly improve accuracy, especially for unstable learners where small changes in the training set cause large changes in the model's output. His original experiments demonstrated substantial gains, particularly with decision trees, establishing bagging as a powerful method to reduce overfitting in high-variance models.

Schapire (1990) provided the seminal theoretical proof that a weak learner—one performing only slightly better than random guessing—can be boosted into a strong learner with arbitrarily high accuracy. Building on this foundation, Freund and Schapire (1997) introduced AdaBoost, which trains a base classifier and then iteratively trains subsequent classifiers on the instances misclassified by earlier models, increasing the weight of hard-to-predict examples at each iteration so that later models focus more on those cases. This sequential process reduces bias and improves overall accuracy, as the ensemble incrementally learns from its mistakes with each additional model. AdaBoost has since become a canonical example of boosting and has proven highly effective in both classification and regression tasks.

Wolpert (1992) introduced stacking, a method that combines multiple base learners through a second-level meta-learner. In a stacking framework, diverse first-level models are trained on the dataset, and then a higher-level learning algorithm is trained on the predictions (or outputs) of those base models to produce the final outcome. This method works by essentially deducing the biases of the individual base learners with respect to the training data and learning how to best integrate their contributions. In other words, instead of simply voting or averaging, stacking learns an optimal way to weight and combine the base models' predictions, often achieving better generalization performance than any single model or simpler ensemble strategy. By using a meta-learner to intelligently blend the strengths of each predictor, stacked generalization aims to minimize the overall error rate of the ensemble, and has been shown to outperform methods like straightforward cross-validation selection in combining models.

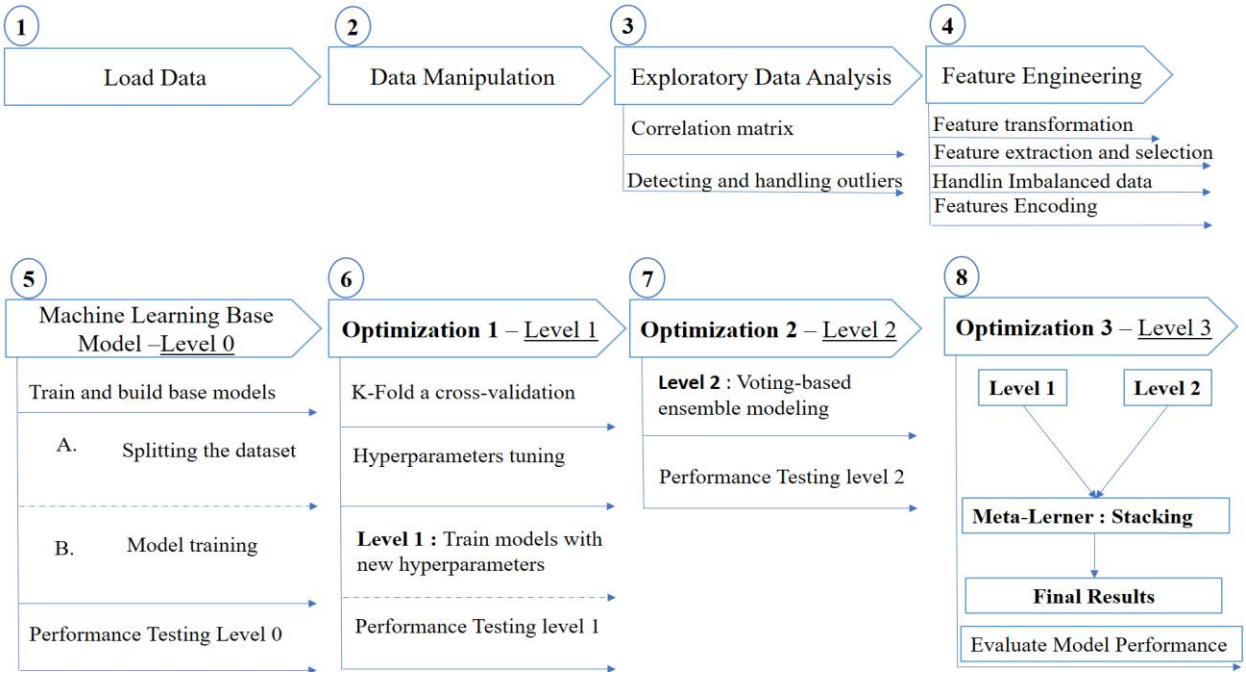
## 4. Methodology

This section describes the proposed design of the HOEL framework to enhance the digital credit scoring in P2P online lending platforms. HOEL framework uses hybrid ensemble classifiers and meta-learners to overcome the traditional frameworks' imbalanced data, model bias, model skewness, and model overall performance. The implementation of the HOEL framework works on gaining the maximum benefits from each component plugged into the framework. The subsections below will discuss the HOEL design, implementation, and evaluation.

### HOEL framework design

The HOEL framework is designed to handle diverse prediction use cases through a flexible, interpretable architecture. Its modular design enables easy adaptation to different data sources while ensuring transparency in predictions. Figure 1 presents the overall framework structure and its main processing blocks, which are described in detail in the following subsections.

**Figure 1**. Architecture of the Proposed HOEL Framework for P2P Lending Prediction




Note: This figure depicts the overall architecture of the HOEL framework, emphasizing its key components and optimization levels across the prediction pipeline, including data input, preprocessing, feature engineering, model optimization, and output interpretation. The framework incorporates a soft voting mechanism at Level 2 and a stacking meta-learner at Level 3 to progressively enhance predictive performance

As shown in Figure 1, the HOEL framework follows a structured multi-stage architecture that integrates data preprocessing, feature engineering, and hierarchical ensemble modeling. The framework is designed to progressively enhance predictive performance through three optimization levels, culminating in a hybrid voting–stacking ensemble that improves classification accuracy and robustness in P2P credit-risk assessment.

### *4.1 Data Preparation and Preprocessing*

The initial stage of implementing the HOEL framework involves data acquisition, manipulation, and transformation to ensure that the dataset is clean, consistent, and suitable for modeling. This process combines data loading, cleaning, feature engineering, and exploratory analysis — all critical for improving model interpretability, accuracy, and robustness. The key steps of this stage are described below.

#### 4.1.1 Data Loading

The data loading step entails importing and consolidating information from multiple heterogeneous sources into the HOEL framework. All data were processed using the Python pandas library within a Jupyter Notebook environment. This step ensures that the framework can seamlessly interface with different structured and semi-structured data formats, supporting diverse business use cases in peer-to-peer (P2P) lending.

#### 4.1.2 Data Manipulation

Data manipulation prepares the dataset for modeling by refining and standardizing its structure. The process begins with data preprocessing, which involves cleaning, removing duplicates, and handling missing values through statistical imputation techniques. Outlier treatment using the Interquartile Range (IQR) method prevents extreme values from distorting results. Normalization rescales numerical features to a common range, enhancing algorithm convergence. Finally, categorical encoding via one-hot encoding transforms qualitative variables into binary representations, improving interpretability and computational efficiency.

#### 4.1.3 Feature Engineering

Feature engineering enhances predictive performance by constructing, transforming, and selecting variables that effectively capture the underlying relationships within the data. In this study, feature creation was applied to generate new variables that enrich model learning. In addition, feature transformation was performed to convert the data into more informative and suitable representations for modeling. Feature extraction techniques were also utilized to condense the information into relevant dimensions while preserving essential patterns. Furthermore, feature importance and selection methods were employed to identify the predictors with the highest contribution to model accuracy. To better understand the relationships among variables, feature correlation analysis was conducted to visualize interdependencies and detect potential redundancy. Finally, imbalanced data handling techniques, including resampling and weighting strategies, were implemented to mitigate skewed class distributions and improve classification stability.

#### 4.1.4 Exploratory Data Analysis (EDA)

Exploratory Data Analysis provides an initial understanding of data distribution, variability, and relationships. It combines graphical visualization (e.g., histograms, boxplots, and correlation heatmaps) with statistical summaries to detect anomalies, assess patterns, and validate assumptions prior to model development. EDA in the HOEL framework thus serves as both a diagnostic and interpretive phase, guiding subsequent model selection and optimization.

### *4.2 Machine Learning Methodologies*

Model selection is a crucial step in machine learning that involves identifying the most appropriate algorithms for a given dataset. The HOEL framework employs a combination of traditional base learners and ensemble methods to enhance predictive performance and address limitations such as bias, variance, and model instability. The following subsections describe the selected models and ensemble strategies.

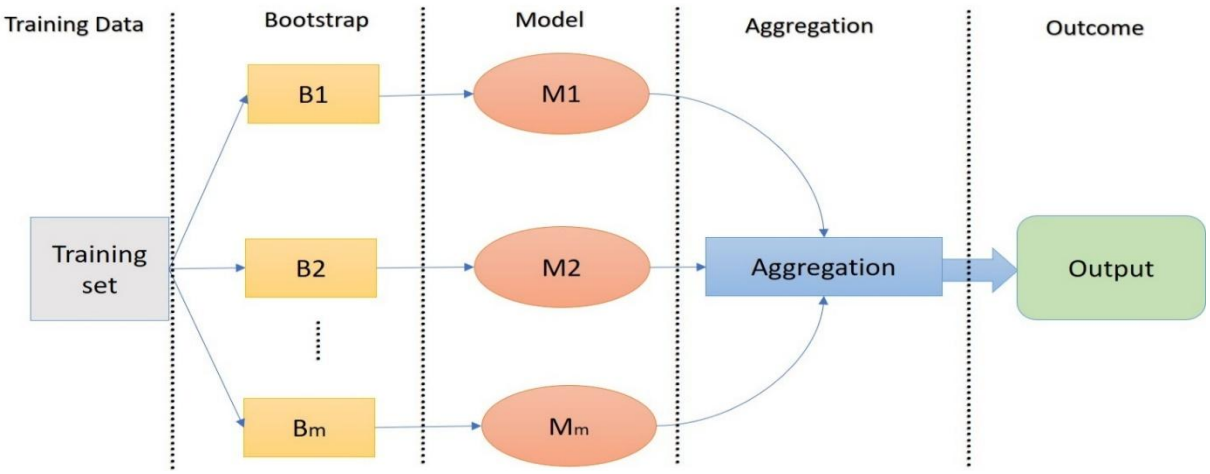
#### 4.2.1 Base ML Models Selection

The modeling process begins with a variety of conventional classifiers to establish a strong foundation for ensemble learning. Commonly adopted algorithms include Decision Trees, Logistic Regression, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN). These base learners provide diverse perspectives on the data and serve as individual contributors in subsequent ensemble stages. These models were chosen to provide diverse decision boundaries and statistical foundations for ensemble learning, ensuring the framework captures both linear and nonlinear relationships within the data.

#### 4.2.2 Bagging Ensemble Modeling

Bagging (Bootstrap Aggregating) is a machine learning technique designed to enhance model stability and accuracy by generating multiple iterations of a training dataset through bootstrap resampling. Each model is trained independently, and the final prediction is obtained by averaging or majority voting across models, thereby reducing variance and mitigating overfitting. Within the HOEL framework, bagging helps stabilize model performance when dealing with heterogeneous lending datasets and reduces variance among base learners. Figure 2 illustrates this process, showing the sequence of bootstrap sampling, parallel model training, and aggregation of results using Random Forest as a representative example (Madaan et al., 2021).

**Figure 2.** Bagging Ensemble Modeling Workflow



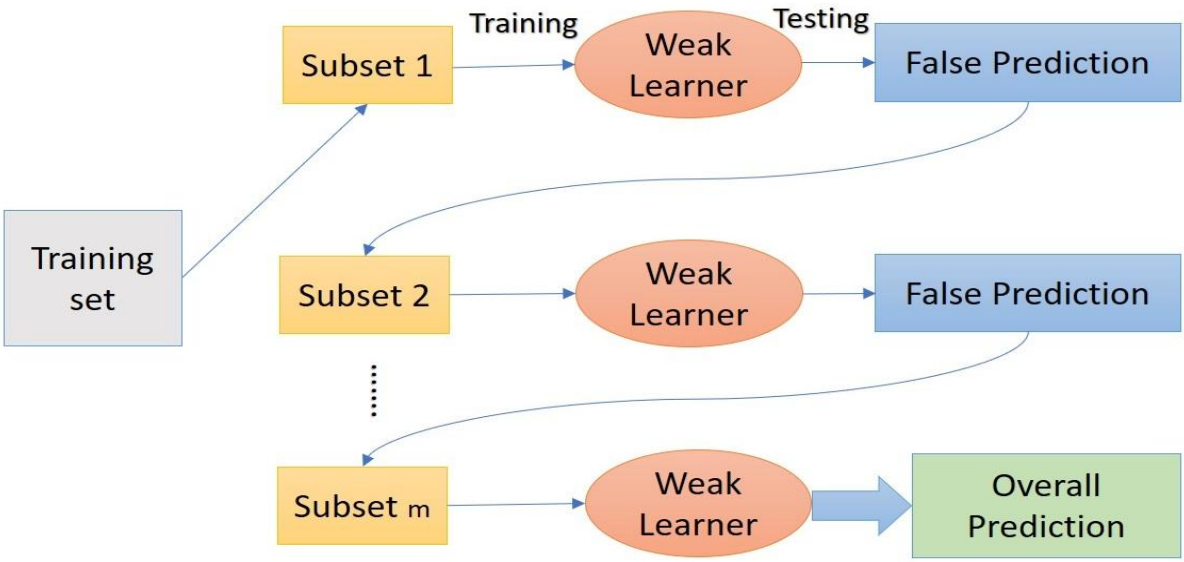

Note: This figure illustrates the bagging process, where the original training dataset is resampled through bootstrap sampling to generate multiple subsets (B1, B2, …, Bm). Each subset is used to train an independent base model (M1, M2, …, Mm) in parallel. The predictions from these models are then aggregated to produce the final output, thereby reducing variance and improving model stability.

As shown in Figure 2, the bagging workflow generates multiple bootstrap samples (B1, B2, …, Bm), each used to train independent base learners (M1, M2, …, Mm) in parallel. This process enhances model diversity by exposing each learner to a different subset of the training data, which in turn improves predictive stability and reduces variance. Within the HOEL framework, this diversity provides a robust foundation for subsequent optimization levels—soft voting and stacking—which further refine and consolidate the final predictions.

### 4.2.3 Boosting ensemble modeling

Boosting is an iterative ensemble approach that converts multiple weak learners into a single strong predictor. The process emphasizes misclassified instances from previous iterations by adjusting their weights, allowing subsequent models to focus on reducing residual errors. This sequential learning continues until optimal accuracy is achieved (Coşkun & Turanli, 2023). Boosting is particularly beneficial in the proposed framework because it improves prediction accuracy by iteratively refining weak classifiers and reducing systematic bias. Figure 3 depicts the sequential training workflow in boosting, where each successive model builds upon the errors of its predecessors to improve overall performance.

**Figure 3.** Boosting Ensemble Modeling Workflow




Note: This figure illustrates the boosting process, where base learners are trained sequentially (M1, M2, …, Mm), with each subsequent model focusing on instances that were misclassified by previous models. The training process iteratively adjusts the importance of observations, allowing later learners to reduce residual errors and progressively improve overall predictive performance.

As shown in Figure 3, the boosting workflow follows a sequential learning process in which each model is trained to address the errors of its predecessor. By emphasizing misclassified instances at each iteration, the approach progressively improves classification accuracy and reduces systematic bias. Within the HOEL framework, this sequential refinement enhances the learning capability of base classifiers before they are integrated into higher-level ensemble optimization.

#### 4.2.4 Testing the baseline model accuracy

There are various metrics used in machine learning to assess the performance of predictive models. One of the most common is accuracy, which measures the proportion of correctly classified instances relative to the total number of instances (Coşkun & Turanli, 2023), defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

where *TP* = true positives, *TN* = true negatives, *FP* = false positives, and *FN* = false negatives.

Another widely used evaluation metric is the receiver operating characteristic–area under the curve (ROC–AUC), which evaluates classifier performance on unseen data. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) across various thresholds. The AUC measures the overall separability between classes, with a value of 1 indicating perfect discrimination (Naidu et al., 2023), defined as follows:

$$TPR = \frac{TP}{TP + FN}, \tag{2}$$

where *TPR* = true positive rate, *TP* = true positives, and *FN* = false negatives.

$$FPR = \frac{FP}{FP + TN}, \tag{3}$$

where *FPR* = false positive rate, *FP* = false positives, and *TN* = true negatives.

### *4.3 Model optimization*

The HOEL framework optimizes the performance of the model through applying three levels of optimization as follows:

#### 4.3.1 Level 1 of Optimization

Level 1 of optimization applies several enhancement methods to improve model performance and generalizability. Cross-validation is employed to evaluate the robustness of the trained model by partitioning the dataset into K folds, training the model K times, and aggregating results (Laborda & Ryoo, 2021), defined as follows:

$$CV_{avg} = \left(\frac{1}{K}\right) \times \sum_{i=1}^{K} M_i, \tag{4}$$

where $CV_{avg}$ is the average performance metric (e.g., accuracy or ROC–AUC), $K$ is the number of folds, $M_i$ is the performance obtained on the $i^{th}$ validation fold, and $i$ denotes the fold index ranging from 1 to $K$.

In addition, hyperparameter tuning is performed to identify the optimal configuration for each model and dataset. This process involves defining a search space, selecting an appropriate optimization methodology, and evaluating performance metrics to avoid suboptimal predictions. To balance optimization effectiveness with computational efficiency, RandomizedSearchCV is utilized, which samples hyperparameter combinations from predefined distributions. Unlike exhaustive GridSearch, RandomizedSearchCV randomly explores a specified number of parameter settings (n_iter), offering a more efficient yet effective optimization strategy (Yang & Shami, 2020).

Furthermore, imbalanced data are handled using the Synthetic Minority Oversampling Technique (SMOTE), as class imbalance can significantly affect model predictions. SMOTE generates synthetic samples for the minority class by interpolating between existing minority instances and their nearest neighbors (Zhao et al., 2024), defined as follows:

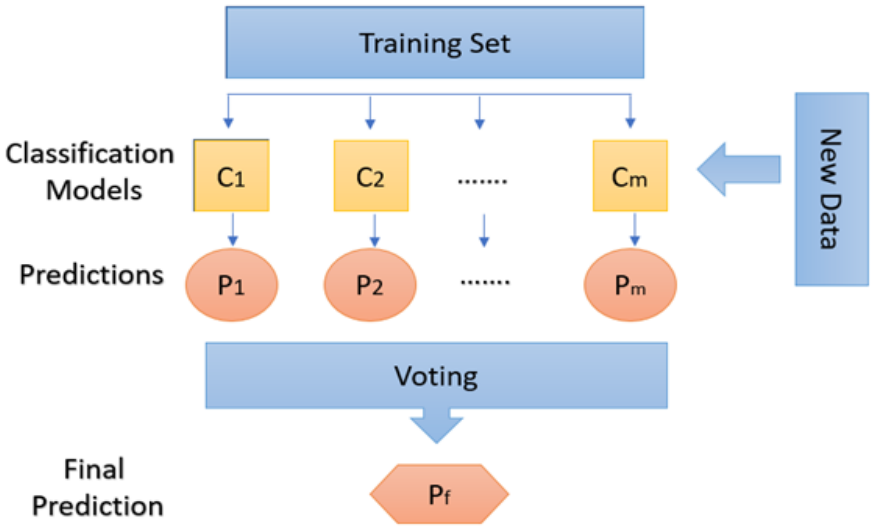
$$x_{new} = x + random(0,1) \times |x - x_n|, \tag{5}$$

where $x_{new}$ is the synthetic minority instance generated by SMOTE, $x$ is an arbitrary minority-class sample, $x_n$ denotes one of the nearest neighbors of $x$, $n$ is the sampling ratio, and random(0, 1) is a uniform random variable between 0 and 1 that introduces variability into the synthetic sample.

### 4.3.2 Level 2 of Optimization

This level of optimization applies a voting ensemble technique to enhance the robustness of classification. This research uses a soft voting approach, which calculates a weighted average of the predicted probabilities from the constituent models, thereby improving the performance of individual learners and addressing limitations associated with base classifiers in ensemble voting, as illustrated in Figure 4 (Kokate & Chetty, 2021).

**Figure 4.** Voting Ensemble Classification Architecture




Note: This figure illustrates the soft voting classification approach, where multiple base learners (C1, C2, …, Cm) generate predicted class probabilities (P1, P2, …, Pm). These predictions are combined through a weighted aggregation mechanism to determine the final predicted class (Pf), as formalized in Equation (6).

As shown in Figure 4, the soft voting mechanism combines the probabilistic outputs of multiple base classifiers to produce a more stable and accurate prediction. By aggregating predictions (P1, P2, …, Pm) from different models (C1, C2, …, Cm), the approach reduces individual model bias and variance, resulting in improved classification robustness within the HOEL framework.

To formalize the soft voting process, the final predicted class is expressed as follows:

$$\hat{y} = \arg\max_{c \in C} \sum_{j=1}^{M} w_j P_j(c \mid x), \tag{6}$$

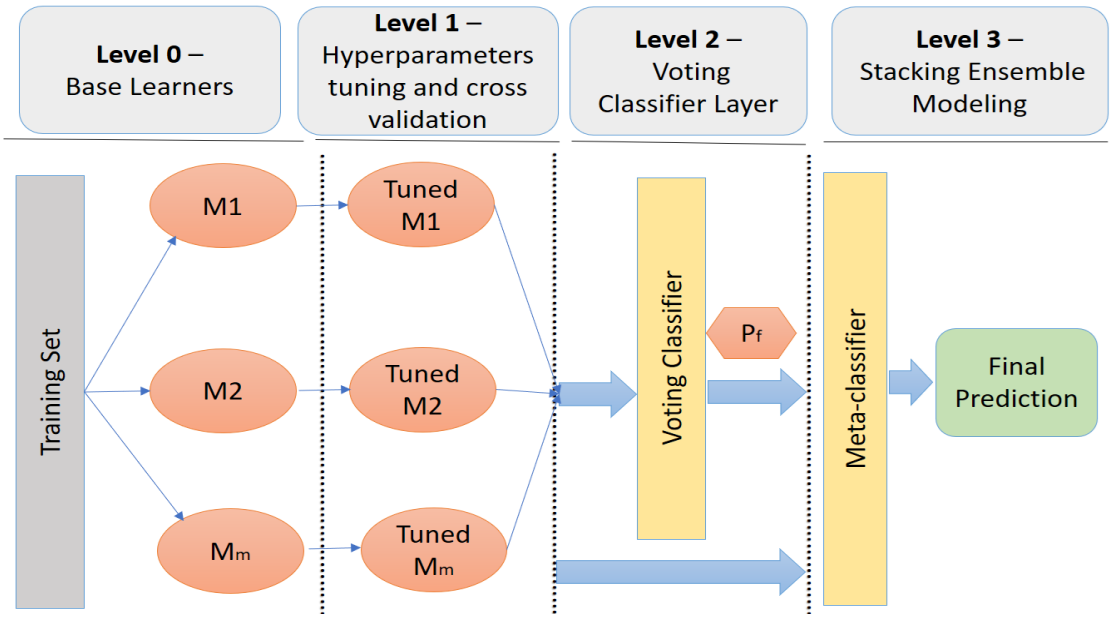
where $\hat{y}$ is the final predicted class label; $C$ denotes the set of possible class labels; $M$ denotes the total number of classifiers; $w_j$ is the weight assigned to classifier $j$; $P_j(c \mid x)$ represents the predicted probability of class $c$ given input $x$ from classifier $j$; and arg max selects the class with the highest aggregated weighted probability.

#### 4.3.3 Level 3 of Optimization

This level of optimization applies stacking ensemble modeling, which operates through a layered architecture where multiple models, known as base learners, are trained on the original training data in the first layer. Each of these models makes predictions independently, and their outputs are collected as new input features. These features are then fed into a second-layer model, often called the meta-learner or meta-classifier, which learns how to best combine the base models' predictions to make a final decision (Munsarif et al., 2022).

The HOEL framework builds a tuned layer at Level 1 of optimization by applying K-fold cross-validation and hyperparameter tuning. Level 2 of optimization employs the voting classifier to construct an automated and interpretable modeling layer. In Level 3 of optimization, the stacking model combines the predictions from Level 1 and Level 2 to build the final predictions using the meta-classifier, as shown in Figure 5.

**Figure 5.** Multi-Level Optimization Structure of the HOEL Framework




Note: This figure illustrates the multi-level optimization structure of the HOEL framework, including Level 1 (tuned base learners M1, M2, …, Mm), Level 2 (voting classifier producing intermediate predictions Pf), and Level 3 (stacking meta-classifier). The meta-classifier integrates predictions from both the tuned base learners and the voting layer to generate the final prediction, as formalized in Equation (7).

As shown in Figure 5, the HOEL framework employs a hierarchical optimization structure in which predictions from tuned base learners (Level 1) and the voting classifier (Level 2) are integrated through a stacking meta-classifier (Level 3). This layered architecture enables the model to capture both individual learner behavior and ensemble consensus, resulting in improved predictive accuracy and robustness. The integration of multiple learning stages reduces both bias and variance, enhancing the overall performance of the credit scoring model.

To formalize the stacking ensemble process, the final predicted class is expressed as follows:

$$\hat{y} = f_{\text{meta}}\left(h_1(x), h_2(x), \ldots, h_K(x), \hat{y}_{\text{vote}}(x)\right), \tag{7}$$

where $\hat{y}$ denotes the final predicted class label; $h_k(x)$ represents the prediction of the $k$-th base learner for input $x$; $\hat{y}_{\text{vote}}(x)$ denotes the output of the voting ensemble from Level 2; $K$ is the total number of base learners; and $f_{\text{meta}}$ represents the meta-classifier that learns the optimal combination of the base learners' predictions.

Notably, unlike conventional stacking approaches, the HOEL framework incorporates the soft voting output $\hat{y}_{\text{vote}}(x)$ from Level 2 as an additional input to the meta-learner, enabling the meta-classifier to leverage both ensemble consensus and individual base learner predictions.

## 5. Data and Variables

The scope of this section is to discuss the data and variables of the P2P lending risk model in the digital lending platforms using the HOEL framework. The dataset used in this model contains 32,581 P2P customers' credit scoring data from an open-source database on the Kaggle website, and the target label detects whether the borrower consumers are defaulters or non-defaulters (Kaggle, 2024). The dataset does not include timestamp information (e.g., loan origination dates or observation periods), and therefore, the temporal coverage of the loans is unspecified. The dataset is structured from 12 fields (input features). The description of these fields is shown in Table 1. The data were loaded using **Pandas** in a Jupyter Notebook environment, where a structured dimensional data mart was created to organize the variables in a tabular format.

**Table 1.** P2P Lending Features Description

| Feature name | Description |
|---|---|
| person_age | Age of the person in years |
| person_income | Annual income |
| person_home_ownership | Type of home ownership (e.g., rent, own, etc.) |
| person_emp_length | Length of current employment in years |
| loan_intent | Loan intent (e.g., personal, educational, etc.) |
| loan_grade | Loan grade (A, B, C, D, E, F, G) |
| loan_amnt: | Loan amount |
| loan_int_rate | Loan interest rate |
| loan_status | Loan status (0 for non-default, 1 for default) |
| loan_percent_income | Percent of income committed to the loan |
| cb_person_default_on_file | Historical default (Yes or No) |
| cb_preson_cred_hist_length | Credit history length in years |

Note: This table summarizes the input features used in the peer-to-peer (P2P) credit-risk dataset obtained from the Kaggle open-source platform. These variables constitute the predictor vector $x$ in the HOEL modeling framework and capture key borrower characteristics, loan attributes, and credit-history indicators.

As shown in Table 1, the dataset includes a diverse set of borrower-related, financial, and credit-history variables that provide a comprehensive basis for modeling credit risk in P2P lending. The combination of demographic attributes (e.g., age and income), loan characteristics (e.g., loan amount and interest rate), and behavioral indicators (e.g., default history) enables the HOEL framework to capture both linear and nonlinear relationships influencing borrower default.

Exploratory Data Analysis (EDA) is a critical pre-processing step in machine learning models that enables a deeper understanding of the dataset distributions, defining patterns, and anomaly detection through statistical analysis and data visualization capabilities, as shown in Table 2.

**Table 2.** P2P Lending Features Statistics

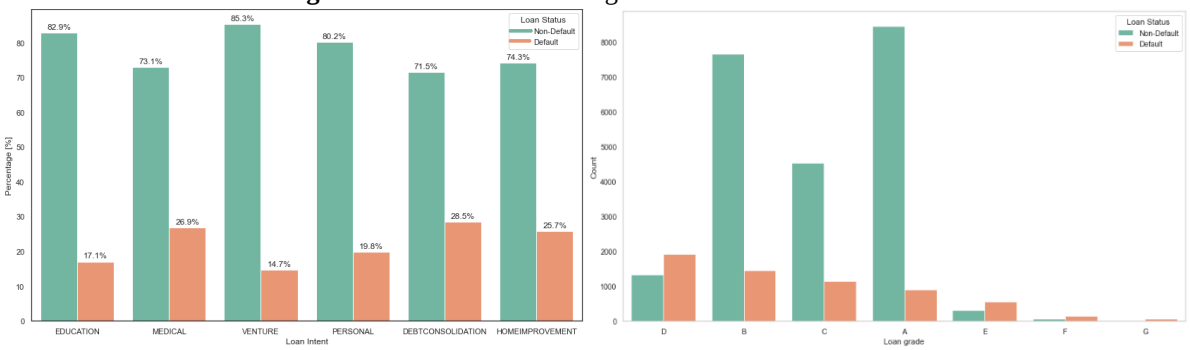
| Feature | Count | Mean | Std | Min | 25% | 50% | 75% | Max |
|---|---|---|---|---|---|---|---|---|
| person_age | 32,581 | 27.7346 | 6.3481 | 20 | 23 | 26 | 30 | 144 |
| person_income | 32,581 | 66,100 | 62,000 | 4,000 | 38,500 | 55,000 | 79,000 | 6,000,000 |
| person_emp_length | 31,686 | 4.7897 | 4.1426 | 0 | 2 | 4 | 7 | 123 |
| loan_amnt | 32,581 | 9589.37 | 6322.08 | 500 | 5000 | 8000 | 12200 | 35000 |
| loan_int_rate | 29,465 | 11.0116 | 3.2405 | 5.42 | 7.9 | 10.99 | 13.47 | 23.22 |
| loan_status | 32,581 | 0.2182 | 0.4130 | 0 | 0 | 0 | 0 | 1 |
| loan_percent_income | 32,581 | 0.1702 | 0.1068 | 0 | 0.09 | 0.15 | 0.23 | 0.83 |
| cb_person_cred_hist_length | 32,581 | 5.8042 | 4.0550 | 2 | 3 | 4 | 8 | 30 |

Note: This table presents descriptive statistics for all continuous variables in the peer-to-peer (P2P) lending dataset, based on the full sample (32,581 observations) prior to model training. Count refers to the number of non-missing observations; Mean and Std represent the average and standard deviation, respectively; and Min, 25%, 50%, 75%, and Max indicate distributional percentiles. These variables correspond to elements of the predictor vector $x$ used in the HOEL framework and provide insight into the scale, variability, and distributional properties of the inputs for subsequent modeling.

As shown in Table 2, the dataset exhibits substantial variability across key financial and demographic variables, with notable dispersion in income and loan-related features. These distributional characteristics indicate the presence of potential outliers and highlight the need for preprocessing. Accordingly, data cleaning procedures were applied, including duplicate removal, which reduced the dataset from 32,581 to 32,416 observations. Missing values, particularly in the *person_emp_length* and *loan_int_rate* variables, were handled using the *dropna* function, resulting in a final dataset of 25,651 observations. Furthermore, analysis of the target variable (*loan_status*) reveals a class imbalance, with 20.7% defaulters and 79.3% non-defaulters. To address this imbalance and improve model performance, SMOTE oversampling was applied to balance the class distribution.

Figure 6 presents the distribution of categorical and ordinal features across defaulters and non-defaulters.

**Figure 6.** Distribution of Categorical and Ordinal Features



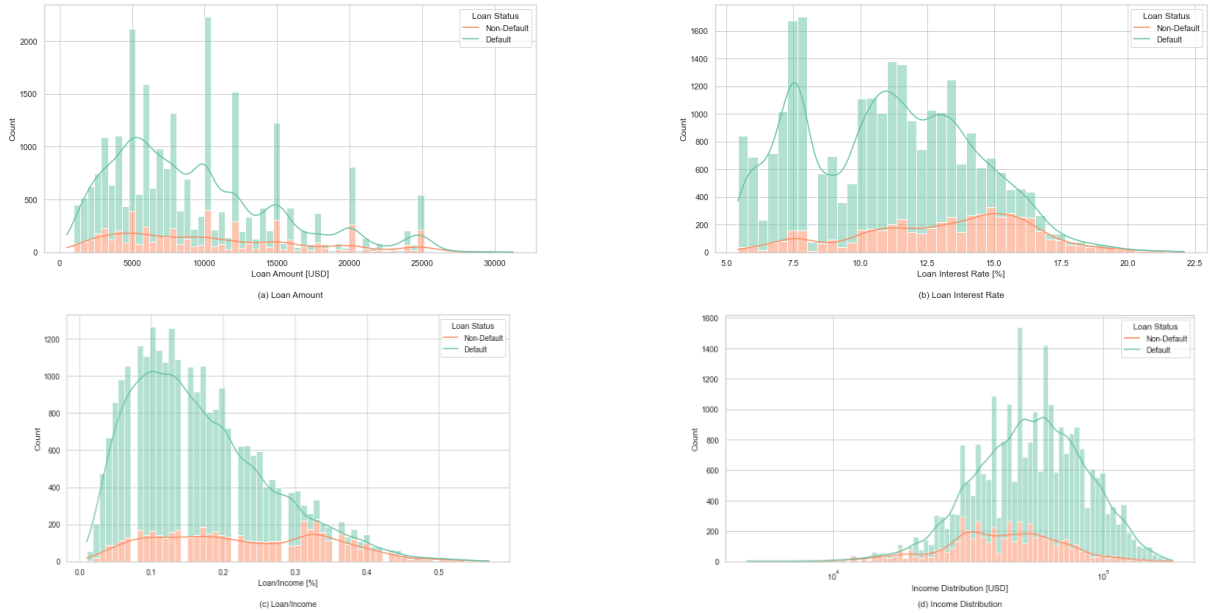

(a)Loan Intent (b) Loan Grade

Note: This figure illustrates the distribution of categorical and ordinal features, including loan intent (panel a) and loan grade (panel b), across defaulters and non-defaulters. It highlights variations in borrower behavior, where certain loan purposes (e.g., debt consolidation and medical expenses) and lower credit grades are more prevalent among default cases. These features constitute key categorical inputs to the HOEL framework and capture important risk segmentation patterns in the dataset.

As shown in Figure 6, analysis of defaulters indicates that borrowers who default are more likely to use loans for debt consolidation, medical expenses, or home improvement purposes. In addition, they tend to have lower loan grades and a prior history of default, suggesting stronger risk concentration within these categories.

Figure 7 presents the distribution of numerical features across defaulters and non-defaulters.

**Figure 7.** Distribution of Numerical Features




Note: This figure illustrates the distribution of numerical features, including loan amount (panel a), loan interest rate (panel b), loan-to-income ratio (panel c), and income (panel d), across defaulters and non-defaulters. The plots indicate that default cases are more prevalent among borrowers with lower income levels, higher loan-to-income ratios, and higher interest rates. These variables represent key continuous inputs to the HOEL framework and provide valuable insights into borrower risk profiles.

As shown in Figure 7, defaulters are more prevalent among borrowers with lower income levels, higher loan-to-income ratios, and higher interest rates. In addition, the distribution indicates a higher proportion of defaulters among younger age groups. These patterns reflect increased financial vulnerability and provide important predictive signals for credit risk modeling within the HOEL framework.

## 6. Experiments, Results, and Discussion

This section outlines the implementation of the HOEL framework through multiple experimental phases. The analysis follows a stepwise approach to ensure rigorous preprocessing, robust modeling, and comprehensive evaluation. Section 6.1 describes the preprocessing and modeling pipeline, Section 6.2 reports diagnostic checks, and Section 6.3 summarizes the results and discusses their implications.

### *6.1 Implementation of the HOEL Framework for the P2P Digital Lending*

This section presents the implementation of the proposed HOEL framework for credit risk assessment in peer-to-peer (P2P) digital lending. It begins with an exploratory analysis to examine the overall structure

of the dataset and provide a general understanding of the relationships among variables. This step supports the interpretation of the data and establishes a foundation for subsequent modeling stages.

To this end, Figure 8 presents the Pearson correlation heatmap of the study variables.

**Figure 8.** Heatmap of Pearson Correlation Coefficients Among P2P Lending Variables


Note: This figure presents the Pearson correlation matrix of the study variables, including borrower characteristics, loan attributes, and credit-history indicators used in the HOEL framework. The values represent pairwise linear correlation coefficients ranging from −1 to 1, where higher absolute values indicate stronger relationships between variables. Statistical significance levels correspond to those reported in Table A1 ($p < 0.10$ (*), $p < 0.05$ (**), $p < 0.01$ (***)).

As shown in Figure 8, the heatmap highlights the strength and direction of relationships among the study variables. As expected, each variable is perfectly correlated with itself (r = 1.0). Among the cross-variable patterns, *person_age* and *cb_person_cred_hist_length* exhibit a particularly strong positive correlation (r = 0.85), reflecting the intuitive relationship between borrower age and accumulated credit history. Similarly, *loan_amnt* shows a moderately strong relationship with *loan_percent_income* (r = 0.66), indicating that larger loan amounts tend to represent a larger proportion of borrower income—an important factor in credit-risk assessment.

To supplement the visualization, a detailed correlation table with statistical significance indicators is presented in the Appendix (Table A1). Given the large sample size, most relationships are statistically significant ($p < .01$); however, the analysis emphasizes practical strength rather than statistical significance alone. It is important to note that, even when variables exhibit stationarity, correlation does not imply causation. Therefore, these relationships should be interpreted cautiously to avoid misleading conclusions or spurious inference.

### 6.1.1 Preprocessing and Data Preparation

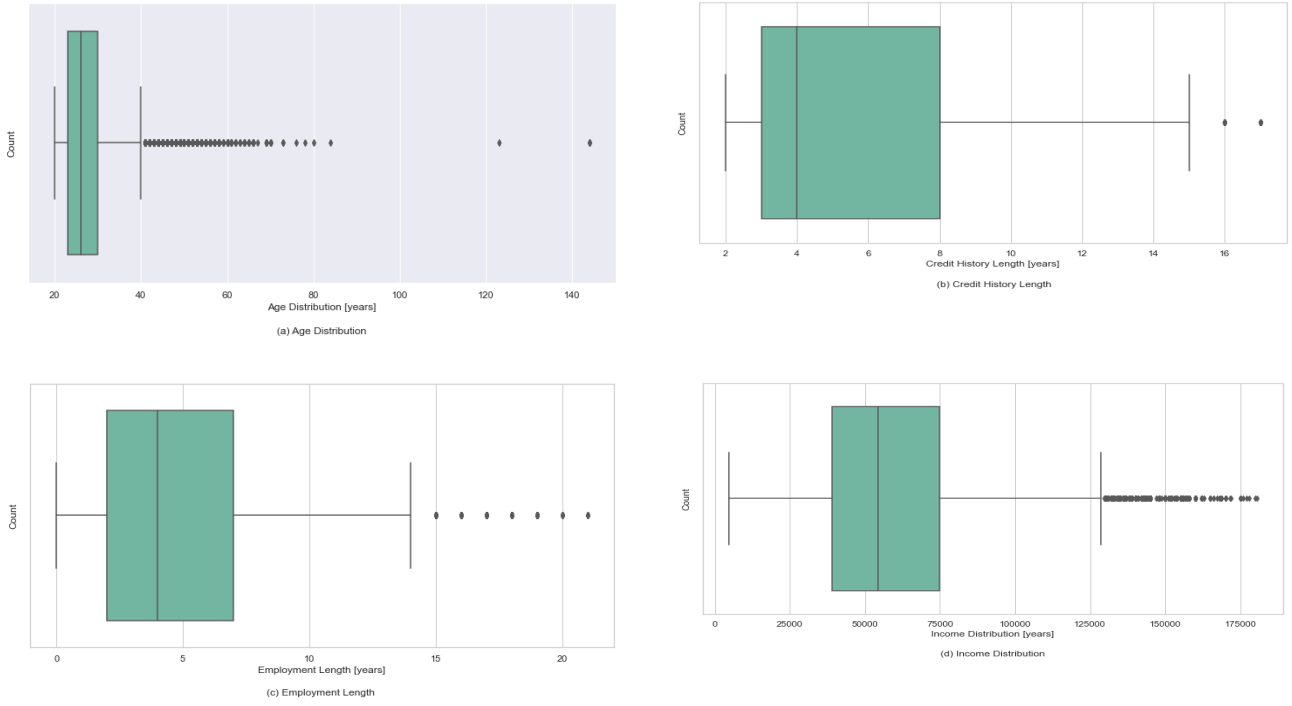
In the preprocessing stage, the dataset is prepared to ensure quality, consistency, and suitability for modeling. This involves outlier detection, feature engineering, and handling class imbalance to enhance model performance and reliability.

As a first step, outliers are identified and addressed. Outliers are extreme values that fall outside the typical range of the data and can distort model estimation and negatively affect predictive performance. To detect such observations, the Elliptic Envelope algorithm is employed, which is an unsupervised method based on covariance estimation under the assumption of a Gaussian distribution. Compared to traditional approaches such as the interquartile range (IQR), this method provides a more robust mechanism for identifying multivariate outliers.

The effectiveness of the preprocessing step is evaluated using the mean squared error (MSE), which measures the average squared difference between predicted and observed values. The computed MSE is 0.126186, and a total of 2,850 outliers are detected. These observations are subsequently removed to prevent distortion of model estimates and to ensure stable and reliable training performance.

Figure 9 presents the boxplot distribution of selected numerical features used in the HOEL framework.

**Figure 9.** Box Plot Distribution of Numerical Features




Note: Note: This figure presents the boxplot distribution of key numerical features, including age (panel a), credit history length (panel b), employment length (panel c), and income (panel d). The boxplots display the median, interquartile range, and potential outliers, providing insight into the distribution, spread, and skewness of the variables. The presence of extreme values indicates variability in borrower characteristics and motivates the application of outlier treatment prior to model training within the HOEL framework.

As shown in Figure 9, several numerical features exhibit noticeable outliers and skewed distributions, particularly in age, employment length, credit history length, and income. These extreme values can distort model estimation and negatively affect predictive performance if not properly addressed. Therefore,

outlier detection and removal are essential preprocessing steps to improve model stability and ensure reliable training outcomes.

As the second step, feature engineering is applied. The purpose of applying the feature engineering step is to transform raw data into a more effective set of inputs for improving the training of machine learning models. For numerical variables, a Simple Imputer is used to manage missing values, and a Robust Scaler is applied to scale features while reducing the impact of outliers. For categorical variables, One-Hot Encoding is used to convert categories into interpretable numerical representations. Lastly, for the ordinal variable, a Min-Max Scaler is employed to normalize values within a specific range. These preprocessing techniques enhance model accuracy and interpretability.

Finally, the data is handled for the existing imbalances. Imbalanced data in the peer-to-peer lending dataset leads to a significant impact on the model performance due to the skewness of the data distribution. The results indicate that manipulating the imbalanced data presents challenges. This research employs the SMOTE technique to address the imbalanced data problem. After applying the SMOTE technique, the sample distribution achieves equal balance at 17,831 for both the default and non-default datasets.

### 6.1.2 The Proposed HOEL Framework Implementation, Testing, and Evaluation

The HOEL framework presented in Figure 1 is evaluated in terms of classification performance and the effectiveness of the proposed optimization stages.

Prior to model training, feature importance is examined to identify the most influential variables affecting credit-risk prediction. Feature importance reflects the contribution of each variable to the model's predictive capability and supports the refinement of the modeling process by highlighting the most relevant predictors. The analysis indicates that *person_income* is the most influential feature, followed by *loan_int_rate*, *loan_intent*, and *loan_percent_income*. These variables capture key aspects of borrower financial stability and loan characteristics, providing valuable insights for credit-risk assessment.

The initial modeling stage employs a set of baseline classifiers, including Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest, CatBoost, LightGBM (LGBM), and XGBoost. Their performance is evaluated using Accuracy and ROC_AUC on a 20% held-out test set.

Table 3 presents the performance results of these baseline models prior to applying Level 1 optimization.

**Table 3.** P2P Lending Model Performance Before Applying Level 1 Optimization

| Model | Accuracy | ROC_AUC |
|---|---|---|
| Logistic Regression | 89.04% | 75.81% |
| Random Forest | 89.34% | 92.68% |
| SVM | 90.00% | 86.77% |
| KNN | 88.03% | 80.46% |
| XGBoost | 92.12% | 94.51% |
| CatBoost | 92.72% | 94.32% |
| LGBM | 91.90% | 94.15% |

Note: This table presents the performance of the seven base machine-learning classifiers before applying any optimization steps. The results are computed on the 20% held-out test set from the Kaggle credit-risk dataset. Accuracy and ROC_AUC are defined according to Equations 1 and 2, respectively, and are used to evaluate classification performance and discriminative ability.

As shown in Table 3, the baseline classifiers exhibit varying levels of performance in terms of both accuracy and discriminative ability. XGBoost achieves the highest ROC_AUC score (94.51%), indicating superior capability in distinguishing between defaulters and non-defaulters. CatBoost follows closely with a ROC_AUC of 94.32% and achieves the highest classification accuracy (92.72%), while LGBM also demonstrates strong performance with a ROC_AUC of 94.15%.

In contrast, traditional models such as Logistic Regression and KNN show comparatively lower discriminative performance, highlighting their limitations in capturing complex nonlinear relationships within the dataset. Overall, the results indicate that gradient boosting models (XGBoost, CatBoost, and LGBM) outperform other classifiers, making them strong candidates for subsequent optimization and ensemble integration within the HOEL framework.

Following the baseline evaluation, Level 1 optimization is applied to enhance model performance. This stage involves 5-fold cross-validation to improve model generalizability, followed by hyperparameter tuning using RandomizedSearchCV to identify optimal configurations. These procedures aim to reduce overfitting and improve predictive accuracy across classifiers.

Table 4 presents the performance of the models after applying Level 1 optimization.

**Table 4.** P2P Lending Models Performance After Applying Level 1 Optimization

| Model | Accuracy | ROC_AUC |
|---|---|---|
| Logistic Regression | 90.54% | 75.96% |
| Random Forest | 90.39% | 93.34% |
| SVM | 91.15% | 86.90% |
| KNN | 88.80% | 81.78% |
| XGBoost | 93.34% | 94.99% |
| CatBoost | 94.88% | 94.72% |
| LGBM | 91.97% | 94.58% |

Note: This table reports the performance of all classifiers after applying Level 1 optimization, which includes 5-fold cross-validation and hyperparameter tuning via RandomizedSearchCV. Results are based on the same 20% held-out test set used throughout the study. Accuracy and ROC_AUC are defined according to Equations (1) and (2), respectively, and are used to evaluate classification performance and discriminative ability.

As shown in Table 4, all classifiers exhibit improved performance following Level 1 optimization. XGBoost achieves the highest ROC_AUC score (94.99%), followed by CatBoost (94.72%) and LGBM (94.58%), indicating enhanced discriminative capability after hyperparameter tuning. CatBoost also attains the highest accuracy (94.88%), demonstrating improved classification performance.

Compared to the baseline results in Table 3, these improvements confirm the effectiveness of cross-validation and hyperparameter tuning in enhancing model generalizability and predictive accuracy. The optimized gradient boosting models remain the strongest performers and are therefore selected for subsequent ensemble stages within the HOEL framework.

Following Level 1 Optimization, additional ensemble techniques are applied to further enhance predictive performance. At Level 2, a voting ensemble model is constructed by combining the best-performing classifiers (XGBoost, CatBoost, and LGBM). At Level 3, a stacking meta-learner is implemented to integrate the outputs of Level 1 models and the Level 2 voting ensemble.

Table 5 summarizes the performance results across all optimization levels of the HOEL framework.

**Table 5. Overall** Optimization Levels and Performance Results of the HOEL Framework

| **Before optimization** | | |
|---|---|---|
| **Model** | **Accuracy** | **ROC_AUC** |
| Logistic Regression | 89.04% | 75.81% |
| Random Forest | 89.34% | 92.68% |
| SVM | 90.00% | 86.77% |
| KNN | 88.03% | 80.46% |
| XGBoost | 92.12% | 94.51% |
| CatBoost | 92.72% | 94.32% |
| LGBM | 91.90% | 94.15% |
| **Level 1 optimization** | | |
| Logistic Regression | 90.54% | 75.96% |
| Random Forest | 90.39% | 93.34% |
| SVM | 91.15% | 86.90% |
| KNN | 88.80% | 81.78% |
| XGBoost | 93.34% | 94.99% |
| CatBoost | 94.88% | 94.72% |
| LGBM | 91.97% | 94.58% |
| **Level 2 Optimization** | | |
| Voting Classifier | 95.35% | 96.21% |
| **Level 3 Optimization** | | |
| Stacking without voting | 95.77% | 97.20% |
| Stacking with voting | 96.15% | 97.62% |

Note: This table summarizes the performance of the HOEL framework across all optimization levels, including baseline models, Level 1 optimized models, Level 2 voting ensemble, and Level 3 stacking meta-learner (with and without voting). Results are evaluated on the 20% held-out test set. Accuracy and ROC_AUC are defined according to Equations (1) and (2), respectively, and serve as measures of classification performance and discriminative ability.

As shown in Table 5, the HOEL framework achieves consistent and substantial improvements across all optimization levels. The baseline models demonstrate moderate results, which are significantly enhanced after Level 1 Optimization through cross-validation and hyperparameter tuning.

At Level 2, the voting ensemble further enhances predictive performance, achieving an ROC_AUC of 96.21% and an accuracy of 95.35%, reflecting the benefit of combining multiple strong classifiers.

At Level 3, the stacking meta-learner achieves the highest performance. Notably, the configuration that incorporates both stacking and the voting layer produces the best results, with an ROC_AUC of 97.62% and an accuracy of 96.15%. This highlights the effectiveness of the proposed multi-level optimization strategy in improving predictive accuracy and robustness.

Overall, these findings confirm that integrating voting and stacking within a unified framework provides significant performance gains over individual models and conventional ensemble approaches.

### *6.2 Diagnostic Checks*

A set of diagnostic tests is conducted to validate the robustness of HOEL framework and ensure that the empirical findings are not spurious. These checks provide complementary evidence to predictive performance and serve to confirm that the chosen models are theoretically well-justified. The diagnostics focused on five areas: 1) Multicollinearity, 2) Residual normality, 3) Nonlinearity, 4) Stationarity diagnostics, and 5) Autocorrelation in Residual diagnostics.

#### 6.2.1 Multicollinearity

Multicollinearity is examined using Variance Inflation Factors (VIF), which measure the extent to which a predictor is linearly related to other predictors in the model (O'Brien, 2007).

Table 6 presents the VIF results for the predictor variables.

**Table 6.** Variance Inflation Factor (VIF) Results for Predictor Variables

| Variable | VIF | Interpretation |
|---|---|---|
| person_age | 3.91 | Moderate, acceptable |
| person_income | 1.50 | No issue |
| person_emp_length | 1.06 | No issue |
| loan_amnt | 2.10 | Acceptable |
| loan_int_rate | 1.03 | No issue |
| loan_percent_income | 2.05 | Acceptable |
| cb_person_cred_hist_length | 3.84 | Moderate, acceptable |

Note: This table reports the Variance Inflation Factor (VIF) values for the continuous predictor variables used in the modeling dataset. VIF measures the degree of multicollinearity among predictors, where values below 10 indicate that multicollinearity is not a significant concern. All variables fall within acceptable limits, supporting their suitability for inclusion in the HOEL framework.

As shown in Table 6, all predictor variables exhibit VIF values well below the conventional threshold of 10, indicating that multicollinearity is not a significant issue in the dataset. The highest values are observed for *person_age* (3.91) and *cb_person_cred_hist_length* (3.84), both of which remain within acceptable ranges.

These results confirm that the predictors are sufficiently independent, ensuring stable model estimation and reliable interpretation. Moreover, the low level of multicollinearity supports the robustness of the applied machine learning models, particularly ensemble methods such as Random Forest and gradient boosting, which are inherently resilient to correlated inputs (Hastie et al., 2009).

#### 6.2.2 Residual Normality (Jarque–Bera)

The Jarque–Bera (JB) test evaluates whether residuals follow a normal distribution based on skewness and kurtosis (Jarque & Bera, 1980). A non-significant result ($p \geq .05$) indicates normality, while significance suggests departures from Gaussian behavior.

Table 7 presents the JB test results for the residuals of all evaluated models.

**Table 7.** Jarque–Bera Normality Test Results for Model Residuals

| Model | JB Statistic | *p*-value | Skew | Kurtosis | Normal (5%)? |
|---|---|---|---|---|---|
| LR | 3,027.94 | < .001 | 1.06 | 5.11 | No |
| RF | 26,240.81 | < .001 | 2.44 | 10.32 | No |
| SVC | 10,653.99 | < .001 | 1.62 | 7.56 | No |
| KNN | 6,025.58 | < .001 | 1.51 | 5.93 | No |
| LGBM | 27,347.83 | < .001 | 2.44 | 10.54 | No |
| XGB | 27,317.74 | < .001 | 2.44 | 10.53 | No |
| CAT | 38,280.27 | < .001 | 2.81 | 12.01 | No |
| Combined (Voting + Stacking) | 33,106.43 | < .001 | 2.6 | 11.39 | No |
| Weighted Voting | 3,140.66 | < .001 | 1.14 | 5 | No |

Note: This table reports the Jarque–Bera (JB) test results for residuals from each model evaluated in the study. The JB test assesses departures from normality using skewness and kurtosis; a significant *p*-value ($p < .05$) indicates non-normal residual distributions. All models show significant deviation from normality, consistent with the heavy-tailed nature of financial risk data.

As reported in Table 7, all models strongly reject the null hypothesis of normality ($p < .001$). Skewness values range from 1.06 to 2.81, while kurtosis values range from 5.00 to 12.01, indicating right-skewed and heavy-tailed residual distributions. These findings are consistent with the well-documented asymmetric and heavy-tailed nature of financial risk data (Cont, 2001).

Importantly, these deviations from normality do not undermine the predictive validity of the machine learning models employed. Ensemble methods such as Random Forest and gradient boosting do not rely on Gaussian assumptions, making them well-suited for capturing complex, nonlinear patterns in financial data. The Jarque–Bera results, therefore, reflect inherent characteristics of the dataset rather than model inadequacy.

Overall, the observed residual behavior further supports the suitability of advanced nonlinear and ensemble approaches within the HOEL framework for credit risk modeling.

### 6.2.3 Nonlinearity and Model Specification (RESET & Link Test)

To assess whether a linear specification is adequate for modeling borrower risk, diagnostic tests are applied to baseline linear models. Specifically, the Ramsey RESET test (Ramsey, 1969) and the Pregibon logistic link test are used to evaluate potential model misspecification and the presence of nonlinear relationships.

Table 8 reports the results of these nonlinearity diagnostics.

**Table 8.** Nonlinearity Diagnostic Test Results

| Test | Statistic | *p*-value | Nonlinearity Detected? |
|---|---|---|---|
| Ramsey RESET (OLS baseline) | F = 482.73 | < .001 | Yes |
| Logistic Link Test (Pregibon) | z = –6.60, $\beta(\hat{p}^2)$ = –2.03 | < .001 | Yes |

Note: This table presents the results of two specification diagnostics applied to baseline linear models: The Ramsey RESET test and the Pregibon logistic link test. Both tests evaluate whether linear functional forms adequately capture the underlying data-generating process. Significant results ($p < .05$) indicate nonlinearity and model misspecification, supporting the use of nonlinear and ensemble learning methods for credit-risk prediction.

As reported in Table 8, both diagnostic tests provide strong evidence against the adequacy of linear model specifications. The Ramsey RESET test is highly significant ($F = 482.73$, $p < .001$), indicating the presence of omitted nonlinear relationships. Similarly, the Pregibon logistic link test—based on augmenting the model with squared fitted values—shows a significant coefficient for the squared term ($\beta = –2.03$, $z = –6.60$, $p < .001$), further confirming model misspecification.

These findings demonstrate that the underlying data-generating process exhibits substantial nonlinearity, which cannot be adequately captured by linear models. Consequently, the results support the adoption of nonlinear and ensemble approaches—such as Random Forest, gradient boosting, and hybrid voting–stacking models—which are better suited for capturing complex borrower behavior and the heavy-tailed nature of financial risk data.

In addition, more advanced nonlinearity testing approaches, such as the method proposed by Hui et al. (2017), provide complementary diagnostics with enhanced capability for detecting complex nonlinear dynamics beyond polynomial-based tests such as RESET.

### 6.2.4 Stationarity Diagnostics

To assess the stationarity properties of the predictors, unit root diagnostics are conducted using the Augmented Dickey–Fuller (ADF) test (Dickey & Fuller, 1979). This test evaluates whether the variables contain a unit root or are stationary in levels.
Table 9 reports the ADF test results for all continuous predictors.

**Table 9.** Augmented Dickey–Fuller (ADF) Test Results for Predictor Variables

| Variable | ADF Statistic | *p*-value | Stationary(5%)? |
|---|---|---|---|
| person_age | -3.502 | .0079 | Yes |
| person_income | -9.568 | < .001 | Yes |
| person_emp_length | -16.056 | < .001 | Yes |
| loan_amnt | -11.007 | < .001 | Yes |
| loan_int_rate | -10.292 | < .001 | Yes |
| loan_percent_income | -13.16 | < .001 | Yes |
| cb_person_cred_hist_length | -3.527 | .0073 | Yes |

Note: This table reports Augmented Dickey–Fuller (ADF) unit root test results for all continuous predictors. The null hypothesis ($H_0$) assumes the presence of a unit root (non-stationarity). A significant p-value ($p < .05$) indicates rejection of $H_0$ and confirms stationarity in levels (I(0)).

As reported in Table 9, all variables significantly reject the null hypothesis of a unit root at the 5% significance level, confirming that the predictors are stationary in levels (I(0)). This indicates that the data are stable and suitable for modeling without requiring differencing transformations.

While the confirmation of stationarity for all predictors is an important diagnostic outcome, it is essential to acknowledge that stationarity reduces—but does not eliminate—the possibility of spurious relationships. Cheng et al. (2021) demonstrate that spurious relationships can emerge even among nearly non-stationary series that technically pass conventional stationarity tests. Wong et al. (2024) show that regressions using purely stationary series can still produce spurious results under certain conditions, particularly when persistent autocorrelation or near-unit-root behavior exists.

Wong and Yue (2024) further demonstrate that regressions combining stationary and non-stationary variables may yield statistically significant yet economically meaningless results. Additionally, Wong and Pham (2022a, 2022b) reveal that even when variables are stationary, autoregressive noise in the error structure can render standard regression tests unreliable. These findings emphasize the importance of carefully considering data-generating processes and model structure. In this context, the use of advanced ensemble methods within the HOEL framework provides additional robustness by capturing complex data patterns beyond traditional linear assumptions.

### 6.2.5 Residual Autocorrelation (Durbin–Watson) Test

Residual autocorrelation is evaluated using the Durbin–Watson (DW) statistic, which tests whether adjacent residuals from a regression or predictive model are correlated (Durbin & Watson, 1950). The DW statistic is traditionally applied to time-series regression models to detect first-order serial correlation. Although the dataset used in this study is cross-sectional, the test is included as a general diagnostic of residual independence.

Table 10 reports the Durbin–Watson test results for all models.

**Table 10.** Durbin–Watson (DW) Test Results for Model Residuals

| Model | Durbin–Watson | Autocorrelation? |
|---|---|---|
| Combined (Voting + Stacking) | 1.987 | No |
| LGBM | 1.989 | No |
| XGB | 1.994 | No |
| CAT | 1.987 | No |
| Weighted Voting | 1.574 | No |
| SVC | 1.999 | No |
| LR | 1.997 | No |
| RF | 1.993 | No |
| KNN | 1.988 | No |

Note: This table reports Durbin–Watson (DW) test statistics for model residuals. Values near 2.0 indicate no autocorrelation; values below 1.5 suggest positive autocorrelation; values above 2.5 suggest negative autocorrelation. All models exhibit DW statistics within the acceptable range (≥ 1.5), indicating no significant serial correlation in residuals.

As reported in Table 10, the Durbin–Watson statistics range between 1.574 and 1.999, with the majority of values clustering close to 2. This indicates that residuals are not serially correlated across models. The Weighted Voting model yields a DW value of 1.574, which lies at the lower bound of the acceptable range, suggesting a slight tendency toward positive dependence; however, this value remains above the threshold for concern and is therefore considered acceptable.

Overall, these results indicate that model residuals are independent across observations, with no significant autocorrelation detected. This supports the reliability of the predictive models and suggests that the observed performance is not driven by spurious dependence structures.

Collectively, the diagnostic results indicate that the data exhibit non-normal, heavy-tailed distributions and that specification tests reject linear adequacy, confirming the presence of nonlinear relationships in the underlying data-generating process. While the stationarity of predictors reduces the risk of misleading inference, it does not fully eliminate it in the presence of complex dependence structures. These findings provide strong empirical justification for the use of nonlinear and ensemble learning approaches, such as the proposed HOEL framework, which are well-suited to capturing the complexity and non-Gaussian characteristics of financial risk data.

### *6.3 Results and Discussion*

This study designed a hybrid meta-learner ensemble framework called HOEL through the use of optimized voting-stacking ensemble models to improve the performance of forecasting in the P2P digital lending platforms. The initial phase of the framework involved utilizing seven machine learning classification models as base learners: Logistic Regression, SVM, KNN, Random Forest, CatBoost, LGBM, and XGBoost classifiers. Through the experiment, the CatBoost classifier recorded a high accuracy, but the XGBoost classifier had the highest performance in ROC_AUC, as shown in Table 5. The second phase of the framework implemented three improvement steps. The first improvement used K-fold cross-validation to fix overfitting by augmenting the training data, and it also applied RandomizedSearchCV to enhance the tuning of hyperparameters. Level 1 of optimization enhanced the performance of all classifiers. The highest classifiers in accuracy and ROC_AUC are CatBoost and XGBoost, respectively. The second improvement that applied an automated voting ensemble classifier resulted in improved prediction accuracy and ROC_AUC score. The final improvement employed the outputs of Level 1 and Level 2 as inputs to the stacking meta-learner technique, achieving a 1.61% increase in model accuracy.

The experimental results indicate that the HOEL framework using the hybrid Voting and Stacking ensemble classifiers yields superior performance compared to solely employing base learners. Utilizing three improvement stages resulted in improved prediction performance and reduced bias and variation in the final forecasts. The voting classifier's automation capability and the optimized meta-learner classifiers are utilized as inputs for the stacking meta-classifier. The novel hybrid HOEL framework achieves an ROC_AUC score of 97.62% and an accuracy score of 96.15%, surpassing the performance of Logistic Regression, SVM, KNN, Random Forest, CatBoost, LGBM, XGBoost classifiers, voting ensemble classifier, and stacking without considering the voting classifier.

To further contextualize the contribution of the proposed framework, a comparative analysis of ensemble learning strategies is conducted. This comparison highlights key methodological differences between traditional ensemble approaches and the proposed hybrid voting–stacking framework.

Table 11 summarizes the key characteristics of these ensemble methods.

**Table 11.** Comparison of Key Characteristics Between the HOEL Framework and Traditional Ensemble Methods

| **Feature** | **Bagging (e.g., Random Forest)** | **Boosting (e.g., XGboost)** | **Stacking** | **Hybrid (Voting + Stacking) [Novel]** |
|---|---|---|---|---|
| Bias Reduction | ✓ Moderate (uses multiple weak learners) | ✓✓ High (corrects previous errors iteratively) | ✓ High (meta-learner optimizes combination) | ✓✓ High (voting stabilizes, stacking refines) |
| Variance Reduction | ✓✓ High (averages predictions) | ✗ Low (boosting can overfit) | ✓ Moderate (depends on meta-learner) | ✓✓ Very High (voting smooths, stacking optimally learns) |
| Model Diversity | ✗ Limited (typically same model type) | ✗ Limited (sequential weak learners) | ✓ High (allows different models) | ✓✓ Very High (voting enables diverse models, stacking optimizes them) |
| Overfitting Risk | ✓ Low (averaging smooths predictions) | ✗ High (sensitive to noise, prone to overfitting) | ✓ Moderate (depends on meta-learner complexity) | ✓✓ Low (voting prevents weak models from influencing stacking) |
| Computational Cost | ✓ Moderate (parallelizable models) | ✗ High (sequential learning, multiple iterations) | ✗ High (two-stage learning) | ✗✗ Very High (both voting & stacking require additional computation) |
| Data Efficiency | ✓ Works well with small-medium data | ✗ Requires large data for best results | ✗ Requires large data (to train meta-learner effectively), | ✓ Works well even with small-medium data (voting pre-processes inputs for stacking) |
| Hyperparameter Sensitivity | ✓ Low (easy tuning) | ✗ High (sensitive to learning rate, iterations) | ✗ High (requires tuning for both base models & meta-learner) | ✓✓ Lower than stacking (voting stabilizes before stacking) |
| Model Interpretability | ✓ Moderate (feature importance available) | ✗ Low (boosting is hard to interpret) | ✗ Low (meta-learner is a black box) | ✓ Moderate (voting provides insights on model agreement before stacking finalizes decision) |
| Real-world Suitability | ✓ Good stable datasets | ✓ Good for complex problems but risky for small data | ✓ Good for structured problems with enough data | ✓✓ Best for diverse and complex problems |

| | |
|---|---|
| | (balances variance & bias effectively) |

Note: This table provides a qualitative comparison of key characteristics among traditional ensemble methods (bagging, boosting, and stacking) and the proposed HOEL hybrid approach. The comparison is conceptual in nature and is based on established properties reported in the literature. It highlights core modeling dimensions—including bias and variance reduction, model diversity, overfitting risk, computational cost, and interpretability—to illustrate how the HOEL framework integrates the strengths of voting and stacking to enhance predictive robustness.

As shown in Table 11, traditional ensemble methods typically focus on specific aspects of model improvement, such as variance reduction (bagging) or bias reduction (boosting). In contrast, the proposed HOEL framework integrates both voting and stacking mechanisms to simultaneously address bias and variance while enhancing model diversity and stability.

The inclusion of a voting layer prior to stacking represents a key innovation, as it stabilizes intermediate predictions before they are passed to the meta-learner. This structured, multi-stage approach improves robustness and interpretability compared to conventional stacking methods, which often rely directly on raw base model outputs. As a result, the HOEL framework provides a more balanced and effective solution for complex prediction tasks such as credit risk modeling.

To further evaluate the performance of the proposed framework, its predictive accuracy is compared with results reported for traditional ensemble methods in the literature.

Table 12 presents this comparative analysis.

**Table 12.** Comparative Accuracy Performance of Traditional Ensemble Methods and the HOEL Approach

| Ensemble Method | Accuracy (%) | Model (Reference) |
|---|---|---|
| Bagging (Random Forest) | 80% | (Madaan et al., 2021) |
| Boosting | 74% | (Coşkun & Turanli, 2023) |
| Voting | 81% | (Kokate & Chetty, 2021) |
| Stacking | 94.54 | (Munsarif et al., 2022) |
| Hybrid Ensemble | 96.15% | HOEL Approach – this research |

Note: This table compares the accuracy of traditional ensemble methods reported in the literature (Coşkun & Turanli, 2023; Kokate & Chetty, 2021; Madaan et al., 2021; Munsarif et al., 2022) with the accuracy achieved by the proposed HOEL framework. Accuracy values for Bagging (Random Forest), Boosting, Voting, and traditional Stacking are taken directly from the respective studies, each of which is based on different datasets and experimental settings. In contrast, the HOEL framework's accuracy (96.15%) is computed using the 20% test set of the Kaggle credit-risk dataset employed in this study. Therefore, the comparison provides contextual benchmarking across ensemble approaches, offering insight into the relative performance of the proposed HOEL framework.

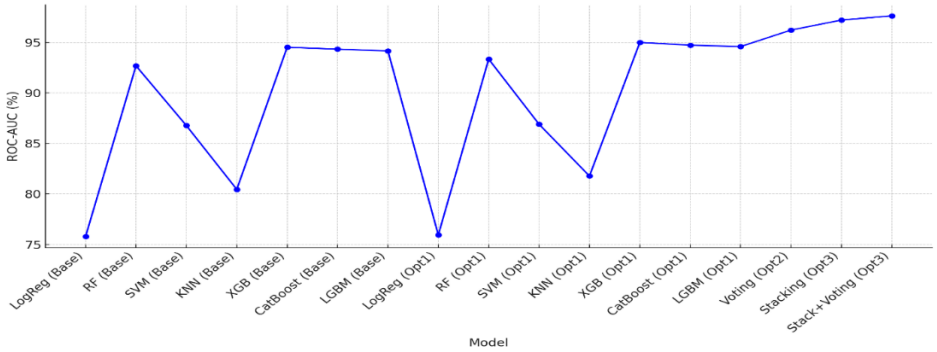
As reported in Table 12, the proposed HOEL framework achieves the highest predictive accuracy (96.15%) compared to traditional ensemble approaches reported in the literature. While direct comparisons should be interpreted with caution due to differences in datasets and experimental conditions, the results indicate that the hybrid voting–stacking approach offers substantial performance improvements over conventional methods.

The integration of voting and stacking enables the model to benefit from both prediction stability and adaptive learning, resulting in improved generalization and robustness. This multi-layered structure

reduces overfitting, enhances interpretability, and performs effectively in complex and heterogeneous data environments, making it particularly suitable for real-world applications such as credit risk assessment.

To further illustrate model performance across different optimization levels, the ROC–AUC results for all evaluated models are visualized in Figure 10.

**Figure 10.** ROC-AUC Comparison Across Models and Optimization Levels

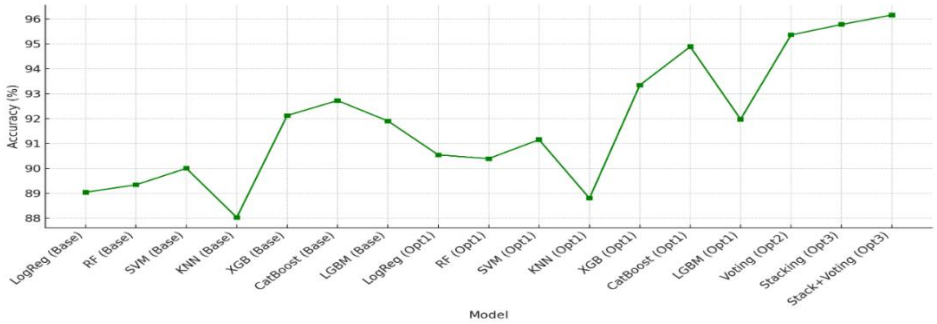



Note: This figure presents the ROC–AUC performance of multiple machine learning models across baseline and successive optimization levels within the HOEL framework, including Level 1 (hyperparameter tuning), Level 2 (voting ensemble), and Level 3 (stacking with voting). The metric reflects the models' ability to discriminate between defaulters and non-defaulters.

As shown in Figure 10, baseline models show strong performance for XGBoost, CatBoost, and Random Forest, with further improvements observed after Level 1 optimization. The introduction of the voting ensemble in Level 2 leads to additional gains in ROC–AUC, reflecting improved predictive stability. The highest performance is achieved at Level 3, where the hybrid voting–stacking configuration reaches an ROC–AUC of 97.62%. This progression demonstrates the effectiveness of the HOEL framework in enhancing model discrimination through multi-level optimization.

Similarly, Figure 11 presents the classification accuracy of the models across the same optimization levels.

**Figure 11.** Accuracy Comparison Across Models and Optimization Levels




Note: This figure presents the classification accuracy of machine learning models across different optimization levels within the HOEL framework, including baseline models, Level 1 optimized models, Level 2 voting ensemble, and Level 3 stacking configurations.

As reported in Figure 11, accuracy improves progressively across optimization levels. CatBoost demonstrates notable gains after Level 1 optimization, while the ensemble approaches in Level 2 and Level 3 achieve the highest performance. The stacking-with-voting configuration attains the best accuracy (96.15%), outperforming all baseline and intermediate models. These results indicate that the hybrid structure enhances generalization and classification precision in credit risk prediction.

Although the hybrid voting–stacking approach yields superior performance, it involves higher computational complexity compared to individual classifiers and standalone ensemble methods, reflecting a trade-off between predictive accuracy and computational cost.

The hybrid voting–stacking configuration consistently outperforms all evaluated models across both ROC–AUC and accuracy metrics. This superior performance can be attributed to the sequential design of the HOEL architecture, where the intermediate voting layer filters weaker classifier outputs before they are passed to the stacking meta-learner. This process enables more precise nonlinear learning and reduces the influence of unstable predictions on the final decision.

While optimization-based approaches such as Hui et al. (2024) and Li et al. (2025) focus on improving financial decision-making through portfolio construction and risk–return optimization, the present study demonstrates the effectiveness of a data-driven classification framework in credit-risk prediction. This distinction underscores the complementary roles of predictive modeling and optimization techniques in financial decision systems.

The diagnostic checks reported in Section 6.2 further reinforce the reliability of the HOEL framework's predictive outcomes. Multicollinearity analyses confirmed that the predictors were statistically appropriate for modeling. Although residuals deviated from normality, this outcome is consistent with the heavy-tailed structure of financial risk data (Cont, 2001) and underscores the suitability of nonlinear approaches. Furthermore, specification tests (RESET and logistic link) strongly rejected linear adequacy, directly supporting the adoption of nonlinear and ensemble learners. In addition, the ADF results confirmed that all continuous predictors were I(0), reducing—but not eliminating—the classical risk of spurious correlations, particularly when interpreted alongside complementary diagnostic checks. Taken together, these diagnostics indicate that the empirical gains achieved by the HOEL framework are unlikely to be artifacts of model misspecification or misleading correlations, but instead reflect its robustness in modeling complex borrower behaviors.

## 7. Conclusion and Future Recommendations

This study is driven by the pressing need to improve credit risk assessment in the rapidly evolving domain of peer-to-peer (P2P) digital lending. As financial technology continues to reshape the lending landscape, traditional credit scoring methods struggle to capture the complexity of borrower behavior, particularly in the presence of data imbalance, nonlinear relationships, and model overfitting. These challenges motivate the development of more accurate, interpretable, and scalable predictive frameworks. In response, this study proposes a hybrid ensemble learning approach (HOEL) that integrates parallel and sequential ensemble strategies within a unified, multi-level architecture.

The empirical findings demonstrate that the proposed HOEL framework significantly enhances predictive performance compared to both individual classifiers and traditional ensemble approaches. The model achieves an accuracy of 96.15% and an ROC–AUC of 97.62%, indicating strong discriminative capability in credit-risk prediction. These results highlight the effectiveness of hybrid ensemble strategies in

capturing complex patterns in financial data and improving the reliability of borrower classification in P2P lending environments.

The proposed HOEL framework makes several key contributions to the methodological literature on credit scoring through a three-level optimized ensemble framework that enhances predictive stability and classification accuracy in P2P credit-risk assessment. By integrating hyperparameter tuning, an intermediate weighted-voting layer, and stacking within a unified architecture, the proposed approach reduces bias, variance, and overfitting while improving generalization across diverse borrower segments. Empirical results further demonstrate that incorporating a voting layer prior to stacking enhances both accuracy and robustness compared with conventional ensemble methods, supporting the framework's practical viability for P2P lending platforms. This contribution is particularly relevant for both academics and practitioners, as it provides a scalable and data-driven framework that enhances credit-risk decision-making under uncertainty while maintaining robustness in imbalanced financial datasets.

From an academic perspective, this research extends existing work on ensemble learning by proposing a hybrid architecture that balances predictive accuracy, stability, and interpretability. From a practical standpoint, the HOEL framework provides financial institutions and P2P lending platforms with a reliable decision-support tool for credit-risk assessment. By enabling more accurate borrower classification and reducing exposure to high-risk lending, the framework supports improved risk management, enhanced portfolio performance, and increased investor confidence in digital lending markets.

The reliability of the HOEL framework is further supported by comprehensive diagnostic testing. The predictors were found to be statistically appropriate, while nonlinearity tests confirmed the suitability of ensemble methods over linear models. Although residuals exhibit non-normal characteristics, this behavior is consistent with the heavy-tailed nature of financial data and does not undermine model validity. These findings reinforce the robustness of the framework for real-world applications.

Despite its advantages, this study has several limitations. The proposed framework involves higher computational complexity compared to conventional models due to its multi-layered structure. In addition, the analysis is based on a single dataset, which may limit the generalizability of the results across different financial contexts. The absence of temporal variables and macroeconomic indicators also restricts the ability to capture time-dependent credit risk dynamics, particularly during periods of economic instability.

Future research can extend this work in several directions. First, improving computational efficiency through parallel processing and cloud-based deployment could enhance scalability for real-time applications. Second, incorporating alternative data sources—such as behavioral or unstructured data—may further improve predictive accuracy. Third, integrating classification-based credit-risk models with portfolio optimization frameworks (Hui et al., 2024; Li et al., 2025) offers a promising avenue for linking borrower-level risk prediction with investment decision-making in financial systems.

## References


Abbasi, K., Alam, A., Brohi, N. A., Brohi, I. A., & Nasim, S. (2021). P2P lending fintechs and SMEs' access to finance. *Economics Letters, 204, 109890.* https://doi.org/10.1016/j.econlet.2021.109890

Abedin, M. Z., Guotai, C., Hajek, P., & Zhang, T. (2022). Combining weighted SMOTE with ensemble learning for the class-imbalanced prediction of small business credit risk. *Complex & Intelligent Systems, 9(4), 3559–3579.* https://doi.org/10.1007/s40747-021-00614-4

Aleksandrova, Y. (2021). Comparing performance of machine learning algorithms for default risk prediction in peer-to-peer lending. *TEM Journal, 10(1), 133–143.* https://doi.org/10.18421/tem101-16

Bone-Winkel, G. F., & Reichenbach, F. (2024). Improving credit risk assessment in P2P lending with explainable machine learning survival analysis. *Digital Finance* https://doi.org/10.1007/s42521-024-00114-3

Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123–140. https://doi.org/10.1023/A:1018054314350

Chang, A.-H., Yang, L.-K., Tsaih, R.-H., & Lin, S.-K. (2022). Machine learning and artificial neural networks to construct P2P lending credit-scoring model: A case using Lending Club data. *Quantitative Finance and Economics, 6(2), 303–325.* https://doi.org/10.3934/qfe.2022013

Chang, C. L., McAleer, M., & Wong, W. K. (2018). Editorial statement of intent for *Advances in Decision Sciences (ADS): 22nd anniversary special issue in 2018* (No. EI2018-40). *Advances in Decision Sciences*. https://doi.org/10.47654/v22y2018iSIp1-2

Cheng, Y., Hui, Y., McAleer, M., & Wong, W. K. (2021). Spurious relationships for nearly non-stationary series. *Journal of Risk and Financial Management, 14*(8), 366. https://doi.org/10.3390/jrfm14080366

Coakley, J., & Huang, W. (2020). P2P lending and outside entrepreneurial finance. *The European Journal of Finance, 1–18.* https://doi.org/10.1080/1351847x.2020.1842223

Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance, 1(2),* 223–236. https://doi.org/10.1080/713665670

Coşkun, S. B., & Turanli, M. (2023). Credit risk analysis using boosting methods. *Journal of Applied Mathematics, Statistics and Informatics, 19(1), 5–18.* https://doi.org/10.2478/jamsi-2023-0001

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, *74*(366a), 427–431. https://doi.org/10.2307/2286348

Dietterich, T. G. (2000). Ensemble methods in machine learning. In J. Kittler & F. Roli (Eds.), *Multiple classifier systems: First International Workshop, MCS 2000* (Lecture Notes in Computer Science, Vol. 1857, pp. 1–15). Springer. https://doi.org/10.1007/3-540-45014-9_1

Duarte, J., Siegel, S., & Young, L. A. (2023). The evolution of P2P lending. *SSRN Electronic Journal* https://doi.org/10.2139/ssrn.4476944

Durbin, J., & Watson, G. S. (1950). Testing for serial correlation in least squares regression. I. *Biometrika, 37*(3–4), 409–428. https://doi.org/10.1093/biomet/37.3-4.409

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1), 119–139. https://doi.org/10.1006/jcss.1997.1504

Hamori, S., & Kume, T. (2018). Artificial intelligence and economic growth. *Advances in Decision Sciences, 22*(1), 1–22. https://doi.org/10.47654/v22y2018i1p1-22

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Hui, Y., Shi, M., Wong, W. K., & Zheng, S. (2024). Pragmatic attitude to large-scale Markowitz's portfolio optimization and factor-augmented derating. *International Review of Financial Analysis, 96*, 103628. https://doi.org/10.1016/j.irfa.2024.103628

Hui, Y., Wong, W. K., Bai, Z., & Zhu, Z. (2017). A new nonlinearity test to circumvent the limitation of Volterra expansion with application. *Journal of the Korean Statistical Society, 46*(3), 365–374. https://doi.org/10.1016/j.jkss.2016.11.006

Jarque, C. M., & Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters, 6(3),* 255–259. https://doi.org/10.1016/0165-1765(80)90024-5

Jayaram, E. S. (2024). Leveraging machine learning techniques for developing robust credit scores for peer-to-peer lending platforms. *Kuwait Journal of Science, 30(5), 12958–12966.* https://doi.org/10.53555/kuey.v30i5.5633

Kaggle. (2024). *Credit risk dataset: Your machine learning and data science community* https://www.kaggle.com/datasets/laotse/credit-risk-dataset/data

Kokate, S., & Chetty, M. S. (2021). Credit risk assessment of loan defaulters in commercial banks using voting classifier ensemble learner machine learning model. *International Journal of Safety and Security Engineering, 11(5), 565–572.* https://doi.org/10.18280/ijsse.110508

Laborda, J., & Ryoo, S. (2021). Feature selection in a credit scoring model. *Mathematics, 9(7), 746.* https://doi.org/10.3390/math9070746

Lenka, S. R., Bisoy, S. K., & Priyadarshini, R. (2024). Multiple optimized ensemble learning for high-dimensional imbalanced credit scoring datasets. *Knowledge and Information Systems, 66(9), 5429–5457.* https://doi.org/10.1007/s10115-024-02129-z

Li, Z., Hui, Y., Wong, W. K., & Lin, R. (2025). Portfolio selection based on mean-generalized variance analysis: Evidence from the G20 stock markets. *Asia-Pacific Journal of Operational Research, 42*(3), 2450016. https://doi.org/10.1142/S021759592450016X

Madaan, M., Kumar, A., Keshri, C., Jain, R., & Nagrath, P. (2021). Loan default prediction using decision trees and random forest: A comparative study. *IOP Conference Series: Materials Science and Engineering, 1022(1), 012042.* https://doi.org/10.1088/1757-899X/1022/1/012042

Meshref, H. (2020). Predicting loan approval of bank direct marketing data using ensemble machine learning algorithms. *International Journal of Circuits, Systems and Signal Processing, 14, 914–922.* https://doi.org/10.46300/9106.2020.14.117

Munsarif, M., Sam'an, M., & Safuan, S. (2022). Peer-to-peer lending risk analysis based on embedded technique and stacking ensemble learning. *Bulletin of Electrical Engineering and Informatics, 11(6), 3483–3489.* https://doi.org/10.11591/eei.v11i6.3927

Naidu, G., Zuva, T., & Sibanda, E. M. (2023). A review of evaluation metrics in machine learning algorithms. *In Computer Science Online Conference (pp. 15–25). Cham: Springer International Publishing* https://doi.org/10.1007/978-3-031-35314-7_2

Najaf, K., Subramaniam, R. K., & Atayah, O. F. (2021). Understanding the implications of FinTech peer-to-peer (P2P) lending during the COVID-19 pandemic. *Journal of Sustainable Finance & Investment, 12(1), 87–102.* https://doi.org/10.1080/20430795.2021.1917225

Nguyen, L., Ahsan, M., & Haider, J. (2024). Reimagining peer-to-peer lending sustainability: Unveiling predictive insights with innovative machine learning approaches for loan default anticipation. *FinTech, 3(1), 184–215.* https://doi.org/10.3390/fintech3010012

Noriega, J. P., Rivera, L. A., & Herrera, J. A. (2023). Machine learning for credit risk prediction: A systematic literature review. *Preprints* https://doi.org/10.20944/preprints202308.0947.v1

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity, 41(5),* 673–690. https://doi.org/10.1007/s11135-006-9018-6

Perera, C. L., & Premaratne, S. C. (2023). An ensemble machine learning approach for forecasting credit risk of loan applications. *WSEAS Transactions on Systems, 23, 31–46.* https://doi.org/10.37394/23202.2024.23.4

Ramsey, J. B. (1969). Tests for specification errors in classical linear least squares regression analysis. *Journal of the Royal Statistical Society: Series B (Methodological), 31(2),* 350–371. https://doi.org/10.1111/j.2517-6161.1969.tb00796.x

Schapire, R. E. (1990). The strength of weak learnability. Machine Learning, 5(2), 197–227. https://doi.org/10.1007/BF00116037

Shih, D.-H., Wu, T.-W., Shih, P.-Y., Lu, N.-A., & Shih, M.-H. (2022). A framework of global credit-scoring modeling using outlier detection and machine learning in a P2P lending platform. *Mathematics, 10(13), 2282.* https://doi.org/10.3390/math10132282

Trinh, L. T. (2024). A comparative analysis of consumer credit risk models in peer-to-peer lending. *Journal of Economics, Finance and Administrative Science, 29(57), 57–73.* https://doi.org/10.1108/jefas-04-2021-0026

Uddin, N., Ahamed, M. K., Uddin, M. A., Islam, M. M., & Aryal, S. (2023). An ensemble machine learning based bank loan approval predictions system with a smart application. *SSRN Electronic Journal* https://doi.org/10.2139/ssrn.4376481

Verified Market Research. (2024, December). *Peer to peer (P2P) lending market by business model (traditional P2P platforms, marketplace lending, blockchain-based P2P lending), loan type (personal loans, business loans, student loans, real estate loans, auto loans), & region for 2024 to 2031* [Market research report]. https://www.verifiedmarketresearch.com/product/peer-to-peer-p2p-lending-market/

Wattanakitrungroj, N., Wijitkajee, P., Jaiyen, S., Sathapornvajana, S., & Tongman, S. (2024). Enhancing supervised model performance in credit risk classification using sampling strategies and feature ranking. *Big Data and Cognitive Computing, 8(3), 28.* https://doi.org/10.3390/bdcc8030028

Wei, Y., Kirkulak-Uludag, B., Zhu, D., & Zhou, Z. (2023). Stacking ensemble method for personal credit risk assessment in P2P lending. *SSRN Electronic Journal* https://doi.org/10.2139/ssrn.4318348

Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Wong, W. K., Cheng, Y., & Yue, M. (2024). Could regression of stationary series be spurious? *Asia-Pacific Journal of Operational Research*. https://doi.org/10.1142/s0217595924400177

Wong, W. K., & Pham, M. T. (2022a). Could the test from the standard regression model make significant regression with autoregressive noise become insignificant? A note. *The International Journal of Finance, 34*(1), 1–18. https://tijof.scibiz.world/ijof-2022_01

Wong, W. K., & Pham, M. T. (2022b). Could the test from the standard regression model make significant regression with autoregressive noise become insignificant? A note. *The International Journal of Finance, 34*, 19–39.

Wong, W. K., & Yue, M. (2024). Could regressing a stationary series on a non-stationary series obtain meaningful outcomes? *Annals of Financial Economics, 19*(3). https://doi.org/10.1142/s2010495224500118

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing, 415, 295–316.* https://doi.org/10.1016/j.neucom.2020.08.013

Yang, R. (2024). Machine learning-based loan default prediction in peer-to-peer lending. *Highlights in Science, Engineering and Technology, 94, 310–318.* https://doi.org/10.54097/qdjd8r65

Yeh, J.-Y., Chiu, H.-Y., & Huang, J.-H. (2024). Predicting failure of P2P lending platforms through machine learning: The case in China. *Finance Research Letters, 59, 104784.* https://doi.org/10.1016/j.frl.2023.104784

Zhang, T., & Sun, W. (2023). Research on online loan default prediction model based on ensemble learning. *In Proceedings of the 2nd International Conference on Mathematical Statistics and Economic Analysis (MSEA 2023) (pp. 26–28).* https://doi.org/10.4108/eai.26-5-2023.2334378

Zhao, Z., Cui, T., Ding, S., Li, J., & Bellotti, A. G. (2024). Resampling techniques study on class imbalance problem in credit risk prediction. *Mathematics, 12(5), 701.* https://doi.org/10.3390/math12050701

## Appendix

**Table A1.** Pearson Correlation Coefficients Among P2P Lending Variables

| Variable | Person Age | Person Income | Employment Length | Loan Amount | Loan Interest Rate | Loan Status | Loan Percent Income | Credit History Length |
|---|---|---|---|---|---|---|---|---|
| Person Age | **1** | 0.069 *** | 0.13 *** | 0.018 *** | 0.002 | -0.04 *** | -0.039 *** | **0.85** *** |
| Person Income | 0.069 *** | **1** | 0.19 *** | **0.45** *** | -0.045 *** | **-0.23** *** | **-0.28** *** | 0.042 *** |
| Employment Length | 0.13 *** | 0.19 *** | **1** | 0.099 *** | -0.075 *** | -0.088 *** | -0.048 *** | 0.11 *** |
| Loan Amount | 0.018 *** | **0.45** *** | 0.099 *** | **1** | 0.12 *** | 0.10 *** | **0.66** *** | 0.008 |
| Loan Interest Rate | 0.002 | -0.045 *** | -0.075 *** | 0.12 *** | **1** | **0.34** *** | 0.13 *** | 0.007 |
| Loan Status | -0.04 *** | **-0.23** *** | -0.088 *** | 0.10 *** | **0.34** *** | **1** | **0.33** *** | -0.03 *** |
| Loan Percent Income | -0.039 *** | **-0.28** *** | -0.048 *** | **0.66 ***** | 0.13 *** | **0.33** *** | **1** | -0.028 *** |
| Credit History Length | **0.85** *** | 0.042 *** | 0.11 *** | 0.008 | 0.007 | -0.03 *** | -0.028 *** | **1** |

Note: This table presents Pearson correlation coefficients (r) among all continuous variables in the cleaned peer-to-peer lending dataset. Statistical significance levels are denoted as: *** $p < .01$, ** $p < .05$, and * $p < .10$. Because the sample size is relatively large, many correlations appear statistically significant; therefore, interpretation emphasizes practical effect size rather than significance alone. Positive coefficients indicate direct linear relationships, while negative coefficients reflect inverse associations. Values in bold ($|r| \geq 0.20$) represent relatively stronger correlations, and diagonal entries equal 1.00 by definition.