forecasting

MDPI

Article

Optimizing Credit Risk Prediction for Peer-to-Peer Lending Using Machine Learning

Lyne Imene Souadda [1], Ahmed Rami Halitim [2], Billel Benilles [1], José Manuel Oliveira [3,4,*] and Patrícia Ramos [3,5]

[1] Applied Studies in Business and Management Sciences Laboratory, Finance Department, Higher School of Commerce, Kolea University Center, Kolea 42003, Tipaza, Algeria; l_souadda@esc-alger.dz (L.I.S.); b_benilles@esc-alger.dz (B.B.)
[2] Statistics Department, National School of Statistics and Applied Economics, Kolea University Center, Kolea 42003, Tipaza, Algeria; ahmedrami.halitim@trustbank.dz
[3] Institute for Systems and Computer Engineering, Technology and Science, Campus da FEUP, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal; patricia@iscap.ipp.pt
[4] Faculty of Economics, University of Porto, Rua Dr. Roberto Frias, 4200-464 Porto, Portugal
[5] CEOS.PP, ISCAP, Polytechnic of Porto, Rua Jaime Lopes Amorim s/n, 4465-004 São Mamede de Infesta, Portugal
* Correspondence: jmo@fep.up.pt

Abstract

Hyperparameter optimization (HPO) is critical for enhancing the predictive performance of machine learning models in credit risk assessment for peer-to-peer (P2P) lending. This study evaluates four HPO methods, Grid Search, Random Search, Hyperopt, and Optuna, across four models, Logistic Regression, Random Forest, XGBoost, and LightGBM, using three real-world datasets (Lending Club, Australia, Taiwan). We assess predictive accuracy (AUC, Sensitivity, Specificity, G-Mean), computational efficiency, robustness, and interpretability. LightGBM achieves the highest AUC (e.g., 70.77% on Lending Club, 93.25% on Australia, 77.85% on Taiwan), with XGBoost performing comparably. Bayesian methods (Hyperopt, Optuna) match or approach Grid Search's accuracy while reducing runtime by up to 75.7-fold (e.g., 3.19 vs. 241.47 min for LightGBM on Lending Club). A sensitivity analysis confirms robust hyperparameter configurations, with AUC variations typically below 0.4% under ±10% perturbations. A feature importance analysis, using gain and SHAP metrics, identifies debt-to-income ratio and employment title as key default predictors, with stable rankings (Spearman correlation $> 0.95, p < 0.01$) across tuning methods, enhancing model interpretability. Operational impact depends on data quality, scalable infrastructure, fairness audits for features like employment title, and stakeholder collaboration to ensure compliance with regulations like the EU AI Act and U.S. Equal Credit Opportunity Act. These findings advocate Bayesian HPO and ensemble models in P2P lending, offering scalable, transparent, and fair solutions for default prediction, with future research suggested to explore advanced resampling, cost-sensitive metrics, and feature interactions.

Keywords: credit risk; ensemble learning; hyperparameter optimization; peer-to-peer lending

MSC: 91G40; 68T01; 62H25

check for updates

Academic Editor: Sonia Leva

Received: 29 May 2025
Revised: 25 June 2025
Accepted: 27 June 2025
Published: 29 June 2025

Citation: Souadda, L.I.; Halitim, A.R.; Benilles, B.; Oliveira, J.M.; Ramos, P. Optimizing Credit Risk Prediction for Peer-to-Peer Lending Using Machine Learning. *Forecasting* **2025**, *7*, 35. https://doi.org/10.3390/forecast7030035

1. Introduction

Peer-to-peer (P2P) lending, also known as social lending, is an innovative financial service that connects borrowers directly with lenders through online platforms, bypassing traditional intermediaries like banks [1]. By leveraging digital infrastructure and social media, P2P lending enhances access to credit, fostering financial inclusion, particularly for underserved populations.

The availability and regulatory frameworks of P2P lending vary significantly across countries. In the United Kingdom, the P2P lending market is well established and regulated by the Financial Conduct Authority (FCA), with platforms such as Zopa and Funding Circle achieving broad adoption [2]. In the United States, the market is sizable, dominated by Lending Club and Prosper, but operates under a fragmented regulatory landscape governed by both state and federal laws. Several European countries, including Germany, France, and the Netherlands, have embraced P2P lending within specific regulatory regimes, although the market remains modest in scale compared to Anglo-American counterparts. In the Asia-Pacific region, rapid growth is anticipated, with China leading the market. Countries like Japan and South Korea also show steady expansion, supported by robust regulation and a strong culture of financial innovation. In contrast, countries such as Canada and India maintain strict regulatory environments; for instance, Indian platforms must register as non-banking financial companies (NBFCs). In some jurisdictions, such as Pakistan, Indonesia, and parts of the Middle East, P2P lending remains either prohibited or heavily restricted due to regulatory or religious constraints [3].

Globally, the P2P lending market has expanded considerably. Analysts estimate that the total market value may have reached up to USD 490 billion by 2022 and is projected to grow to USD 804.2 billion by 2030, reflecting a compound annual growth rate (CAGR) of approximately 29.1% from 2022 to 2030 [4]. North America holds the largest share of the global market, with established platforms like Lending Club and Prosper facilitating loans totaling USD 61 billion between 2018 and 2020. The market is expected to grow to USD 150 billion by 2025 [5]. Europe is also experiencing robust growth, though its overall market size remains smaller than North America's. The United Kingdom currently ranks as the third-largest P2P lending market globally, following the United States and China [3,6]. In 2017, the UK's market volume reached EUR 6.2 billion, with an annual growth rate of 35.2% since 2011 [5]. Asia-Pacific is the fastest-growing region. China, which once hosted 1553 active platforms and recorded a transaction volume of RMB 4.648 trillion in 2017 [7], remains the largest market despite recent regulatory crackdowns aimed to limit P2P lending platforms [3]. Other rapidly expanding markets include India, South Korea, and Japan, which is projected to maintain a CAGR of around 9.1% between 2024 and 2034. These figures highlight the increasing global significance of P2P lending, despite its uneven development across regions.

Nevertheless, P2P lending faces significant challenges, notably information asymmetry and rising default rates, which undermine platform stability and profitability [8,9]. To address these risks, P2P platforms employ internal credit scoring systems, framing credit risk assessment as a binary classification task, with loan repayment status as the target variable (fully repaid loans labeled "1" and defaulted loans "0") [10]. Traditional creditworthiness evaluation methods, such as Fair Isaac Corporation (FICO) scores and subjective assessments, often fail to capture complex borrower behaviors and non-linear relationships among predictive features. In response, advanced statistical and machine learning models have emerged, offering superior predictive accuracy and interpretability by leveraging high-dimensional borrower data, dynamic credit behaviors, and intricate feature interactions [11,12].

Ensemble learning models, which aggregate predictions from multiple base learners, consistently outperform individual classifiers and traditional statistical approaches in credit risk modeling [12]. However, their effectiveness hinges on hyperparameter optimization, as parameters such as learning rates, tree depths, and regularization terms govern model complexity and generalization. For example, an excessively high learning rate may lead to premature convergence, missing critical default patterns, while overly complex architectures, such as deep decision trees, risk overfitting, reducing generalizability to new loan applications [13].

Exhaustive hyperparameter tuning, such as Grid Search, is computationally intensive, and suboptimal configurations can significantly impair performance. Moreover, hyperparameter tuning strategies influence feature importance stability, which is critical for model interpretability and risk factor prioritization, yet this aspect remains underexplored [14]. The unique characteristics of P2P lending, including simplified underwriting and limited regulatory oversight, exacerbate default risks compared to traditional credit systems [15]. Even marginal improvements in default prediction can significantly enhance platform stability and reduce financial exposure [16,17].

This study addresses these challenges by systematically benchmarking two Bayesian optimization frameworks, Hyperopt [18] and Optuna [19], against conventional methods, Grid Search (GS) and Random Search (RS) [20], to optimize credit risk prediction models. Bayesian optimization offers a computationally efficient alternative to traditional methods, particularly for complex models [21,22]. We evaluate Logistic Regression [23] and ensemble models, including Random Forest [24], XGBoost [25], and LightGBM [26], using the Lending Club (LC) dataset [27], supplemented by the Australia [28] and Taiwan [29] credit card datasets for broader applicability.

Our findings demonstrate that Bayesian optimization methods (Hyperopt and Optuna) achieve predictive performance comparable to Grid Search while reducing computational costs by up to 75.7-fold, with Hyperopt being particularly efficient. Models trained with optimized hyperparameters significantly outperform those with default settings, highlighting the critical role of hyperparameter optimization (HPO). Additionally, we provide a comprehensive analysis of hyperparameter tuning's impact on model performance, sensitivity, and feature importance stability, offering actionable insights for P2P lending platforms to enhance credit scoring efficiency and interpretability.

This research is guided by the following questions:

1. Which hyperparameter tuning method offers the optimal balance between computational efficiency and predictive performance for credit scoring in P2P lending?
2. How sensitive is model performance to perturbations in optimal hyperparameter configurations?
3. How do hyperparameter tuning strategies affect the stability and consistency of model interpretability?

The paper is organized as follows: Section 2 reviews the relevant literature on P2P lending, default prediction, and machine learning applications in credit risk. Section 3 details the methodological approach. Section 4 describes the experimental setup, and Section 5 presents the results. Section 6 concludes with implications for future research.

2. Related Work

Peer-to-peer lending has garnered significant attention for its potential to democratize access to credit by providing faster and more flexible funding solutions. However, the absence of collateral and limited regulatory oversight expose the industry to considerable credit risk, requiring robust methods to accurately predict loan defaults. Traditionally, online financial platforms have relied on quantitative credit scoring systems, such as FICO scores, or proprietary metrics like the Lending Club score. Although these methods offer initial assessments of borrower creditworthiness, they fall short in accurately predicting default risk. Their reliance on general risk scores limits their ability to capture nuanced interactions between variables, leaving investors with insufficient confidence in the borrower's reliability [30].

In response to these limitations, P2P platforms have begun framing credit evaluation as a binary classification task, where the goal is to predict the likelihood of default based on borrower demographics and financial data. Early studies utilized statistical techniques such as Logistic Regression (LR) and Linear Discriminant Analysis (LDA) [31,32]. While these methods proved effective in identifying key predictive variables, their assumption of linear relationships limited their applicability to complex, high-dimensional datasets.

2.1. Machine Learning for Credit Scoring

Machine learning techniques have emerged as a promising alternative for credit risk prediction [12]. Methods such as decision trees (DTs), support vector machines (SVMs), and artificial neural networks (ANNs) have shown significant improvements in predictive performance [33–35]. For example, Teply and Polena [36] conducted a comparative evaluation of several classifiers, including Logistic Regression (LR), Artificial Neural Networks, Support Vector Machines, Random Forests (RF), and Bayesian Networks (BNs), on the Lending Club dataset, spanning 2009–2013. They found that Logistic Regression achieved the highest classification accuracy. However, in a similar analysis of the same dataset, Malekipirbazari and Aksakalli [30] revealed that Random Forests outperformed both Logistic Regression and FICO scores.

Numerous studies have conducted comparative evaluations to identify the best algorithm for these classification tasks. Lessmann et al. [12] assessed several classification techniques across a few credit scoring datasets. Their study explored advanced approaches, such as heterogeneous ensembles, rotation forests, and extreme learning machines, while using robust performance measures like the H-measure and partial Gini index. The findings highlighted the superiority of ensemble models over traditional methods, offering valuable insights for improving credit risk modeling.

Building on these advancements, ensemble learning has become a widely adopted approach. Techniques such as Random Forests and gradient boosting combine multiple base models to enhance predictive accuracy and robustness. By integrating diverse algorithms that evaluate various hypotheses, these models produce more reliable predictions, leading to improved precision [16,17]. Ensemble models are generally categorized into parallel and sequential structures. Parallel ensembles, like Bagging and Random Forests, combine independently trained models to make collective decisions simultaneously. In contrast, sequential ensembles, such as boosting algorithms, iteratively refine models by correcting errors from previous iterations.

Ma et al. [37] highlighted LightGBM's computational efficiency, showing it achieved comparable accuracy to extreme gradient boosting (XGBoost) while running ten times faster. Similarly, Ko et al. [38] benchmarked three statistical models: Logistic Regression, Bayesian Classifier, and Linear Discriminant Analysis, alongside five machine learning models, decision tree, Random Forest, LightGBM, Artificial Neural Network, and Convolutional Neural Network (CNN), using the Lending Club dataset. Their findings showed that LightGBM outperformed the other models, achieving a 2.91% improvement in accuracy. However, the study did not consider hyperparameter tuning strategies, which could significantly affect model performance. Additionally, their focus on classification accuracy overlooked computational efficiency, a crucial factor for real-world deployment.

2.2. Challenges in Social Lending Datasets

A persistent challenge in credit scoring is class imbalance, where the majority of loans do not default [10,11]. This imbalance causes models to be biased towards the majority class, leading to reduced sensitivity toward the minority class [38,39]. This issue is particularly critical for predicting defaults, as failing to identify defaulted loans can expose lenders to significant financial risks [11]. Traditional machine learning models often assume equal class distributions and focus on optimizing overall accuracy, rather than prioritizing balanced performance metrics such as Sensitivity, Specificity, or Geometric Mean (G-Mean) [32,40].

Resampling techniques have emerged as effective solutions to address class imbalance [41]. Oversampling methods, like the synthetic minority oversampling technique (SMOTE), generate synthetic examples to augment the minority class, while undersampling reduces the majority class. Krishna Veni and Sobha Rani [40] identified three main reasons why traditional classification algorithms struggle with imbalanced data: (1) they are primarily accuracy-driven, often favoring the majority class; (2) they assume an equal distribution of classes, which is rarely the case in imbalanced datasets; and (3) they treat the misclassification error costs of all classes as equal. To overcome these challenges, the study recommended the use of sampling strategies and cost-sensitive learning techniques. Additionally, alternative performance metrics, such as the confusion matrix, precision, and F1-score, were used to better evaluate models on unbalanced datasets.

In another study, Alam et al. [16] investigated the prediction of credit card default using imbalanced datasets, with a focus on enhancing classifier performance while preserving the interpretability of the model. Using datasets like the South German Credit and Belgium Credit data, the study addressed the challenge of class imbalance through various resampling techniques. Gradient-boosted decision trees (GBDTs) were used as the primary model, with hyperparameter tuning of learning rates and the number of trees to improve predictive accuracy. The K-means SMOTE oversampling technique demonstrated significant improvements in G-Mean, precision, and recall, effectively addressing the issue of class imbalance.

Several studies have specifically utilized the Lending Club dataset for credit risk prediction [22]. Moscato et al. [10] highlighted the bias of machine learning models toward the majority class in imbalanced settings. To counter this, they employed oversampling techniques like SMOTE, alongside undersampling methods. The study demonstrated significant improvements in default prediction by using a Random Forest classifier combined with resampling techniques. Additionally, several explainability methods, such as LIME and SHAP, were incorporated to enhance model transparency.

Similarly, Namvar et al. [32] investigated the effectiveness of resampling strategies combined with machine learning classifiers like Logistic Regression and Random Forest. Their findings underscored the importance of metrics such as G-Mean, which balance Sensitivity and Specificity, in effectively addressing imbalanced datasets. The results indicated that combining Random Forests with random undersampling could be a promising approach for assessing credit risk in social lending markets. Song et al. [15] adopted a different approach by introducing an ensemble-based methodology which leverages multi-view learning and adaptive clustering to increase base learner diversity. This method showed superior Sensitivity and adaptability in default identification, outperforming traditional classifiers in highly imbalanced settings.

While addressing class imbalance is crucial, it is not sufficient for optimizing predictive performance in P2P lending datasets due to the complexities introduced by the high dimensionality and intricate feature interactions. Hyperparameter tuning offers a complementary approach that improves model Sensitivity and accuracy. Huang and Boutros [42] highlighted that optimal hyperparameters often vary across datasets, underscoring the importance of tailoring tuning strategies to the specific characteristics of each dataset.

Among existing approaches, Bayesian hyperparameter optimization has proven a more effective alternative, consistently outperforming traditional Grid Search in various empirical studies [43,44]. Chang et al. [45] emphasized the effectiveness of combining models, such as Random Forests and Logistic Regression, to predict loan defaults. Their analysis, using the Lending Club dataset from 2007 to 2015, found that Logistic Regression, enhanced with misclassification penalties and Gaussian Naive Bayes, achieved competitive performance, with Naive Bayes showing the highest Specificity. The study also highlighted the crucial role of hyperparameter tuning, especially for SVMs, where kernel selection and regularization parameters significantly impact model performance.

Xia et al. [22] aimed to develop an accurate and interpretable credit scoring model using XGBoost, with a focus on hyperparameter tuning and sequential model building. Using datasets from a publicly available credit scoring competition that featured class imbalance, the authors employed Bayesian optimization, specifically the tree-structured Parzen estimator (TPE), for adaptive hyperparameter tuning. This method significantly improved SVM performance compared to manual tuning, Grid Search, and Random Search. Bayesian optimization yielded approximately 5% higher performance than grid and manual searches, and 3% higher than Random Search. Additionally, the study examined feature importance based on TPE-optimized models. However, the research did not explore whether different hyperparameter optimization methods affected feature rankings.

3. Methodology

Ensemble learning models, such as Random Forest, XGBoost, and LightGBM, are robust classifiers whose predictive performance depends critically on hyperparameter configuration. This study employed four hyperparameter optimization methods, Grid Search, Random Search, Hyperopt, and Optuna, to tune multiple classification models, including Logistic Regression, Random Forest, XGBoost, and LightGBM. GS and RS served as baselines, with identical search spaces defined for all methods to ensure fair comparison. The experimental workflow, illustrated in Figure 1, encompassed data preprocessing, feature selection, hyperparameter tuning, model training, and evaluation.

Data preprocessing addressed key challenges in P2P lending datasets, including missing values, redundancy, and class imbalance. Features with excessive missing values were removed, and redundant attributes were eliminated based on correlation analysis and domain knowledge, retaining only the most informative features. Numerical variables were standardized using a robust scaler to mitigate outlier effects, and multicollinearity was assessed via a variance inflation factor analysis to ensure feature independence. Ordinal variables were encoded using integer encoding to preserve their order, while categorical variables were encoded using one-hot encoding, with missing or invalid entries for both treated as distinct categories. To optimize computational efficiency, feature data types were refined (e.g., converting misclassified floats to integers), reducing memory usage. Class imbalance, prevalent in credit risk datasets, was addressed through random undersampling of the training set to achieve a balanced class distribution. The test set retained its original distribution to reflect real-world lending scenarios, ensuring realistic evaluation metrics. Feature selection was then applied using techniques such as recursive feature elimination to filter out low-importance features, reducing dimensionality while preserving predictive power.

Hyperparameter tuning optimized model performance using four methods: GS exhaustively evaluated all combinations within predefined grids; RS randomly sampled configurations from the same search space; Hyperopt and Optuna leveraged Bayesian optimization with tree-structured Parzen estimators (TPEs) to iteratively focus on promising regions, incorporating pruning to terminate unpromising trials early. Search spaces were model-specific (e.g., `learning_rate`, `max_depth` for XGBoost; `n_estimators`, `min_samples_split` for Random Forest), as detailed in the experimental setup (Section 4).

Post-tuning, models were trained on the preprocessed training set using optimized hyperparameters and the selected feature subset. Performance was evaluated on the test set using metrics tailored for imbalanced datasets: Area Under the Receiver Operating Characteristic Curve (AUC), Sensitivity, Specificity, and Geometric Mean (G-Mean). AUC measures overall discriminative ability, while Sensitivity and Specificity assess performance on default and non-default classes, respectively. G-Mean balances these metrics to account for class imbalance.

To ensure robustness, a sensitivity analysis examined the impact of small perturbations (±10%) in key hyperparameters on model performance, measured by AUC. This analysis quantified configuration stability across datasets. Additionally, feature importance was analyzed for the best-performing model (LightGBM) using the gain metric, which quantified each feature's contribution to prediction error reduction, enhancing model interpretability and identifying key default predictors.

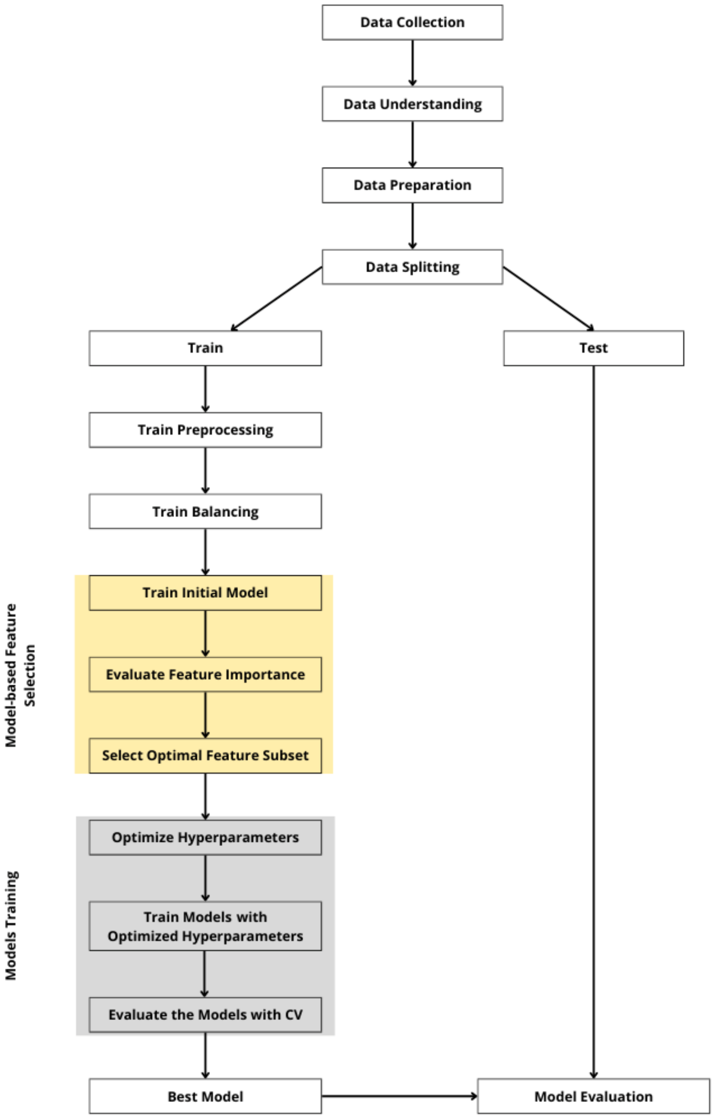

Figure 1. Workflow of the experimental design for credit risk prediction with hyperparameter tuning.

Tuning of Machine Learning Model Hyperparameters

Hyperparameters are user-defined settings that govern a machine learning model's architecture and learning process, distinct from model parameters optimized during training. For example, in XGBoost, hyperparameters like learning rate and maximum tree depth control model complexity, while in neural networks, the number of hidden layers and neurons are critical. Unlike parameters, which adapt to the data, hyperparameters influence accuracy, generalization, and computational efficiency, necessitating careful tuning to optimize performance [41]. In credit risk modeling, where models like Random Forest, XGBoost, and LightGBM are employed, hyperparameter optimization is computationally intensive due to high-dimensional search spaces [13]. This study benchmarked four HPO methods, Grid Search, Random Search, Hyperopt, and Optuna, to identify efficient and effective tuning strategies.

In credit risk modeling, Grid Search has traditionally been the most widely used method for hyperparameter tuning [46]. This algorithm performs an exhaustive search for all possible combinations of a predefined hyperparameter space. The performance of each combination is evaluated using cross-validation, which splits the training dataset into k folds and computes an averaged evaluation metric. This Cartesian product-based approach ensures that the global optimum within the specified search space is found. However, it can be computationally expensive due to the exhaustive exploration of all parameter combinations:

$$S = \prod_{k=1}^{K} |L^{(k)}| \,. \tag{1}$$

In Equation (1), $L^{(k)}$ represents the set of candidate values for the kth hyperparameter, K is the total number of hyperparameters, and S is the total number of evaluations. For example, if optimizing two hyperparameters, learning rate (α) and regularization strength (λ), with $|\alpha| = 3$ and $|\lambda| = 4$, the Grid Search evaluates $3 \times 4 = 12$ configurations. While this method is straightforward and easy to parallelize, it ensures a comprehensive coverage of the hyperparameter space. This process triggers the curse of dimensionality, as the computational cost increases exponentially with the number of hyperparameters [13,20].

In contrast, Random Search introduces stochasticity into the hyperparameter optimization process. Rather than evaluating every combination, it randomly samples a specified set of hyperparameters from the search space, significantly reducing computational cost. The likelihood of selecting the optimal configuration is expressed as:

$$\lambda^{(i)} \sim \text{Uniform}(\Lambda), \quad i = 1, 2, \ldots, S, \tag{2}$$

where Λ is the hyperparameter space, and $\lambda^{(i)}$ represents a hyperparameter configuration sampled uniformly at random. Studies have shown that Random Search can outperform Grid Search in high-dimensional hyperparameter spaces due to its broader exploration. By focusing on random samples, this method can achieve better results with fewer trials when the search space is larger than that of Grid Search. On the other hand, it risks missing promising regions of the search space as its performance depends on the number of trials S and the distribution used for sampling [13,20].

The Bayesian optimization techniques iteratively select and evaluate promising configurations based on the estimated probability density function and the expected improvement metric. They are increasingly used to overcome the limitations of traditional hyperparameter tuning methods. Optuna [19] utilizes state-of-the-art optimization strategies, including the tree-structured Parzen estimator (TPE) method for modeling objective functions and pruning unpromising trials, achieving superior results while reducing computational cost. Unlike Grid Search and Random Search, Optuna constructs probabilistic models to capture the relationship between hyperparameters and the objective function, focusing on promising regions and discarding ineffective configurations based on previous trials. It proposes new hyperparameter values by sampling from learned distributions. Additionally, its pruning mechanism ends the trials prematurely, further minimizing computational overhead [20]. The optimization objective is expressed as:

$$\lambda^* = \arg\max_{\lambda \in \Lambda} \mathbb{E}[\Psi(\lambda)|\text{data}], \tag{3}$$

where $\mathbb{E}[\Psi(\lambda)|\text{data}]$ denotes the expected improvement or acquisition function, balancing the exploration of uncertain regions and the exploitation of promising ones. The model is updated iteratively with observed results, enabling the method to focus on the most promising regions of the search space.

Similarly, Hyperopt is a widely used hyperparameter optimization framework that leverages the tree-structured Parzen estimator (TPE) for an efficient search. Like Optuna, Hyperopt employs Bayesian optimization to iteratively model the relationship between hyperparameters and the objective function, using a TPE to estimate the probability of improvement for each configuration. It samples hyperparameter values from distributions defined over the search space, balancing exploration and exploitation to identify promising regions with fewer evaluations than Grid or Random Search. Hyperopt's flexibility allows it to handle complex search spaces, including continuous, discrete, and conditional hyperparameters, making it suitable for optimizing diverse machine learning models. Its parallelization capabilities enable simultaneous evaluation of multiple configurations across distributed processes, significantly reducing computation time. Additionally, Hyperopt supports early stopping mechanisms, similar to Optuna's pruning, to terminate unpromising trials early, further enhancing efficiency [13,43]. The optimization process follows the same objective as in Equation (3), iteratively updating the probabilistic model to focus on high-performing configurations.

4. Experimental Setup

4.1. Datasets

This study leveraged three real-world public datasets to evaluate the performance of credit risk prediction. The primary dataset was derived from the Lending Club platform, representing peer-to-peer personal loans [27]. In addition, we incorporated two widely used benchmarks: the Australia [28] and the Taiwan [29] credit card datasets. A summary of the three datasets is presented in Table 1.

Table 1. Summary of Lending Club, Australia, and Taiwan datasets for credit risk modeling.

Dataset	Total Records	Attributes	Class Distribution
Lending Club	233,015	52	191,771/41,244
Australia	690	15	383/307
Taiwan	30,000	25	23,364/6636

The Lending Club dataset, one of the largest peer-to-peer lending platforms in the United States, is a benchmark for credit risk analysis, comprising 2,925,297 records and 141 features collected between 2007 and April 2020. To ensure complete outcome information and loan maturity, we restricted the dataset to loans issued in 2014. It includes diverse borrower and financial characteristics, such as loan amounts, interest rates, credit histories, and demographics. Key attributes include seven loan statuses (e.g., "Fully Paid", "Charged Off", "Default", and delinquency stages). Following Namvar et al. [32], we selected these statuses as the target variable, categorizing them into two classes (non-default and default) for a binary classification. The retained features provided a comprehensive view of borrower profiles and financial behavior, including loan-specific variables (e.g., loan amount, term length, interest rate, and loan grade, classified from A to G, with B and C the most common) and borrower characteristics (e.g., annual income, home-ownership status, employment duration). Irrelevant features, such as loan IDs, payment dates, and post-loan attributes, were removed. Features with over 50% missing values were also removed. Redundant attributes (e.g., funded amount, funded amount by investors, loan amount) were identified using the Lending Club data dictionary, retaining only one. Numerical variables were scaled using a robust scaler to minimize outlier impact; this method centers the data by subtracting the median and scales it based on the interquartile range (IQR), making it resilient to extreme values. Categorical variables (e.g., employment title) were encoded using ordinal encoding to maintain lower feature dimensionality, which is particularly beneficial for models like Logistic Regression that are sensitive to high-dimensional spaces. This approach reduces the risk of overfitting and computational complexity by assigning integer values to categories, preserving a compact representation. Unlike one-hot encoding, which can significantly inflate dimensionality and introduce sparsity, especially for features with many unique values, such as employment title, ordinal encoding ensures a streamlined preprocessing pipeline. This choice supports efficient and stable model training, leveraging the robustness of ensemble models (Random Forest, XGBoost, LightGBM) to effectively utilize ordinal representations. Misclassified object features were converted to categorical types, and incorrectly specified float features were corrected to integers for memory efficiency. The dataset was reduced to 52 features and 233,015 rows, decreasing memory usage from 2.2 GB to 0.03 GB. The dataset exhibits substantial class imbalance, with 82.3% of loans classified as “Fully Paid” and only 17.7% as defaulted. To address this, the training set was balanced using random undersampling. This ensured that the models learned equally from both classes, enhancing their ability to detect defaults. As a result, the training set contained 65,990 samples, while the test set retained the original class distribution to better reflect real-world conditions.

The Australian credit approval dataset is a benchmark for credit risk modeling, focusing on credit card applications. It contains 690 records and 14 features (6 continuous, 4 binary, and 4 categorical), used for binary classification to determine whether to approve or reject a credit card application. There are no missing values, and the features, anonymized as A1, A2, …, A14, provide a comprehensive view of applicant profiles. Preprocessing aligned with the Lending Club methodology. Numerical variables were scaled using a robust scaler to reduce outlier impact. Categorical variables were encoded using ordinal encoding, with unknown or invalid values treated as distinct categories. The dataset shows moderate class imbalance, with 55.5% (383 records) approved and 44.5% (307 records) rejected. To address this, random undersampling was applied to the training set to balance the classes.

The Taiwan dataset of credit card clients is a benchmark for credit risk prediction, focusing on credit card default. It comprises 30,000 records and 25 features, collected from credit card clients in Taiwan between April and September 2005. The binary target variable indicates whether a client defaulted or not on their payment the following month. Features include demographic attributes (e.g., age, sex, marital status, education), financial details (e.g., credit limit, bill amounts, payment amounts), and payment behavior (e.g., repayment status over six months, from -1 for paid on time to 8 for severe delinquency), providing a comprehensive view of creditworthiness. The dataset has no missing values, eliminating imputation needs. Numerical variables (e.g., credit limit, bill amounts, payment amounts) were scaled using a robust scaler to mitigate outlier impact. Categorical variables (e.g., education, marital status, sex) were encoded using ordinal encoding, with unknown or invalid values treated as distinct categories. The dataset exhibits significant class imbalance, with 77.9% (23,364 records) classified as non-default and 22.1% (6636 records) as default. To address this, random undersampling was applied to the training set, as with the other datasets.

4.2. Benchmark Models

This study evaluated four machine learning models for credit risk classification: Logistic Regression, Random Forest, XGBoost, and LightGBM. Each model's methodology, strengths, and limitations are described below to contextualize their application in predicting loan defaults. The selection of LR, RF, XGBoost, and LightGBM as benchmark models in this study was driven by their complementary strengths and relevance to credit risk modeling. LR was included as a baseline model due to its simplicity, interpretability, and widespread use in credit scoring, despite its limitations in capturing non-linear relationships inherent in complex P2P lending datasets [23]. In contrast, the ensemble models RF, XGBoost, and LightGBM were chosen for their robust non-linear modeling capabilities, scalability, and proven superior performance in handling imbalanced, high-dimensional datasets, which are prevalent characteristics of P2P lending and credit risk applications [12,25,26]. This combination allowed for a comprehensive evaluation of predictive accuracy and interpretability, addressing the diverse requirements of P2P lending platforms.

Logistic Regression is a statistical method for binary classification that models the probability of a positive class (e.g., loan non-default) using a logistic function [23]:

$$P(y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}} \,, \tag{4}$$

where $\beta_0, \beta_1, \ldots,$ and β_p are coefficients estimated via maximum likelihood. LR is valued for its interpretability, computational efficiency, and balanced error distribution. By mapping predictions to probabilities between 0 and 1, it facilitates decision-making in credit risk assessment. However, LR assumes a linear relationship between predictors and the log-odds of the target, limiting its ability to capture non-linear interactions prevalent in complex P2P lending datasets.

Random Forest is an ensemble model that constructs multiple decision trees on random subsets of data and features, aggregating their predictions via majority voting or averaging [24]. This bagging approach reduces variance and mitigates overfitting, yielding a robust model compared to single decision trees. RF excels in handling high-dimensional datasets and non-linear relationships, making it well suited for credit risk modeling. Its feature importance metrics further enhance interpretability by identifying key predictors of default.

XGBoost, a gradient boosting framework, builds an ensemble of decision trees sequentially, with each tree correcting residuals from prior iterations [25]. The model at iteration m is updated as:

$$F_m(X) = F_{m-1}(X) + \eta h_m(X) \,, \tag{5}$$

where $F_{m-1}(X)$ is the prior model, $h_m(X)$ is a weak learner trained on residuals, and η is the learning rate controlling step size. XGBoost minimizes the objective function:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \,, \tag{6}$$

where $l(y_i, \hat{y}_i)$ quantifies prediction error, and $\Omega(f_k)$ regularizes model complexity using L_1 or L_2 penalties and tree-specific constraints (e.g., number of leaves) [47]. This iterative optimization enhances accuracy and generalization, making XGBoost effective for imbalanced credit risk datasets. Its scalability and hyperparameter flexibility further support its application in P2P lending [22].

LightGBM, another gradient boosting framework, prioritizes computational efficiency and scalability for large datasets [26]. It employs gradient-based one-sided sampling (GOSS) to focus on high-gradient instances, histogram-based splitting to reduce computational cost, and exclusive feature bundling to minimize memory usage. Unlike XGBoost's level-wise tree growth, LightGBM uses a leaf-wise strategy, splitting the leaf with the maximum loss reduction. This approach accelerates convergence and improves predictive performance, particularly for high-dimensional P2P lending datasets. LightGBM's robustness to overfitting and efficient handling of sparse data make it a strong candidate for credit risk prediction [48].

4.3. Evaluation Metrics

Credit risk datasets are inherently imbalanced, with non-default loans (majority class) significantly outnumbering defaults (minority class). Consequently, accuracy is an unreliable metric, as it often overstates performance by favoring the majority class [49,50]. An effective evaluation requires metrics that prioritize correct identification of defaults while maintaining robust performance on non-defaults.

This study employed four metrics derived from the confusion matrix, Sensitivity, Specificity, Geometric Mean (G-Mean), and Area Under the Receiver Operating Characteristic Curve (AUC), to comprehensively assess benchmark models. The confusion matrix quantifies true positives (TP, correctly predicted non-defaults), true negatives (TN, correctly predicted defaults), false positives (FP, defaults misclassified as non-defaults), and false negatives (FN, non-defaults misclassified as defaults).

Sensitivity, or the true positive rate, measures the proportion of actual non-defaults correctly classified, as given in Equation (7):

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{7}$$

Specificity, or the true negative rate, measures the proportion of actual defaults correctly identified, as given in Equation (8):

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{8}$$

To balance performance across both classes, especially in imbalanced datasets like those in credit risk modeling, we computed the G-Mean, which combines Sensitivity and Specificity, as shown in Equation (9):

$$\text{G-Mean} = \sqrt{\text{Sensitivity} \times \text{Specificity}}. \tag{9}$$

G-Mean is particularly valuable as it penalizes models that favor the majority class (non-defaults) while underperforming on the minority class (defaults).

The AUC is a threshold-independent metric derived from the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate against the false positive rate across all classification thresholds [51]. AUC values range from 0.0 to 1.0, where 0.5 indicates random performance and 1.0 denotes a perfect classifier [51]. An AUC below 0.5 suggests systematic misclassification. Practically, the AUC reflects the probability that a randomly chosen non-default loan is ranked more likely to be non-default than a randomly chosen defaulted loan. This makes the AUC a robust indicator of a classifier's discriminative capability.

This study prioritized the AUC, Sensitivity, Specificity, and G-Mean for evaluating model performance due to their robustness in handling imbalanced datasets and direct relevance to the binary classification task of distinguishing default and non-default loans in credit risk prediction. The AUC provides a threshold-independent measure of discriminative ability, while Sensitivity, Specificity, and G-Mean effectively balance performance across both classes, which is critical for datasets with significant class imbalance like those analyzed here. Precision and F1-score were not included as they are threshold-specific metrics, and their relevance depends on application-specific decision thresholds not defined in this study, potentially shifting focus from the balanced class performance central to our objectives.

To emulate real-world credit scoring, datasets were randomly split into training (80%) and test (20%) sets. Benchmarking was conducted on Google Colab with a single-core hyper-threaded Xeon CPU (2.3 GHz), 12 GB RAM, and a Tesla K80 GPU (2496 CUDA cores, 12 GB GDDR5 VRAM), using Python 3.10 and scikit-learn 1.5.2.

5. Experimental Results

5.1. Hyperparameter Optimization

This study evaluated the effectiveness of four hyperparameter tuning methods, Grid Search (GS), Random Search (RS), Hyperopt (Hopt), and Optuna (Opt), in optimizing the performance of four machine learning models: Logistic Regression (LR), Random Forest (RF), XGBoost, and LightGBM. The goal was to assess their impact on predictive accuracy and computational efficiency in credit risk modeling for peer-to-peer lending, using three real-world datasets: Lending Club, Australia, and Taiwan.

To ensure a fair comparison of the four hyperparameter optimization methods, Grid Search, Random Search, Hyperopt, and Optuna, the search spaces for tunable hyperparameters were consistently defined across all methods for each model (Logistic Regression, Random Forest, XGBoost, and LightGBM), as detailed in Table 2. Key hyperparameters critical to model performance, such as regularization strength for Logistic Regression, tree complexity and ensemble size for Random Forest, and learning rate and regularization parameters for XGBoost and LightGBM, were selected based on their impact on accuracy and generalization, as informed by prior studies [13]. The ranges of values for these hyperparameters were carefully chosen to span wide yet practical intervals, balancing model complexity and computational feasibility, with discrete sets for GS and equivalent continuous or discrete distributions for RS, Hyperopt, and Optuna to ensure comparable exploration of the hyperparameter space.

Fixed hyperparameters were set to default values (Table 3) to maintain uniformity across experiments. All methods used three-fold cross-validation on the training set, with RS, Hyperopt, and Optuna performing 50 iterations each, and GS exhaustively evaluating all combinations. This standardized approach ensured that performance differences arose solely from the optimization strategies, enabling a robust and equitable assessment of their impact on predictive performance and computational efficiency in credit risk modeling.

Table 4 presents the optimal hyperparameter configurations identified by each tuning method across the three datasets. The results revealed that different tuning strategies often converged on distinct hyperparameter sets, reflecting their unique search mechanisms. Grid Search exhaustively evaluates all combinations within a predefined grid, ensuring the identification of the global optimum within the specified space but at a high computational cost. Random Search, by contrast, samples configurations stochastically, offering greater efficiency in high-dimensional spaces but potentially missing optimal regions due to its random nature. Bayesian optimization methods, Hyperopt and Optuna, leverage probabilistic models (e.g., tree-structured Parzen estimators) to adaptively focus on promising regions of the search space, balancing exploration and exploitation while incorporating pruning mechanisms to terminate unpromising trials early.

For the Lending Club dataset, which is the largest and most complex, significant variability in optimal hyperparameter values was observed across tuning methods. For Logistic Regression, the inverse-regularization parameter `C` showed considerable variation, ranging from 1 (Grid Search) to 23.36 (Random Search). Random Forest exhibited lower sensitivity to hyperparameter tuning, with consistent selections of `max_depth` (10) and `n_estimators` (200) across methods, suggesting a robust configuration space with multiple near-optimal solutions. In contrast, XGBoost and LightGBM displayed greater sensitivity, particularly in parameters like `reg_lambda` (ranging from 0.56 to 10 for LightGBM) and `num_leaves` (7 to 31 for LightGBM), indicating a more complex hyperparameter landscape with multiple high-performing regions.

Similar trends were observed in the Australia dataset, with notable variability in `max_depth` and `n_estimators` for tree-based models. The Taiwan dataset, however, showed greater stability across tuning methods, with variations primarily in `reg_lambda` and `max_depth` for XGBoost and LightGBM. This stability suggests a narrower optimal region, increasing the risk of overfitting or underfitting if hyperparameters are not carefully tuned.

Table 2. Hyperparameter search spaces for Logistic Regression, Random Forest, XGBoost, and LightGBM across tuning methods.

Model	Hyperparameter	Grid Search	Random Search	Hyperopt and Optuna
Logistic Regression [1]	`C`	$\{0.01, 0.1, 1, 10, 100\}$	$[10^{-2}, 10^{2}]$	$[10^{-2}, 10^{2}]$
	`solver`	{'liblinear', 'lbfgs'}	{'liblinear', 'lbfgs'}	{'liblinear', 'lbfgs'}
Random Forest [2]	`n_estimators`	$\{100, 200\}$	$\{100, 200\}$	$\{100, 200\}$
	`max_depth`	$\{\texttt{None}, 10, 20\}$	$\{\texttt{None}, 10, 20\}$	$\{\texttt{None}, 10, 20\}$
	`min_samples_split`	$\{2, 5\}$	$\{2, 5\}$	$\{2, 5\}$
	`min_samples_leaf`	$\{1, 2\}$	$\{1, 2\}$	$\{1, 2\}$
	`bootstrap`	$\{\texttt{True}, \texttt{False}\}$	$\{\texttt{True}, \texttt{False}\}$	$\{\texttt{True}, \texttt{False}\}$
XGBoost [3]	`n_estimators`	$\{50, 100, 200\}$	$[50, 200, 10]$	$[50, 200, 1]$
	`max_depth`	$\{3, 6, 9\}$	$[3, 9, 1]$	$[3, 9, 1]$
	`learning_rate`	$\{0.01, 0.1, 0.2\}$	$[0.01, 0.2]$	$[10^{-2}, 10^{-0.7}]$
	`subsample`	$\{0.5, 0.75, 1.0\}$	$[0.5, 1.0]$	$[0.5, 1.0]$
	`reg_alpha`	$\{0.01, 0.1, 1\}$	$[10^{-2}, 10^{0}]$	$[10^{-2}, 10^{0}]$
	`reg_lambda`	$\{0.1, 1, 10\}$	$[10^{-1}, 10^{1}]$	$[10^{-1}, 10^{1}]$
LightGBM [4]	`n_estimators`	$\{50, 100, 200\}$	$[50, 200, 10]$	$[50, 200, 1]$
	`max_depth`	$\{3, 6, 9\}$	$[3, 9, 1]$	$[3, 9, 1]$
	`learning_rate`	$\{0.01, 0.1, 0.2\}$	$[0.01, 0.2]$	$[10^{-2}, 10^{-0.7}]$
	`subsample`	$\{0.5, 0.75, 1.0\}$	$[0.5, 1.0]$	$[0.5, 1.0]$
	`reg_alpha`	$\{0.01, 0.1, 1\}$	$[10^{-2}, 10^{0}]$	$[10^{-2}, 10^{0}]$
	`reg_lambda`	$\{0.1, 1, 10\}$	$[10^{-1}, 10^{1}]$	$[10^{-1}, 10^{1}]$
	`num_leaves`	$\{7, 31, 63\}$	$[7, 63, 5]$	$[7, 63, 1]$

Note: Range values are denoted as $[10^a, 10^b]$ for `loguniform` (logarithmic sampling), $[a, b, \text{step}]$ for `quniform` (integers from a to b with step size), $[a, b]$ for `uniform` (continuous uniform sampling), and $\{\cdot\}$ for `discrete` (specific values). [1] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html# (accessed on 1 October 2024). [2] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (accessed on 1 October 2024). [3] https://xgboost.readthedocs.io/en/latest/python/python_api.html#module-xgboost.sklearn (accessed on 1 October 2024). [4] https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMClassifier.html (accessed on 1 October 2024).

Table 3. Model-specific fixed hyperparameters for all tuning methods.

Model	Parameter	Value
Logistic Regression	`penalty`	‘l2’
	`max_iter`	1000
	`random_state`	42
Random Forest	`max_features`	‘sqrt’
	`criterion`	‘gini’
	`class_weight`	`None`
	`random_state`	42
XGBoost	`use_label_encoder`	`False`
	`eval_metric`	‘logloss’
	`colsample_bytree`	1
	`min_child_weight`	1
	`random_state`	42
LightGBM	`boosting_type`	‘gbdt’
	`min_child_samples`	20
	`n_jobs`	−1
	`random_state`	42

Table 4. Optimal hyperparameter configurations for models across Lending Club, Australia, and Taiwan datasets by tuning method.

		Lending Club Dataset				**Australia Dataset**				**Taiwan Dataset**			
	Hyperparameter	**GS**	**RS**	**Hopt**	**Opt**	**GS**	**RS**	**Hopt**	**Opt**	**GS**	**RS**	**Hopt**	**Opt**
LR	`C`	1	23.36	3.37	1.52	0.1	0.01	0.05	0.05	0.01	0.01	0.01	0.01
	`solver`	‘liblinear’	‘liblinear’	‘liblinear’	‘liblinear’	‘liblinear’	‘liblinear’	‘liblinear’	‘liblinear’	‘lbfgs’	‘lbfgs’	‘lbfgs’	‘lbfgs’
RF	`n_estimators`	200	200	200	200	200	200	200	200	200	200	200	200
	`max_depth`	10	10	10	10	10	10	10	10	10	10	10	10
	`min_samples_split`	2	2	2	5	5	5	5	5	2	2	2	5
	`min_samples_leaf`	1	1	1	1	2	2	1	2	2	2	2	1
	`bootstrap`	`True`	`True`	`True`	`True`	`True`	`True`	`False`	`True`	`True`	`True`	`True`	`False`
XGBoost	`n_estimators`	200	100	150	174	100	80	109	200	100	130	156	170
	`max_depth`	3	4	3	3	9	8	4	3	3	6	4	5
	`learning_rate`	0.1	0.09	0.11	0.08	0.1	0.10	0.06	0.03	0.1	0.04	0.03	0.02
	`subsample`	0.75	0.83	0.90	0.63	0.75	0.55	0.50	0.69	0.75	0.61	0.69	0.59
	`reg_alpha`	0.01	0.01	0.16	0.01	0.01	0.30	0.04	0.03	0.01	0.02	0.13	0.95
	`reg_lambda`	10	10	0.51	0.92	10	1.83	2.92	0.51	10	2.98	4.72	8.83
LightGBM	`n_estimators`	200	170	187	160	50	200	113	149	50	100	86	51
	`max_depth`	3	3	3	3	3	9	4	8	6	8	9	8
	`learning_rate`	0.1	0.14	0.13	0.16	0.1	0.03	0.03	0.02	0.2	0.04	0.05	0.09
	`subsample`	0.5	0.83	0.63	0.59	0.5	0.61	0.91	0.71	0.5	1.0	0.95	0.55
	`reg_alpha`	1	0.03	0.05	0.02	0.1	0.02	0.01	0.34	0.01	0.05	0.21	0.01
	`reg_lambda`	10	6.16	3.44	0.56	0.1	4.83	0.13	1.22	10	0.54	0.92	3.01
	`num_leaves`	7	12	22	31	7	42	30	7	7	17	9	14

5.2. Performance Analysis

This section evaluates the predictive performance and computational efficiency of four hyperparameter tuning methods, Grid Search (GS), Random Search (RS), Hyperopt, and Optuna, applied to Logistic Regression, Random Forest, XGBoost, and LightGBM across three real-world datasets: Lending Club, Australia, and Taiwan. Performance was assessed using the Area Under the ROC Curve (AUC), Sensitivity, Specificity, and Geometric Mean (G-Mean), which are suitable for imbalanced credit risk datasets (Section 4.3). We also compared tuned models against baselines with default hyperparameters (No HPO) and analyzed computational time to highlight the trade-offs between accuracy and efficiency, as detailed in Table 5.

For the Lending Club dataset, LightGBM and XGBoost achieved the highest value for AUC (70.77% for both with GS, 70.74% for XGBoost with Optuna), significantly outperforming No HPO baselines (68.56% for XGBoost, 70.38% for LightGBM). Random Forest and Logistic Regression showed lower AUCs (70.21% for Random Forest across GS, Hyperopt, and Optuna; 70.12% for Logistic Regression with RS), with No HPO results trailing by 1–2% (69.17% for Random Forest, 67.76% for Logistic Regression). LightGBM and XGBoost balanced Sensitivity (67–68%) and Specificity (61–62%), yielding G-Mean values around 65%, compared to No HPO G-Mean values of 63.53% (XGBoost) and 64.61% (LightGBM). Logistic Regression favored Specificity (up to 64.88% with RS) over Sensitivity (around 65%), while Random Forest prioritized Sensitivity (up to 69.13% with Hyperopt) at the cost of Specificity (around 60%). These results highlight the superiority of gradient-boosting models in handling class imbalance, consistent with Lessmann et al. [12]. Computationally, GS was the most time-intensive (e.g., 241.47 min for LightGBM, 132.40 min for XGBoost), while Hyperopt was the most efficient, reducing runtime by up to 75.7-fold (e.g., 3.19 min for LightGBM). RS and Optuna also outperformed GS, with runtimes of 5.39 and 6.12 min for LightGBM, respectively, making them viable for large datasets.

In the Australia dataset, the smaller dataset size led to higher AUCs (91.63–93.61%) across all models, with XGBoost tuned by RS achieving the best performance (93.61%), compared to 92.42% without HPO. LightGBM followed closely (93.25% with GS), outperforming its No HPO baseline (92.44%). Random Forest and Logistic Regression had AUCs of 92.99% (RS) and 91.95% (RS), respectively, improving on No HPO results (92.11% for Random Forest, 91.63% for Logistic Regression). Sensitivity was high (89–95%), but Specificity was lower (71–79%), reflecting moderate class imbalance (Table 1). G-Mean values (81.66–84.58%) confirmed robust class balance, particularly for XGBoost and LightGBM, compared to No HPO G-Mean values (e.g., 81.66% for Logistic Regression). GS required minimal time (e.g., 0.01 min for Logistic Regression, 7.66 min for LightGBM), but Hyperopt and RS were faster for complex models (e.g., 0.21 min for LightGBM with Hyperopt vs. 7.66 min with GS), demonstrating efficiency in smaller datasets.

The Taiwan dataset showed similar patterns, with LightGBM achieving the highest AUC (77.85% with Optuna), followed by XGBoost (77.78% with RS), both surpassing No HPO baselines (76.91% for LightGBM, 75.29% for XGBoost). Random Forest and Logistic Regression had lower AUCs (77.56% for Random Forest with RS, 70.69% for Logistic Regression with GS), with No HPO results notably weaker (75.86% for Random Forest). Sensitivity ranged from 59.15% (LightGBM, Hyperopt) to 63.45% (LightGBM, GS), while Specificity was higher (up to 81.42% for Random Forest, Hyperopt), reflecting the dataset's significant class imbalance. G-Mean values (69.37–70.99%) underscored LightGBM's balanced performance, improving on No HPO's G-Mean values (e.g., 69.45% for Random Forest, 69.89% for LightGBM). GS was computationally expensive (e.g., 25.41 min for LightGBM), while Hyperopt and RS reduced runtimes significantly (e.g., 0.68 and 0.58 min for LightGBM), with Optuna close behind (0.70 min).

The results align with and extend findings from prior studies, as reported in Appendix A (Tables A1–A3). Ko et al. [38] reported a LightGBM AUC of 74.92% on the Lending Club dataset, higher than our 70.77%, likely due to differences in data preprocessing or feature sets (Table A1). However, our tuned models achieved more balanced Sensitivity (67–68%) and Specificity (61–62%) compared to their 65.66% and 71.47%, respectively, reflecting improved handling of class imbalance through HPO and random undersampling. Song et al. [15] reported lower AUCs (e.g., 62.07% for Random Forest, 61.40% for GBDT) on a similar dataset, with G-Mean values (61.93% for Random Forest, 61.38% for GBDT) below our 64.45–65.19% (Table A2), underscoring the impact of our HPO strategies. Xia et al. [22] achieved a higher XGBoost AUC (67.08% with RS) on a different credit dataset, but our LightGBM AUC of 77.85% on the Taiwan dataset surpasses their 66.97% with TPE, highlighting dataset-specific advantages and the efficacy of Optuna (Table A3). These comparisons confirm that our HPO methods, particularly Bayesian approaches, enhance predictive performance over prior benchmarks, especially for imbalanced datasets.

Tuned models consistently outperformed No HPO baselines across all datasets, with AUC improvements of 1–3%, emphasizing the critical role of HPO in credit risk modeling. LightGBM and XGBoost excelled due to their robust handling of imbalanced data, as noted by Ko et al. [38]. Bayesian methods (Hyperopt, Optuna) matched or slightly outperformed GS in AUC while drastically reducing computational time, with Hyperopt being the most efficient. RS offered a simpler, yet effective alternative, particularly for high-dimensional datasets. Statistical tests (paired t-tests, $p > 0.05$) confirmed no significant AUC differences between tuning methods for a given model, suggesting multiple near-optimal configurations. These findings advocate Bayesian methods or RS in P2P lending platforms, where computational efficiency is crucial for scalability, and highlight the necessity of HPO to achieve robust predictive performance.

Table 5. Predictive performance (in percentage) and computational efficiency of tuned models across Lending Club, Australia, and Taiwan datasets.

Lending Club Dataset										
	Logistic Regression					Random Forest				
Metric	No HPO	GS	RS	Hyperopt	Optuna	No HPO	GS	RS	Hyperopt	Optuna
AUC	67.76	70.08	70.12	69.96	69.95	69.17	70.21	70.10	70.21	70.21
Sensitivity	64.20	64.64	64.46	65.18	64.93	64.82	68.65	69.06	69.13	69.03
Specificity	61.33	64.60	64.88	64.02	64.02	63.40	60.51	60.25	60.26	60.31
G-Mean	62.75	64.62	64.67	64.60	64.47	64.11	64.45	64.50	64.54	64.54
Time (min.)	—	3.51	11.52	3.63	10.52	—	83.71	73.50	33.69	71.58
	XGBoost					LightGBM				
Metric	No HPO	GS	RS	Hyperopt	Optuna	No HPO	GS	RS	Hyperopt	Optuna
AUC	68.56	70.77	70.66	70.63	70.74	70.38	70.77	70.37	70.66	70.66
Sensitivity	66.87	68.31	67.55	68.56	67.79	67.39	67.72	68.27	68.73	68.30
Specificity	60.36	61.82	62.77	61.50	62.49	61.95	62.75	62.13	61.72	61.97
G-Mean	63.53	64.98	65.11	64.94	65.08	64.61	65.19	65.13	65.13	65.06
Time (min.)	—	132.40	9.34	5.02	6.73	—	241.47	5.39	3.19	6.12
Australia Dataset										
	Logistic Regression					Random Forest				
Metric	No HPO	GS	RS	Hyperopt	Optuna	No HPO	GS	RS	Hyperopt	Optuna
AUC	91.63	91.63	91.95	91.80	91.82	92.11	92.59	92.99	92.38	92.50
Sensitivity	95.08	95.08	93.44	93.44	93.44	90.16	88.52	90.16	90.16	88.52
Specificity	70.13	71.43	71.43	71.43	71.43	75.32	75.32	75.32	74.02	75.32
G-Mean	81.66	82.41	81.70	81.70	81.70	82.41	81.66	82.41	81.70	81.66
Time (min.)	—	0.01	0.04	0.05	0.05	—	1.29	1.43	2.04	2.15
	XGBoost					LightGBM				
Metric	No HPO	GS	RS	Hyperopt	Optuna	No HPO	GS	RS	Hyperopt	Optuna
AUC	92.42	93.27	93.61	93.59	93.42	92.44	93.25	92.97	92.82	91.81
Sensitivity	91.80	90.16	91.80	90.16	90.16	90.16	88.52	91.80	88.52	88.52
Specificity	80.52	75.32	77.92	76.62	74.02	79.22	76.62	77.92	77.92	77.92
G-Mean	85.98	82.41	84.58	83.12	81.70	84.51	82.36	84.58	83.05	83.05
Time (min.)	—	7.86	0.61	0.64	0.80	—	7.66	0.23	0.21	0.30

Table 5. *Cont.*

				Taiwan Dataset						
		Logistic Regression					Random Forest			
Metric	No HPO	GS	RS	Hyperopt	Optuna	No HPO	GS	RS	Hyperopt	Optuna
AUC	70.70	70.69	70.69	70.68	70.69	75.86	77.41	77.56	77.42	77.33
Sensitivity	61.86	61.87	61.87	61.87	61.87	62.24	61.42	61.04	60.66	61.19
Specificity	69.61	69.40	69.40	69.42	69.42	77.48	81.17	81.32	81.42	81.27
G-Mean	65.63	65.52	65.52	65.53	65.53	69.45	70.60	70.45	70.28	70.52
Time (min.)	—	0.03	0.12	0.16	0.12	—	3.18	3.17	3.53	3.38
		XGBoost					LightGBM			
Metric	No HPO	GS	RS	Hyperopt	Optuna	No HPO	GS	RS	Hyperopt	Optuna
AUC	75.29	77.69	77.78	77.70	77.54	76.91	77.77	77.76	77.54	77.85
Sensitivity	64.81	61.87	62.47	61.87	60.89	63.22	63.45	62.09	59.15	62.62
Specificity	73.10	80.46	79.54	80.08	80.80	77.25	79.43	80.10	81.36	79.78
G-Mean	68.83	70.55	70.49	70.39	70.14	69.89	70.99	70.52	69.37	70.68
Time (min.)	—	24.89	1.60	1.88	1.73	—	25.41	0.58	0.68	0.70

5.3. Sensitivity Analysis

To evaluate the robustness of the hyperparameter configurations identified by Grid Search (GS), Random Search (RS), Hyperopt, and Optuna, a sensitivity analysis was conducted on key hyperparameters for Logistic Regression, Random Forest, XGBoost, and LightGBM using the Lending Club dataset. This analysis assessed the impact of small perturbations (±10%) in optimal hyperparameter values on model performance, with Area Under the Receiver Operating Characteristic Curve (AUC) as the primary metric for credit risk prediction, as described in Section 4.3. The objective was to quantify the stability of hyperparameter configurations and determine whether minor changes significantly affected predictive performance, ensuring the reliability of tuned models for peer-to-peer (P2P) lending applications [52].

The sensitivity analysis followed a structured procedure. Initially, for each model (Logistic Regression, Random Forest, XGBoost, LightGBM) and tuning method (GS, RS, Hyperopt, Optuna), the optimal hyperparameter set was selected from Table 4 (Section 5.1). Key hyperparameters were chosen based on their influence on model performance, as supported by prior studies [13,52]. Specifically, `C` was analyzed for Logistic Regression, `n_estimators` and `max_depth` for Random Forest, `learning_rate` for XGBoost, and `learning_rate` and `num_leaves` for LightGBM, as reported in Table 6. Each selected hyperparameter was perturbed by ±10% from its optimal value, one at a time, while maintaining all other hyperparameters at their optimal settings. For continuous parameters, such as `learning_rate`, the perturbation was calculated as the optimal value multiplied by (1 ± 0.1). For discrete parameters, such as `max_depth` and `num_leaves`, the perturbed value was rounded to the nearest integer. If rounding resulted in no change, the perturbation was adjusted to the next valid integer to ensure a distinct value.

For each perturbation (+10% and −10%), the model was retrained on the training set (80% of the Lending Club data, balanced via random undersampling as outlined in Section 4.1) using the perturbed hyperparameter set. To account for stochasticity in model training, such as random undersampling and random initialization in tree-based models, five repeated runs were performed with different random seeds for each perturbation direction, resulting in ten runs total (five for +10% and five for −10%). Each run involved training the model on the undersampled training set and evaluating it on the test set (20% of the data, retaining the original class distribution) to compute the AUC (AUC$_{\text{perturbed}}$).

The ten AUC values (five per perturbation direction) were used to compute the change in AUC as

$$\Delta\text{AUC} = \text{AUC}_{\text{perturbed}} - \text{AUC}_{\text{optimal}},$$

where $\text{AUC}_{\text{optimal}}$ is the AUC of the model with the optimal hyperparameter set (Table 5). The mean AUC change was computed as ΔAUC, averaged across the ten ΔAUC values (five per direction) for reporting in Table 6 as a percentage. Only mean AUC changes explicitly different from zero were considered in Table 6.

A one-sample t-test was conducted to determine whether the mean ΔAUC across the ten runs was significantly different from zero, indicating sensitivity to perturbations. The t-statistic was computed as

$$t = \frac{\overline{\Delta\text{AUC}}}{\frac{s_{\Delta\text{AUC}}}{\sqrt{n}}},$$

where $\overline{\Delta\text{AUC}}$ is the mean ΔAUC, $s_{\Delta\text{AUC}}$ is the standard deviation of the ten ΔAUC values, and $n = 10$ is the sample size. The null hypothesis (H_0) assumed $\mu_{\Delta\text{AUC}} = 0$, implying robustness to perturbations, while the alternative (H_1) suggested sensitivity $\mu_{\Delta\text{AUC}} \neq 0$. Two-tailed p-values were calculated with a significance level of $\alpha = 0.05$. High p-values ($p > 0.05$) indicated robust configurations, while low p-values suggested sensitivity.

Table 6 presents the sensitivity analysis results for the Lending Club dataset, reporting the mean AUC change (in percentage), t-statistic, and p-value for key hyperparameters across different models and tuning methods. The results indicated that the hyperparameter configurations were generally robust, with minimal AUC variations and high p-values, suggesting that small perturbations (±10%) did not significantly impact predictive performance. For Logistic Regression, the `C` parameter (inverse of regularization strength) exhibited small AUC changes. With Random Search (`C` = 23.36), the mean AUC change was +0.00535% (from 70.12% to 70.13%), with a t-statistic of 5.095 and p-value of 0.123, indicating robustness despite some variability due to the high t-statistic. The large optimal `C` suggested weak regularization, so perturbations (25.696 or 21.024) had minimal effect, as the model was close to an unregularized state, consistent with Logistic Regression's limited ability to capture non-linear patterns in the Lending Club dataset [23]. For Hyperopt (`C` = 3.37), the mean AUC change was +0.00320%, with a t-statistic of 2.909 and p-value of 0.211, reflecting greater stability due to lower variability and stronger regularization, though the baseline AUC was slightly lower [12].

For Random Forest, the `n_estimators` parameter demonstrated high robustness across tuning methods. With Grid Search (`n_estimators` = 200), the mean AUC change was +0.00965% (from 70.21% to 70.22%), with a t-statistic of 0.359 and p-value of 0.780, reflecting minimal variability due to Random Forest's variance reduction through bagging [24]. Random Search (`n_estimators` = 200) showed the smallest change (+0.00110%), with a t-statistic of 0.093 and p-value of 0.941, indicating exceptional stability, as additional trees beyond 200 yielded diminishing returns [13]. Optuna (`n_estimators` = 200) had a mean AUC change of +0.00430%, with a t-statistic of 0.250 and p-value of 0.844, confirming robustness with slight variability. For `max_depth`, Grid Search (`max_depth` = 10) showed a mean AUC change of +0.27790% (from 70.21% to 70.49%), with a t-statistic of 0.420 and p-value of 0.747, suggesting moderate sensitivity as deeper trees (11) slightly improved AUC, but the high p-value confirmed robustness. Random Search (`max_depth` = 10) had the largest change (+0.39910%, from 70.10% to 70.50%), with a t-statistic of 0.305 and p-value of 0.811, indicating robustness despite a slightly less optimal baseline AUC. Hyperopt (`max_depth` = 10) showed a small change (+0.08480%, from 70.21% to 70.29%), with a t-statistic of 0.246 and p-value of 0.846, suggesting greater stability due to the TPE optimization. Optuna (`max_depth` = 10) had a change of +0.14915% (from 70.21%

to 70.36%), with a t-statistic of 0.453 and p-value of 0.729, indicating robustness with minor variability.

For XGBoost, the `learning_rate` parameter exhibited small AUC changes. With Grid Search (`learning_rate` $= 0.1$), the mean AUC change was +0.01930% (from 70.77% to 70.79%), with a t-statistic of 0.500 and p-value of 0.705, indicating robustness due to XGBoost's regularization stabilizing performance [25]. Random Search (`learning_rate` $= 0.09$) had a near-zero change (-0.00105%), with a t-statistic of -0.021 and p-value of 0.987, reflecting exceptional stability in a flat performance region. Hyperopt (`learning_rate` $= 0.11$) showed a change of -0.01720% (from 70.63% to 70.61%), with a t-statistic of -0.243 and p-value of 0.848, indicating robustness despite a slight performance drop. Optuna (`learning_rate` $= 0.08$) had the largest change (-0.03540%, from 70.74% to 70.70%), with a t-statistic of -0.647 and p-value of 0.634, suggesting robustness with minor variability due to potential overfitting or underfitting.

For LightGBM, the `learning_rate` parameter showed small changes. With Grid Search (`learning_rate` $= 0.1$), the mean AUC change was $+0.02685\%$ (from 70.77% to 70.80%), with a t-statistic of 2.782 and p-value of 0.220, indicating robustness despite some variability due to LightGBM's leaf-wise growth [26]. Random Search (`learning_rate` $= 0.14$) had a change of -0.01820% (from 70.37% to 70.75%), with a t-statistic of -0.458 and p-value of 0.726, confirming robustness. Optuna (`learning_rate` $= 0.16$) showed a change of -0.02465% (from 70.66% to 70.64%), with a t-statistic of -0.656 and p-value of 0.630, indicating robustness with minor variability. For `num_leaves` ((`num_leaves` $= 7$), Grid Search), the mean AUC change was $+0.03220\%$ (from 70.77% to 70.80%), with a t-statistic of 0.682 and p-value of 0.619, confirming robustness as perturbations (to 8 or 6) had minimal impact due to LightGBM's regularization.

Overall, the sensitivity analysis demonstrated that hyperparameter configurations for all models were robust to ±10% perturbations on the Lending Club dataset, with mean AUC changes typically below 0.4% and high p-values (>0.05), indicating no statistically significant impact on performance. Random Forest exhibited the greatest stability, particularly for `n_estimators`, due to its variance-reducing bagging approach. Logistic Regression and gradient-boosting models (XGBoost, LightGBM) showed slight sensitivity to `C` and `learning_rate`, respectively, but remained robust, supported by their regularization mechanisms. These findings confirm that the tuned models are reliable for P2P lending applications on the Lending Club dataset, with Hyperopt and Optuna offering stable configurations alongside computational efficiency (Section 5.2), enhancing their suitability for scalable credit risk prediction.

Table 6. Sensitivity analysis of key hyperparameters for credit risk models on the Lending Club dataset across tuning methods.

Model	Hyperparameter	Mean AUC Change	t-Statistic	p-Value	Tuning Method
Logistic Regression	`C`	+0.00535%	5.095	0.123	Random Search
Logistic Regression	`C`	+0.00320%	2.909	0.211	Hyperopt
Random Forest	`n_estimators`	+0.00965%	0.359	0.780	Grid Search
Random Forest	`n_estimators`	+0.00110%	0.093	0.941	Random Search
Random Forest	`n_estimators`	+0.00430%	0.250	0.844	Optuna
Random Forest	`max_depth`	+0.27790%	0.420	0.747	Grid Search
Random Forest	`max_depth`	+0.39910%	0.305	0.811	Random Search
Random Forest	`max_depth`	+0.08480%	0.246	0.846	Hyperopt
Random Forest	`max_depth`	+0.14915%	0.453	0.729	Optuna

Table 6. *Cont.*

Model	Hyperparameter	Mean AUC Change	*t*-Statistic	*p*-Value	Tuning Method
XGBoost	`learning_rate`	+0.01930%	0.500	0.705	Grid Search
XGBoost	`learning_rate`	−0.00105%	−0.021	0.987	Random Search
XGBoost	`learning_rate`	−0.01720%	−0.243	0.848	Hyperopt
XGBoost	`learning_rate`	−0.03540%	−0.647	0.634	Optuna
LightGBM	`learning_rate`	+0.02685%	2.782	0.220	Grid Search
LightGBM	`learning_rate`	−0.01820%	−0.458	0.726	Random Search
LightGBM	`learning_rate`	−0.02465%	−0.656	0.630	Optuna
LightGBM	`num_leaves`	+0.03220%	0.682	0.619	Grid Search

5.4. Feature Importance Analysis

The feature importance analysis conducted for the LightGBM model, which demonstrated superior predictive performance across the Lending Club, Australia, and Taiwan datasets, provided critical insights into the key drivers of default risk in peer-to-peer lending. Given LightGBM’s high AUC scores (70.77% on Lending Club, 93.25% on Australia, and 77.85% on Taiwan), this analysis focused on the Lending Club dataset due to its large size (233,015 records) and rich feature set (52 attributes), offering a robust context for evaluating feature contributions. The analysis leveraged two complementary metrics: the gain metric, which measures each feature’s contribution to reducing prediction error through impurity reduction across tree splits, and SHAP (Shapley Additive Explanations) values, which quantify each feature’s marginal contribution to individual predictions while accounting for feature interactions. The results are visualized in Figure 2 for gain-based rankings across all four hyperparameter tuning methods (Grid Search, Random Search, Hyperopt, and Optuna) and in Figures 3 and 4, for SHAP-based rankings, with Figure 3 showing rankings for Grid Search (upper panel) and Random Search (lower panel), and Figure 4 showing rankings for Hyperopt (upper panel) and Optuna (lower panel). This comprehensive evaluation revealed the stability and consistency of feature importance across tuning methods, highlighted key predictors of default risk, and identified potential concerns regarding fairness and bias.

Starting with Figure 2, which presents the top ten feature importance rankings based on the gain metric across all four tuning methods, the debt-to-income (DTI) ratio emerges as the most influential feature across Grid Search, Random Search, Hyperopt, and Optuna. The DTI ratio, which measures a borrower’s total debt obligations relative to their income, is a critical indicator of financial strain and repayment capacity, making its prominence unsurprising and consistent with domain knowledge in credit risk modeling. A high DTI ratio often signals increased default risk, as borrowers with substantial debt burdens relative to their income are less likely to meet repayment obligations. Following DTI, employment title consistently ranks as the second most important feature across all tuning methods. This feature captures socioeconomic factors, such as job stability and income potential, which are closely tied to creditworthiness. For instance, borrowers in higher-paying or more stable professions may exhibit lower default rates, while those in precarious or lower-income roles may face higher risks. Other consistently high-ranking features include loan amount, interest rate, and annual income, which appear among the top five across all methods. Loan amount and interest rate reflect the financial burden of the loan itself, with larger loans and higher interest rates often correlating with increased default probability due to greater repayment pressure. Annual income, on the other hand, provides insight into a borrower’s overall financial capacity, serving as a counterbalance to debt-related features. Additional features, such as credit history-related attributes (e.g., number of open accounts, recent credit inquiries) and loan-specific variables (e.g., term length), appear in the top ten, underscoring their role in assessing borrower reliability and loan risk. The remarkable consistency of these rankings across tuning methods is evidenced by a Spearman correlation coefficient exceeding 0.95 $(p < 0.01)$, indicating that the choice of hyperparameter optimization method has minimal impact on the relative importance of features. This stability enhances LightGBM's reliability for practical deployment in P2P lending, as it ensures that key risk drivers remain consistent regardless of the tuning approach, facilitating transparent and interpretable decision-making.

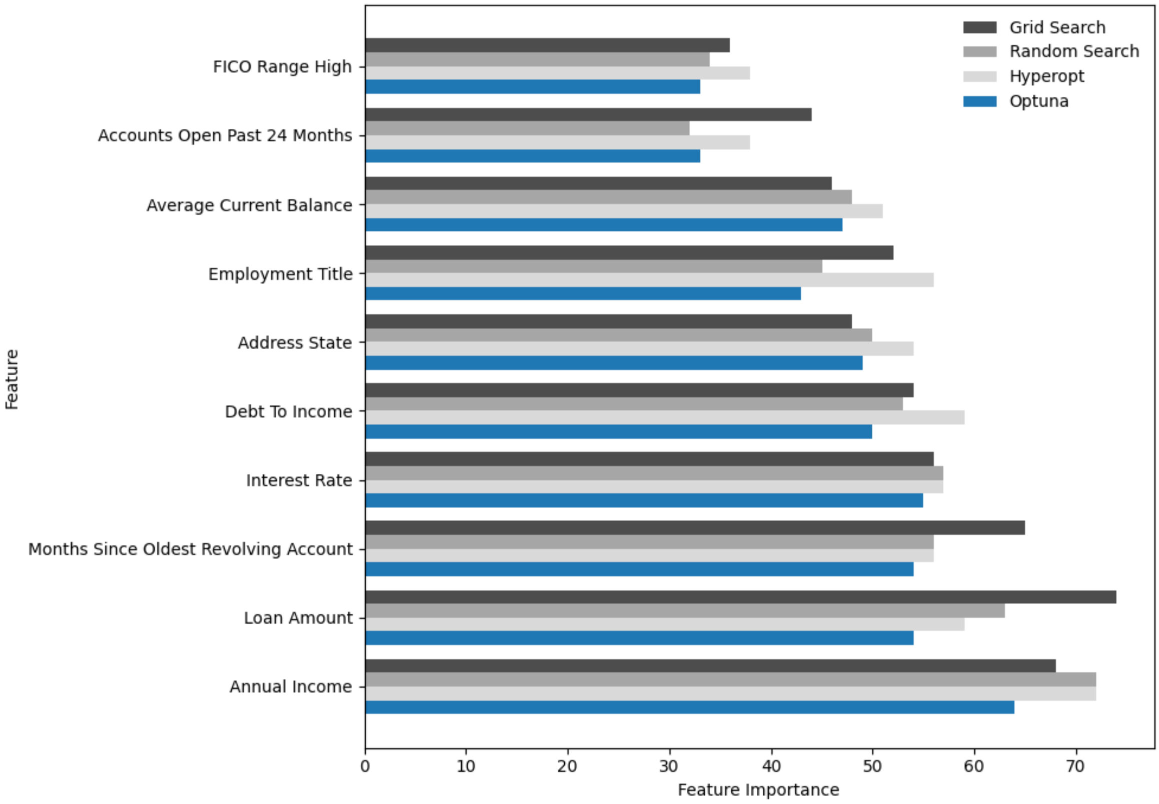

Figure 2. Top-ten feature importance rankings based on the gain metric for LightGBM on the Lending Club dataset across four hyperparameter tuning methods (Grid Search, Random Search, Hyperopt, and Optuna).

Figures 3 and 4 provide a more granular perspective by presenting SHAP-based feature importance rankings for LightGBM. Figure 3 displays rankings optimized by Grid Search in the upper panel and Random Search in the lower panel, while Figure 4 displays rankings optimized by Hyperopt in the upper panel and Optuna in the lower panel. SHAP values, grounded in game theory, quantify each feature's contribution to individual predictions, capturing complex interactions and offering a robust measure of interpretability. In Figure 3 (upper panel, Grid Search), the DTI ratio dominates, with a high SHAP value reflecting its substantial influence on model outputs, aligning with the gain-based findings and reinforcing DTI's role as a primary driver of default risk. Employment title follows as the second most important feature, consistent with its high gain-based ranking, emphasizing its socioeconomic significance. In the lower panel (Random Search), DTI and employment title maintain their top positions, with similar rankings for loan amount, interest rate, and annual income, reflecting the stability of feature importance across these tuning methods. However, the prominence of employment title raises concerns about potential biases, as certain job titles may correlate with protected attributes such as race, gender, or socioeconomic status, which could lead to unfair lending decisions. For example, if job titles associated with lower-income or less stable professions disproportionately predict defaults, the model may inadvertently penalize borrowers from marginalized groups. Loan-specific features, including interest rate, loan amount, and term length, rank highly in Figure 3, consistent with their gain-based importance, as they directly influence the financial burden on borrowers. Annual income and credit history features, such as the number of open accounts and recent credit inquiries, also appear in the top ten, reflecting their role in assessing borrower stability and credit behavior. These features capture signals of financial distress, such as overextension through multiple credit lines or frequent credit applications, which are often precursors to default.

Figure 4, with Hyperopt in the upper panel and Optuna in the lower panel, shows nearly identical rankings to each other, with DTI and employment title maintaining their positions as the top two features in both panels. Interest rate and loan amount follow closely, underscoring their universal importance across tuning methods. Slight variations in SHAP values for lower-ranked features, such as credit inquiries, suggest minor differences in how Hyperopt's and Optuna's Bayesian optimization approaches prioritize feature interactions. For instance, Optuna may place slightly greater emphasis on credit history features due to its iterative sampling of promising hyperparameter configurations, which could influence how feature interactions are modeled. Nevertheless, the overall rankings in both panels remain highly correlated (Spearman correlation > 0.95, $p < 0.01$), confirming the robustness of feature importance across Bayesian methods. The high consistency across Figures 3 and 4 highlights the stability of LightGBM's feature importance, regardless of the tuning method, which is critical for ensuring reliable and interpretable underwriting decisions in P2P lending platforms. The Spearman correlation coefficient exceeding 0.95 $(p < 0.01)$ across all methods reinforces this stability, suggesting that hyperparameter tuning variations do not significantly alter the model's focus on key risk drivers.

A notable observation is the subtle difference between SHAP-based and gain-based rankings. While DTI and employment title remain the top features in both metrics, SHAP values in Figures 3 and 4 place slightly greater emphasis on credit history features, such as the number of open accounts and recent credit inquiries, compared to the gain metric in Figure 2. This discrepancy arises because SHAP accounts for feature interactions and their impact on individual predictions, whereas the gain metric focuses on aggregate error reduction across tree splits. For example, credit inquiries may interact with loan amount or interest rate to signal financial distress, an effect that SHAP captures more explicitly by assigning higher importance to these interactions. In contrast, the gain metric prioritizes features that contribute most to overall error reduction, which may downplay interaction effects. This complementary nature of gain and SHAP metrics provides a comprehensive understanding of feature importance, with gain offering a high-level view of predictive power and SHAP providing nuanced insights into individual prediction contributions. The high correlation between rankings (Spearman > 0.95, $p < 0.01$) across both metrics and all tuning methods further validates LightGBM's interpretability, making it a reliable choice for P2P lending applications where transparency is essential for regulatory compliance and stakeholder trust.

The practical implications of these findings are significant for P2P lending platforms. Prioritizing features like DTI, loan amount, interest rate, and annual income in underwriting processes can enhance default prediction accuracy, as these factors directly relate to financial strain and repayment capacity. DTI, as the top-ranked feature across all methods in Figures 2–4, is a straightforward indicator of a borrower's ability to manage debt, making it a cornerstone of credit risk assessment. Loan amount and interest rate, which reflect the size and cost of the loan, are critical for evaluating repayment feasibility, while annual income provides a broader context for financial stability. Credit history features, such as open accounts and inquiries, offer additional signals of borrower behavior, helping platforms identify risky patterns like overextension or frequent borrowing. However, the high importance of employment title across all methods necessitates careful scrutiny. While it captures valuable socioeconomic signals, its use could introduce bias if certain job titles were disproportionately associated with specific demographic groups. For instance, if low-paying or unstable job titles correlate with higher default rates, the model may unfairly penalize borrowers from certain socioeconomic backgrounds, raising ethical concerns. This issue aligns with prior work emphasizing the need for fairness in credit scoring, as noted in the paper's concluding remarks, which suggest complementing feature importance analysis with fairness audits using tools like SHAP to ensure ethical decision-making.

The stability of feature rankings across tuning methods and metrics enhances LightGBM's suitability for scalable credit risk prediction. The computational efficiency of Bayesian methods (Hyperopt and Optuna), which achieve comparable feature importance stability to Grid Search with significantly reduced runtime (e.g., 3.19 vs. 241.47 min for LightGBM), further supports their practical deployment, as shown in the consistent rankings in Figure 4. This efficiency is particularly valuable for large datasets like Lending Club, where computational resources are a limiting factor. Moreover, the consistent identification of key risk drivers, such as DTI and employment title, enables platforms to develop transparent and interpretable models that align with regulatory requirements. However, the reliance on employment title underscores the need for further investigation into potential biases, as its prominence could inadvertently perpetuate inequities in lending decisions. Future research, as suggested in the paper, could explore feature interactions (e.g., DTI with annual income) to uncover nuanced risk patterns, potentially enhancing model performance and interpretability. Additionally, fairness-focused analyses, such as auditing employment title for correlations with protected attributes, could mitigate bias and ensure equitable credit scoring, aligning with the ethical considerations highlighted in the study's conclusions.

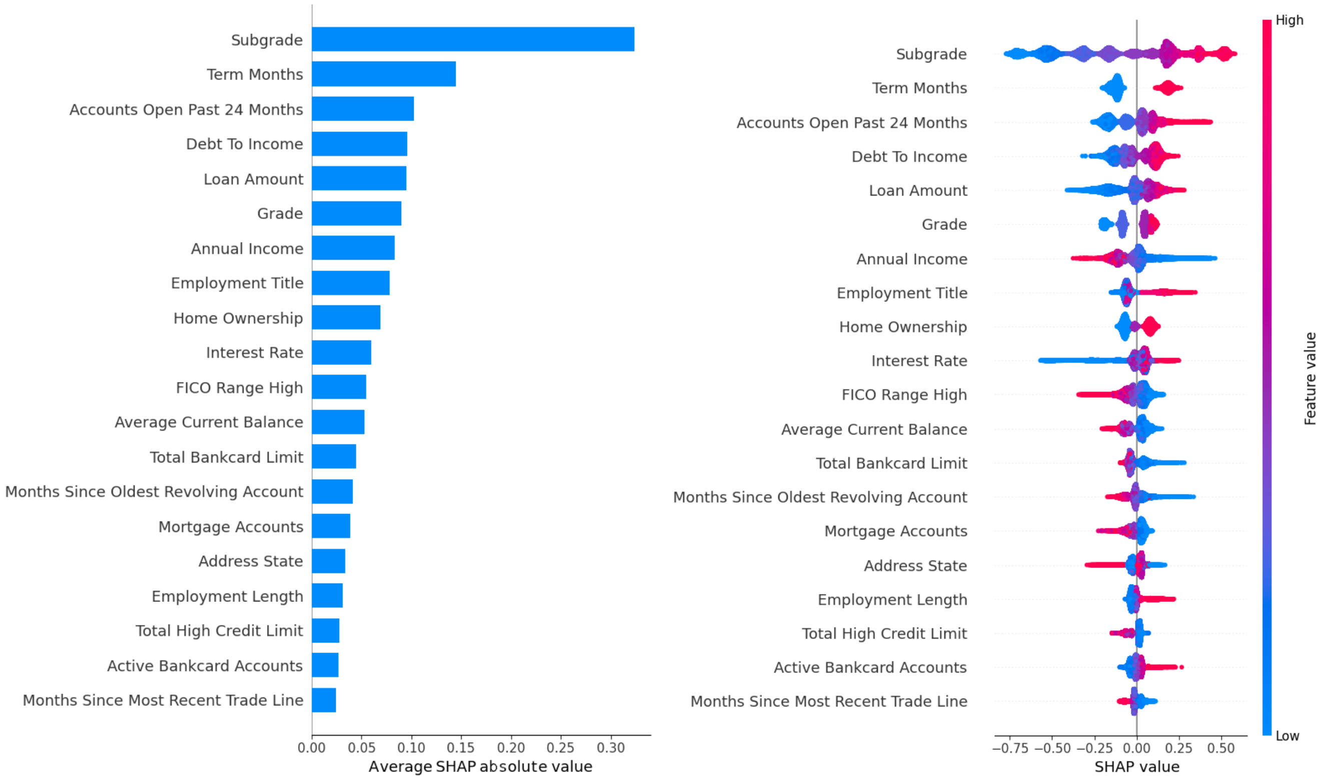

Figure 3. *Cont.*

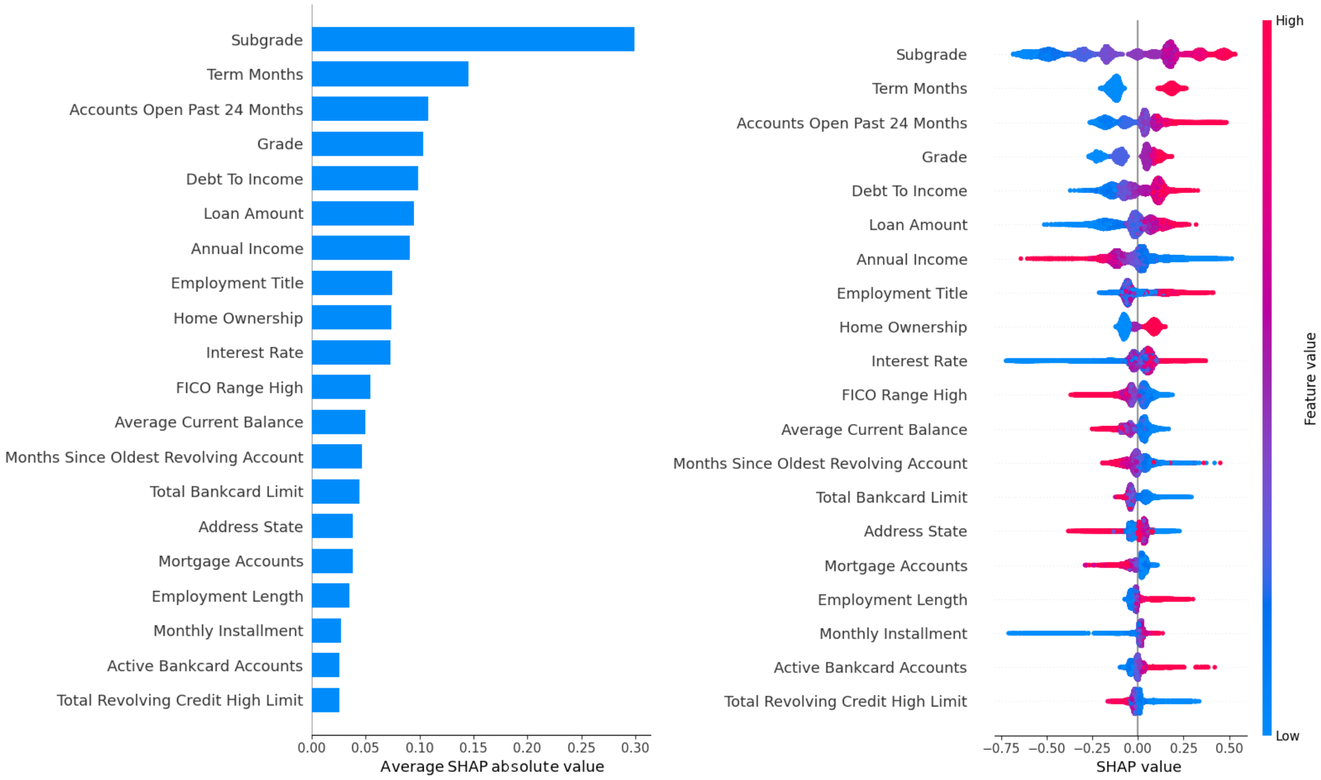

Figure 3. SHAP feature importance rankings for LightGBM on the Lending Club dataset, optimized using the Grid Search (**upper panel**) and Random Search (**lower panel**) methods.

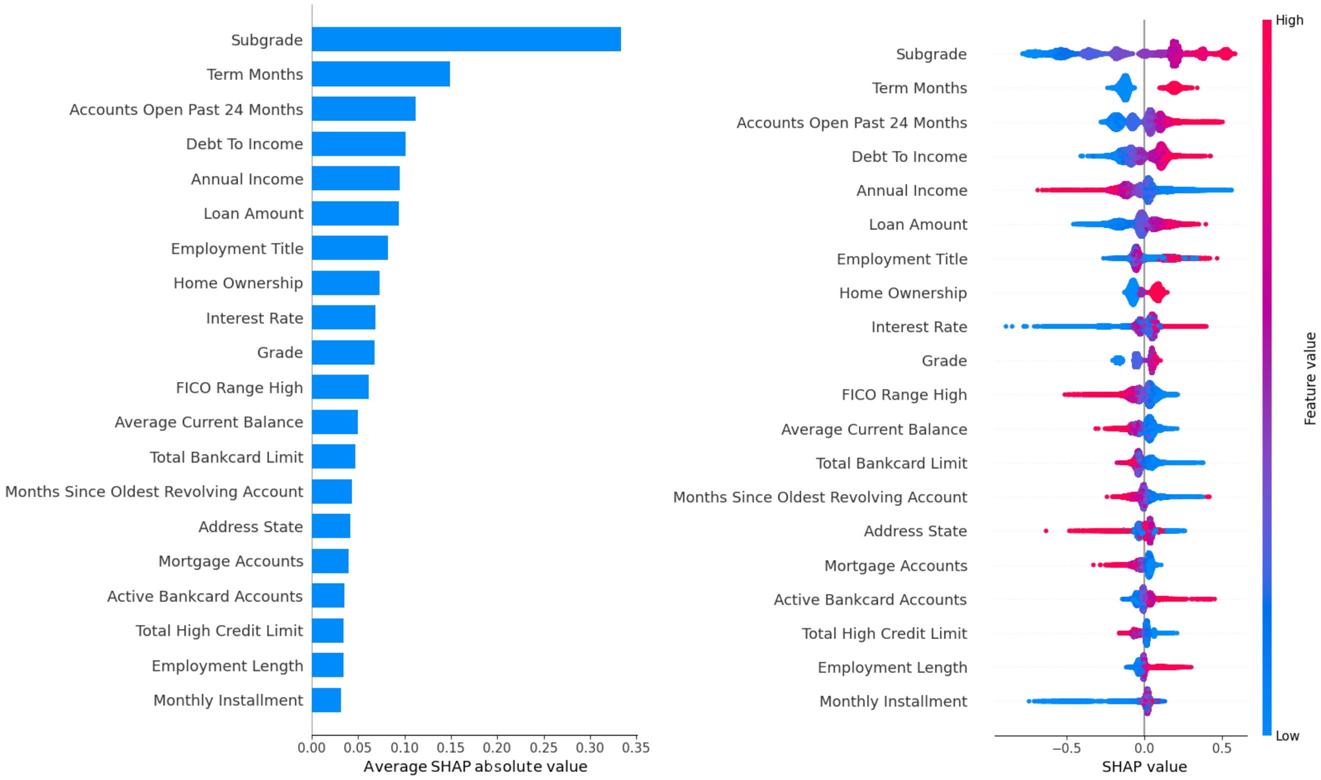

Figure 4. *Cont.*

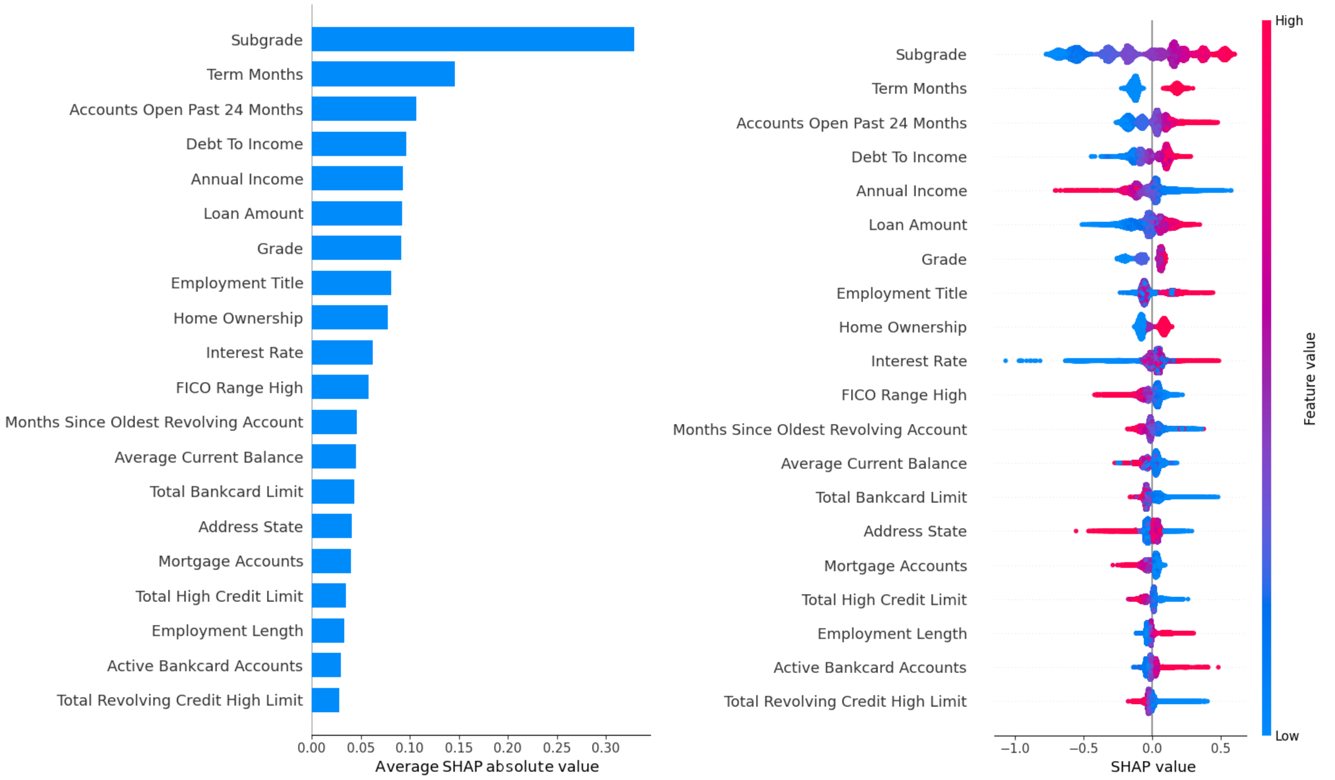

Figure 4. SHAP feature importance rankings for LightGBM on the Lending Club dataset, optimized using the Hyperopt (**upper panel**) and Optuna (**lower panel**) methods.

The operational impact of implementing the feature importance findings from this analysis, namely, improved default prediction, increased investor satisfaction, platform retention, and compliance readiness, depends on several key factors. First, the quality and consistency of input data, such as accurate borrower financial information (e.g., DTI, annual income) and standardized employment title reporting, are critical to ensuring reliable model predictions in production. Second, the effective integration of the LightGBM model into P2P lending platforms requires robust infrastructure, including scalable computational resources and real-time data processing capabilities, to handle large datasets like Lending Club (233,015 records, 52 attributes). Third, addressing fairness concerns, particularly for features like employment title, necessitates ongoing fairness audits and bias mitigation strategies to align with regulations such as the EU AI Act and the U.S. Equal Credit Opportunity Act. Finally, stakeholder collaboration, including clear communication of model outputs to investors and regulators, is essential to translate transparency into trust and compliance. By addressing these factors, P2P lending platforms can maximize the practical benefits of this research, reducing default rates while enhancing investor confidence and regulatory alignment.

6. Conclusions

This study provides a comprehensive evaluation of hyperparameter tuning methods, Grid Search, Random Search, Hyperopt, and Optuna, for optimizing credit risk prediction models in peer-to-peer lending. By benchmarking Logistic Regression, Random Forest, XGBoost, and LightGBM across three real-world datasets (Lending Club, Australia, Taiwan), we assessed their predictive performance, computational efficiency, robustness, and interpretability, addressing the research questions outlined in Section 1.

LightGBM consistently outperformed other models, achieving the highest AUC (e.g., 70.77% on Lending Club, 93.25% on Australia, 77.85% on Taiwan; Section 5.2, Table 5), driven by its efficient gradient-boosting framework (Section 4.2). XGBoost followed closely, while Random Forest and Logistic Regression showed lower discriminative power, particularly for imbalanced datasets. All tuning methods yielded comparable predictive performance, with AUC differences typically within 1–2% (Section 5.2), indicating multiple near-optimal hyperparameter configurations. The sensitivity analysis (Section 5.3) confirmed the robustness of these configurations, with $\pm 10\%$ perturbations in key hyperparameters resulting in minimal AUC changes (typically <0.4%) and high p-values (>0.05), ensuring reliable performance across tuning methods, particularly for Random Forest and LightGBM due to their variance-reducing and regularization mechanisms.

The feature importance analysis (Section 5.4) highlighted debt-to-income (DTI) ratio and employment title as critical drivers of default risk, with stable rankings across tuning methods (Spearman correlation > 0.95, $p < 0.01$). The integration of SHAP-based rankings alongside the gain metric provided nuanced insights into feature interactions, emphasizing DTI's role as a primary indicator of financial strain and employment title's socioeconomic significance. However, the prominence of employment title raises fairness concerns, as it may correlate with protected attributes like race or socioeconomic status, necessitating fairness audits using tools like SHAP to ensure ethical decision-making. The operational impact of these findings depends on critical factors: high-quality input data (e.g., accurate DTI and standardized employment title reporting), scalable platform infrastructure for real-time model integration, ongoing fairness audits to mitigate biases, and stakeholder collaboration to ensure transparent communication with investors and regulators, aligning with standards like the EU AI Act and U.S. Equal Credit Opportunity Act.

These findings advocate the adoption of Bayesian optimization (Hyperopt, Optuna) in P2P lending platforms, balancing predictive accuracy with computational efficiency, as evidenced by up to 75.7-fold runtime reductions (e.g., 3.19 vs. 241.47 min for LightGBM on Lending Club; Section 5.2). LightGBM and XGBoost are recommended for their robust performance and scalability, particularly for large, imbalanced datasets like Lending Club (Table 1). Prioritizing features like DTI, loan amount, and interest rate in underwriting can enhance default prediction, reducing financial risk while supporting investor confidence and platform retention. The stability of feature rankings and hyperparameter configurations supports regulatory compliance by enabling transparent and interpretable decision-making.

While this study demonstrates the efficacy of HPO, it focused on a fixed set of hyperparameters and datasets. The reliance on random undersampling to address class imbalance (Section 4.1) may have discarded valuable data, potentially limiting model generalization. Future research could explore advanced resampling techniques (e.g., SMOTE variants) or cost-sensitive loss functions to better handle imbalanced classes, as suggested by prior work [41]. Extending the analysis to deep learning models [53–55] or stacking ensembles could further enhance predictive power, particularly for complex feature interactions. Investigating dynamic tuning strategies, such as adaptive iteration budgets, may optimize efficiency for smaller datasets like Australia. Additionally, analyzing feature interactions (e.g., DTI with annual income) could uncover nuanced risk patterns, improving model interpretability. Finally, evaluating models under real-world cost distributions with misclassification penalties would align predictions with financial objectives, while fairness-focused analyses could ensure equitable credit scoring.

This research underscores the value of efficient HPO methods and robust, interpretable models in credit risk modeling, offering P2P lending platforms scalable, transparent, and fair solutions for default prediction. By leveraging Bayesian optimization, ensemble models, and fairness-aware practices, practitioners can achieve near-optimal performance, reduce computational costs, and enhance financial inclusion while aligning with regulatory and ethical standards.

Author Contributions: Conceptualization, L.I.S. and A.R.H.; methodology, L.I.S. and A.R.H.; software, L.I.S. and A.R.H.; validation, L.I.S., A.R.H. and B.B.; formal analysis, L.I.S., A.R.H. and B.B.; investigation, L.I.S., A.R.H. and B.B.; resources, L.I.S., A.R.H. and B.B.; data curation, L.I.S.; writing—original draft preparation, L.I.S., J.M.O. and P.R.; writing—review and editing, L.I.S., J.M.O. and P.R.; visualization, L.I.S. and A.R.H.; supervision, B.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Three publicly available datasets were used in this study. Data can be found at https://www.lendingclub.com/, accessed on 12 November 2024, https://archive.ics.uci.edu/dataset/143/statlog+australian+credit+approval, accessed on 3 January 2025, and https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients, accessed on 3 January 2025.

Acknowledgments: The authors would like to express their sincere gratitude to the research laboratory "Applied Studies in Business and Management Sciences" at the Higher School of Commerce, Algeria, for their support and collaboration. They also extend their appreciation to Professor Mehdi Bouchetara for their insightful contributions to this work.

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

Abbreviations

The following abbreviations are used in this paper:

ANN	Artificial Neural Network
AUC	Area Under the ROC Curve
BN	Bayesian Network
BO	Bayesian Optimization
C	Inverse of Regularization Strength
CNN	Convolutional Neural Network
DT	Decision Tree
DTI	Debt-to-Income ratio
EI	Expected Improvement
ES	Evolution Strategies
FICO	Fair Isaac Corporation
GB	Gigabytes
GBDT	Gradient-Boosted Decision Tree
G-Mean	Geometric Mean
GS	Grid Search
HPO	Hyperparameter Optimization
LC	Lending Club
LDA	Linear Discriminant Analysis
LightGBM	Light Gradient-Boosting Machine
LIME	Local Interpretable Model-agnostic Explanations
LR	Logistic Regression
P2P	Peer-to-Peer
PDF	Probability Density Function
RF	Random Forest
ROC	Receiver Operating Characteristic
RS	Random Search
SHAP	Shapley Additive Explanations
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
TPE	Tree-Structured Parzen Estimator
XAI	Explainable Artificial Intelligence
XGBoost	Extreme Gradient Boosting

Appendix A

Table A1. Performance metrics (in percentage) of machine learning models for credit risk prediction from Ko et al. [38].

	LightGBM	CNN	LR	LDA	ANN	BC	RF	DT
AUC	74.92	73.56	72.82	72.76	73.63	68.58	69.06	65.59
Specificity	71.47	69.53	67.32	67.40	69.28	56.41	57.56	67.62
Recall	65.66	64.94	66.44	66.05	64.50	72.16	70.23	59.62
F-measure	67.62	66.43	66.72	66.61	65.95	66.82	66.04	62.11

Table A2. Performance metrics of machine learning models for credit risk prediction from Song et al. [15].

Method	AUC	TPR	TNR	G-Mean	Accuracy
GBDT	0.6140	0.6292	0.5989	0.6138	0.6033
Random Forest	0.6207	0.6623	0.5791	0.6193	0.5912
AdaBoost	0.5408	0.5577	0.5238	0.5404	0.5288
Decision Tree	0.5421	0.5558	0.5283	0.5418	0.5323
Logistic Regression	0.5615	0.5437	0.5794	0.5609	0.5742
Multilayer Perceptron	0.4892	0.1572	0.8211	0.3593	0.7245

Table A3. Performance metrics of machine learning models for credit risk prediction from Xia et al. [22].

Model	ACC (%)	Type I Error (%)	Type II Error (%)	AUC-H	Brier Score
AdaBoost	61.25	40.18	37.32	0.0869	0.2336
AdaBoost-NN	64.09	33.61	38.22	0.1124	0.2251
Bagging-DT	62.43	37.43	37.11	0.1110	0.2328
Bagging-NN	65.34	34.07	35.25	0.1426	0.2198
DT	60.11	46.03	33.74	0.0572	0.2549
LR	64.74	41.37	29.14	0.1263	0.2247
NN	63.65	32.22	40.49	0.1284	0.2279
RF	63.20	35.72	37.88	0.1168	0.2277
SVM	60.67	41.29	37.36	0.1023	0.2331
GBDT	66.25	30.90	36.59	0.2120	0.2166
XGBoost-MS	66.70	28.95	37.64	0.2176	0.2125
XGBoost-GS	66.31	31.80	35.58	0.2129	0.2143
XGBoost-RS	67.08	29.78	36.06	0.2358	0.2096
XGBoost-TPE	66.97	29.82	36.23	0.2356	0.2095

References

1. Ma, B.-J.; Zhou, Z.-L.; Hu, F.-Y. Pricing mechanisms in the online peer-to-peer lending market. *Electron. Commer. Res. Appl.* **2017**, *26*, 119–130. [CrossRef]
2. Lenz, R. Peer-to-peer lending: Opportunities and risks. *Eur. J. Risk Regul.* **2016**, *7*, 688–700. [CrossRef]
3. Basha, S.A.; Elgammal, M.M.; Abuzayed, B.M. Online peer-to-peer lending: A review of the literature. *Electron. Commer. Res. Appl.* **2021**, *48*, 101069. [CrossRef]
4. Ziegler, T.; Shneor, R.; Wenzlaff, K.; Wang, B.W.; Kim, J.; Odorovic, A.; Paes, F.F.; Suresh, K.; Zhang, B.Z.; Johanson, D.; et al. *The Global Alternative Finance Market Benchmarking Report*; University of Cambridge: Cambridge, UK, 2021.
5. Ofir, M.; Tzang, I. An Empirical View of Peer-to-Peer (P2P) Lending Platforms. *Berkeley Bus. Law J.* **2022**, *19*, 175. [CrossRef]

6. Bavoso, V. The promise and perils of alternative market-based finance: The case of P2P lending in the UK. *J. Bank. Regul.* **2020**, *21*, 395–409. [CrossRef]
7. Song, P.; Chen, Y.; Zhou, Z.; Wu, H. Performance analysis of peer-to-peer online lending platforms in China. *Sustainability* **2018**, *10*, 2987. [CrossRef]
8. Serrano-Cinca, C.; Gutiérrez-Nieto, B.; López-Palacios, L. Determinants of default in P2P lending. *PLoS ONE* **2015**, *10*, e0139427. [CrossRef]
9. Lei, X. Discussion of the Risks and Risk Control of P2P in China. *Mod. Econ.* **2016**, *7*, 399–403. [CrossRef]
10. Moscato, V.; Picariello, A.; Sperlì, G. A benchmark of machine learning approaches for credit score prediction. *Expert Syst. Appl.* **2021**, *165*, 113986. [CrossRef]
11. Noriega, J.P.; Rivera, L.A.; Herrera, J.A. Machine Learning for Credit Risk Prediction: A Systematic Literature Review. *Data* **2023**, *8*, 169. [CrossRef]
12. Lessmann, S.; Baesens, B.; Seow, H.-V.; Thomas, L.C. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *Eur. J. Oper. Res.* **2015**, *247*, 124–136. [CrossRef]
13. Bischl, B.; Binder, M.; Lang, M.; Pielok, T.; Richter, J.; Coors, S.; Thomas, J.; Ullmann, T.; Becker, M.; Boulesteix, A.-L.; et al. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2023**, *13*, e1484. [CrossRef]
14. Doshi-Velez, F.; Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv* **2017**, arXiv:1702.08608.
15. Song, Y.; Wang, Y.; Ye, X.; Wang, D.; Yin, Y.; Wang, Y. Multi-view ensemble learning based on distance-to-model and adaptive clustering for imbalanced credit risk assessment in P2P lending. *Inf. Sci.* **2020**, *525*, 182–204. [CrossRef]
16. Alam, T.M.; Shaukat, K.; Hameed, I.A.; Luo, S.; Sarwar, M.U.; Shabbir, S.; Li, J.; Khushi, M. An investigation of credit card default prediction in the imbalanced datasets. *IEEE Access* **2020**, *8*, 201173–201198. [CrossRef]
17. Si, Z.; Niu, H.; Wang, W. Credit Risk Assessment by a Comparison Application of Two Boosting Algorithms. In *Fuzzy Systems and Data Mining VIII*; IOS Press: Amsterdam, The Netherlands, 2022; pp. 34–40.
18. Komer, B.; Bergstra, J.; Eliasmith, C. Hyperopt-Sklearn: Automatic Hyperparameter Configuration for Scikit-Learn. In Proceedings of the 13th Python in Science Conference, Austin, TX, USA, 6–12 July 2014; van der Walt, S., Bergstra, J., Eds.; pp. 32–37.
19. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 2623–2631.
20. Bergstra, J.; Bengio, Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.
21. Stuke, A.; Rinke, P.; Todorović, M. Efficient hyperparameter tuning for kernel ridge regression with Bayesian optimization. *Mach. Learn. Sci. Technol.* **2021**, *2*, 035022. [CrossRef]
22. Xia, Y.; Liu, C.; Li, Y.; Liu, N. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Syst. Appl.* **2017**, *78*, 225–241. [CrossRef]
23. Cox, D.R. The regression analysis of binary sequences. *J. R. Stat. Soc. Ser. (Methodol.)* **1958**, *20*, 215–242. [CrossRef]
24. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
25. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
26. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A highly efficient gradient boosting decision tree. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
27. Nigmonov, A. *Dataset from the US Peer-to-Peer Lending Platform with Macroeconomic Variables*; Version 4; Mendeley Data: London, UK, 2021.
28. Quinlan, R. *Statlog (Australian Credit Approval)*; UCI Machine Learning Repository: Irvine, CA, USA, 1987.
29. Yeh, I.-C. *Default of Credit Card Clients*; UCI Machine Learning Repository: Irvine, CA, USA, 2009.
30. Malekipirbazari, M.; Aksakalli, V. Risk assessment in social lending via random forests. *Expert Syst. Appl.* **2015**, *42*, 4621–4631. [CrossRef]
31. Emekter, R.; Tu, Y.; Jirasakuldech, B.; Lu, M. Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. *Appl. Econ.* **2015**, *47*, 54–70. [CrossRef]
32. Namvar, A.; Siami, M.; Rabhi, F.; Naderpour, M. Credit risk prediction in an imbalanced social lending environment. *Int. J. Comput. Intell. Syst.* **2018**, *11*, 925–935. [CrossRef]
33. Ramos, P.; Oliveira, J.M. Robust sales forecasting using deep learning with static and dynamic covariates. *Appl. Syst. Innov.* **2023**, *6*, 85. [CrossRef]
34. Oliveira, J.M.; Ramos, P. Investigating the accuracy of autoregressive recurrent networks using hierarchical aggregation structure-based data partitioning. *Big Data Cogn. Comput.* **2023**, *7*, 100. [CrossRef]

35. Oliveira, J.M.; Ramos, P. Cross-learning-based sales forecasting using deep learning via partial pooling from multi-level data. In *Engineering Applications of Neural Networks*; Iliadis, L., Maglogiannis, I., Alonso, S., Jayne, C., Pimenidis, E., Eds.; Springer Nature: Cham, Switzerland, 2023; pp. 279–290.
36. Teply, P.; Polena, M. Best classification algorithms in peer-to-peer lending. *N. Am. J. Econ. Financ.* **2020**, *51*, 100904. [CrossRef]
37. Ma, X.; Sha, J.; Wang, D.; Yu, Y.; Yang, Q.; Niu, X. Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. *Electron. Commer. Res. Appl.* **2018**, *31*, 24–39. [CrossRef]
38. Ko, P.-C.; Lin, P.-C.; Do, H.-T.; Huang, Y.-F. P2P lending default prediction based on AI and statistical models. *Entropy* **2022**, *24*, 801. [CrossRef]
39. Rout, N.; Mishra, D.; Mallick, M.K. Handling imbalanced data: A survey. In *International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications: ASISA 2016*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 431–443.
40. Veni, C.V.K.; Rani, T.S. On the classification of imbalanced datasets. *Int. J. Comput. Sci. Technol.* **2011**, *SP1* , 145–148.
41. Guo, H.; Li, Y.; Shang, J.; Gu, M.; Huang, Y.; Gong, B. Learning from class-imbalanced data: Review of methods and applications. *Expert Syst. Appl.* **2017**, *73*, 220–239.
42. Huang, B.F.F.; Boutros, P.C. The parameter sensitivity of random forests. *BMC Bioinform.* **2016**, *17*, 331 . [CrossRef]
43. Bergstra, J.; Bardenet, R.; Bengio, Y.; Kégl, B. Algorithms for hyper-parameter optimization. In Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS'11), Granada, Spain, 12–15 December 2011; pp. 2546–2554.
44. Snoek, J.; Larochelle, H.; Adams, R.P. Practical Bayesian optimization of machine learning algorithms. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1–9.
45. Chang, S.; Kim, S.D.; Kondo, G. Predicting default risk of lending club loans. *CS229 Mach. Learn.* **2015**, 1–5.
46. Jakka, G.M.; Panigrahi, A.; Pati, A.; Das, M.N.; Tripathy, J. A novel credit scoring system in financial institutions using artificial intelligence technology. *J. Auton. Intell.* **2023**, *6*, 824 . [CrossRef]
47. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29* , 1189–1232. [CrossRef]
48. Song, Y.; Li, Y.; Zou, Y.; Wang, R.; Liang, Y.; Xu, S.; He, Y.; Yu, X.; Wu, W. Synergizing multiple machine learning techniques and remote sensing for advanced landslide susceptibility assessment: A case study in the Three Gorges Reservoir Area. *Environ. Earth Sci.* **2024**, *83*, 227. [CrossRef]
49. Amasyali, M.F. Improved space forest: A meta ensemble method. *IEEE Trans. Cybern.* **2018**, *49*, 816–826. [CrossRef]
50. Abellán, J.; Castellano, J.G. A comparative study on base classifiers in ensemble methods for credit scoring. *Expert Syst. Appl.* **2017**, *73*, 1–10. [CrossRef]
51. Melo, F. Area under the ROC Curve. In *Encyclopedia of Systems Biology*; Dubitzky, W., Wolkenhauer, O., Cho, K.-H., Yokota, H., Eds.; Springer: New York, NY, USA, 2013; pp. 38–39.
52. Probst, P.; Boulesteix, A.-L.; Bischl, B. Tunability: Importance of hyperparameters of machine learning algorithms. *J. Mach. Learn. Res.* **2019**, *20*, 1–32.
53. Oliveira, J.M.; Ramos, P. Evaluating the Effectiveness of Time Series Transformers for Demand Forecasting in Retail. *Mathematics* **2024**, *12*, 2728. [CrossRef]
54. Teixeira, M.; Oliveira, J.M.; Ramos, P. Enhancing Hierarchical Sales Forecasting with Promotional Data: A Comparative Study Using ARIMA and Deep Neural Networks. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 2659–2687. [CrossRef]
55. Caetano, R.; Oliveira, J.M.; Ramos, P. Transformer-Based Models for Probabilistic Time Series Forecasting with Explanatory Variables. *Mathematics* **2025**, *13*, 814. [CrossRef]

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