J. Math. Math. Sci. (2025) 105–129

Balancing accuracy and interpretability in credit risk modeling: Evidence from peer-to-peer lending

Dinh Thuy Tien (Hanoi, Vietnam)

Nguyen Thi Tra My (Hanoi, Vietnam)

(Received Apr. 5, 2025; accepted May 22, 2025)

Abstract. Accurate credit risk assessment is crucial for the stability and growth of peer-to-peer (P2P) lending platforms. This study investigates the effectiveness of machine learning models in predicting loan defaults using historical Lending Club data. We evaluate logistic regression, decision tree, and random forest, employing feature engineering techniques like one-hot and weight of evidence encoding. Model performance is assessed using K-fold cross-validation and metrics such as accuracy and AUC. To enhance model interpretability, we utilize explainable AI techniques like LIME and SHAP, enabling lenders and borrowers to understand the factors driving loan defaults. Our findings demonstrate that while complex models offer higher predictive accuracy, simpler models like logistic regression with WoE encoding provide a suitable balance between accuracy and interpretability, fostering trust and responsible lending within the P2P lending ecosystem.

1. Introduction

The emergence of financial technology (Fintech) is widely recognized as one of the most significant innovations in the financial sector, reshaping the delivery and consumption of financial services at an unprecedented pace [1]. Broadly, Fintech solutions fall into two categories: those designed for individual consumers such as personal financial management, investment, and lending

Key words and phrases: Peer-to-Peer Lending, Credit Risk Assessment, Logistic Regression, Random Forest, Weight of Evidence Encodings, Explainable AI, LIME, SHAP, Model Interpretability, Lending Club

and those developed for financial institutions, offering solutions like customer identification and credit scoring (ISB, 2025).

Among consumer-facing Fintech innovations, Peer-to-Peer (P2P) lending has gained considerable traction as a disruptive force in traditional lending markets. In this model, individual lenders provide unsecured loans directly to borrowers through online platforms, bypassing conventional financial intermediaries. P2P lending platforms operate similarly to marketplace disruptors like Uber and Grab, facilitating connections between lenders and borrowers at scale.

Since the launch of the first P2P platform, Zopa, in the UK in 2005, the industry has grown rapidly across the globe [2]. The Credit Committee on the Global Financial System and the Financial Stability Board have identified China, the United States, and the United Kingdom as the largest P2P markets, with outstanding P2P credit in China reaching USD 99.7 billion, followed by USD 34.3 billion in the U.S. and USD 4.1 billion in the UK.

Several factors have contributed to this surge. The aftermath of the 2008 global financial crisis led to stricter regulatory capital requirements and constrained lending by traditional banks [3]. Simultaneously, declining interest rates on savings accounts pushed investors to seek higher-yield alternatives. P2P lending has also appealed to underserved segments, such as small and medium enterprises (SMEs) and rural populations who are often excluded from formal banking channels. In addition, the proliferation of mobile technology and the internet has created an infrastructure that supports the scalability of digital lending platforms.

Despite its rapid growth, the P2P lending industry faces significant challenges, particularly in ensuring robust credit risk assessment to maintain platform stability and protect stakeholders. The motivation for this study stems from the critical need to develop accurate and interpretable credit risk models that can effectively predict loan defaults while meeting regulatory and ethical standards. In P2P lending, where individual investors bear the financial risk of borrower defaults, inaccurate credit assessments can lead to substantial losses, erode investor confidence, and undermine the sustainability of lending platforms. Moreover, the lack of transparency in credit decision-making processes can exacerbate issues of trust and fairness, particularly for borrowers who may be denied loans without clear justifications. These challenges are compounded by the increasing complexity of modern credit datasets, which include diverse borrower attributes and require sophisticated analytical approaches to uncover meaningful patterns.

Traditional credit scoring methods, such as logistic regression and statistical scoring models, have long been valued for their simplicity and interpretability, making them suitable for regulatory compliance and stakeholder communication. However, these methods often struggle to capture the non-linear relationships and high-dimensional interactions present in large-scale P2P lending datasets. In contrast, machine learning (ML) techniques, such as decision trees and random forests, excel at modeling complex patterns and improving predictive accuracy, but their "black-box" nature poses challenges in regulated financial environments where explainability is paramount. For instance, regulations like the Fair Credit Reporting Act (1970) in the United States and the General Data Protection Regulation (GDPR) (2018) in the European Union mandate that lenders provide clear explanations for credit decisions, a requirement that complex ML models struggle to meet without additional interpretability tools. The integration of machine learning and mathematical models offers a promising solution to address these dual objectives of accuracy and interpretability. By combining the predictive power of advanced ML algorithms with the transparency of traditional mathematical frameworks, such as logistic regression with Weight of Evidence (WoE) encoding, this study aims to develop credit risk models that are both highly accurate and easily interpretable. Furthermore, the incorporation of explainable AI (XAI) techniques, such as Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP), allows us to bridge the interpretability gap for complex ML models, enabling lenders to understand and communicate the factors driving credit decisions.

This study makes several key contributions to the field of credit risk assessment in P2P lending. First, we provide a comprehensive evaluation of machine learning models (logistic regression, decision trees, and random forests) under two preprocessing strategies-Weight of Evidence (WoE) encoding and one-hot encoding with min-max scaling-using a real-world dataset from Lending Club. This analysis identifies optimal modeling approaches that balance predictive accuracy with interpretability, offering practical guidance for P2P lending platforms. Second, we demonstrate the effectiveness of integrating traditional mathematical models with advanced ML techniques, showing that logistic regression with WoE encoding achieves a desirable trade-off between performance and transparency, while random forests enhanced with XAI tools deliver superior accuracy with actionable explanations. Third, we apply LIME and SHAP to interpret complex ML models, providing both local and global insights into the factors driving credit decisions, which supports regulatory compliance and enhances stakeholder trust. Finally, our findings contribute to the responsible deployment of ML in P2P lending by proposing a framework that aligns with regulatory expectations, promotes fair lending practices, and fosters transparency in credit scoring, thereby supporting the sustainable growth of the P2P lending ecosystem.

However, the growth of P2P lending raises concerns around regulatory oversight, consumer protection, and systemic risk. Countries like China and the U.S. have implemented regulatory safeguards-such as prohibiting platforms from holding client funds or disbursing loans directly in order to reduce risks such as fraud, mismanagement, and financial exclusion. Legal risks, including Ponzi-like schemes and investor discrimination, remain ongoing challenges, and concerns about transparency, liquidity, and platform viability persist [4] [5]. These developments underscore the need for robust credit risk models that can assess borrower quality and support the responsible expansion of P2P lending.

Given the scale and risks involved in digital lending ecosystems, credit default prediction has become an essential task in modern financial services. Building reliable models that can identify borrowers likely to default is crucial for mitigating financial loss, maintaining investor trust, and complying with regulatory standards. This paper focuses on building and interpreting machine learning models for predicting credit default risk using a real-world dataset from Lending Club, one of the largest P2P lending platforms in the United States.

2. Literature Review

Credit risk modeling has long been central to financial decision-making, with early work relying on traditional statistical techniques [6]. As credit markets expanded and digital lending platforms emerged, machine learning (ML) approaches have increasingly been used to enhance prediction accuracy and scale model deployment. Among these, logistic regression, decision trees, and random forests remain some of the most widely applied methods in consumer credit risk, including in peer-to-peer (P2P) lending [7] [8], mortgage default [9] [10], and credit card repayment modeling [11].

The rapid growth of digital lending platforms, particularly peer-to-peer (P2P) lending, has spurred research into advanced machine learning (ML) techniques for credit risk assessment, with a shared goal of improving predictive accuracy while addressing interpretability challenges in regulated financial environments. Ma et al. (2018) [7] made a significant contribution by addressing the critical problem of predicting loan defaults in P2P lending networks, aiming to enhance risk assessment for online platforms. They employed gradient boosting algorithms, specifically LightGBM and XGBoost, on a high-dimensional dataset from a Chinese P2P lending platform, incorporating borrower demographics, credit history, and transaction records, with preprocessing to handle missing values and outliers. Their results demonstrated superior performance, with LightGBM achieving an AUC of 0.85, underscoring the power of gradient boosting to capture complex, non-linear patterns in P2P lending data and setting a benchmark for predictive modeling in this domain, which directly informs our study's exploration of ML in P2P credit scoring.

Similarly, Duan (2019) [8] tackled credit default prediction across various lending contexts, focusing on modeling financial system risk. The study utilized deep neural networks (DNNs) on a proprietary consumer loan dataset, including features like credit scores, debt-to-income ratios, and payment histories. The DNNs achieved an accuracy of 0.82, surpassing logistic regression, but their complexity highlighted the need for interpretability in regulated settings, a challenge that aligns with our emphasis on explainable models for P2P lending.

In the mortgage sector, Sirignano et al. (2016) [9] addressed the problem of predicting default risk, a priority following the 2008 financial crisis. They applied recurrent neural networks (RNNs) to a large U.S. mortgage loan dataset, incorporating time-series data on payment behavior and macroeconomic indicators. Their model achieved an AUC of 0.78, demonstrating the ability of deep learning to model temporal dependencies, though the lack of interpretability posed limitations, reinforcing the need for explainable AI (XAI) methods in our work.

Kvamme et al. (2018) [10] also focused on mortgage default prediction, aiming to improve risk assessment for financial institutions. Using convolutional neural networks (CNNs) on a Norwegian mortgage dataset with borrower financials and loan characteristics, they achieved a recall of 0.71, outperforming traditional models. However, the black-box nature of CNNs underscored the importance of interpretability, a concern central to our study's use of XAI techniques like LIME and SHAP.

Butaru et al. (2016) [11] investigated credit card repayment risk, seeking to identify delinquency drivers in consumer credit. They applied logistic regression and random forests to a large dataset from a U.S. credit card issuer, including transaction and payment data. The random forest model yielded an AUC of 0.80, outperforming logistic regression, but regulatory demands for transparency favored the interpretable logistic regression, a finding that shapes our model selection strategy for balancing accuracy and explainability in P2P lending. These studies collectively highlight the potential of advanced ML to enhance credit risk prediction across diverse lending contexts, with Ma et al. (2018) [7] providing a particularly relevant framework for P2P lending through their high-performing gradient boosting approach. However, the recurring challenge of model interpretability, especially for complex models, underscores the need for integrating predictive power with transparency, a core objective of our study.

Logistic regression continues to be popular in the financial industry due to its simplicity and transparency. The model's coefficients can be directly interpreted as indicators of a feature's effect on the likelihood of default, making it highly suitable in regulated environments. Decision trees also offer transparency through their rule-based structure but are prone to instability when faced with noisy or imbalanced data. To improve predictive performance, many studies and industry applications turn to random forests, which aggregate predictions from multiple decision trees trained on randomized subsets of the data and features [12]. Although random forests generally outperform simpler models, they are less transparent, making them difficult to interpret a key concern in finance.

This lack of interpretability presents serious challenges in regulated credit environments. In the United States, the Fair Credit Reporting Act (1970) requires lenders to disclose the main reasons behind a loan rejection. In the European Union, the General Data Protection Regulation (GDPR) (2018) provides individuals with a "right to explanation" for algorithmic decisions [13]. In Vietnam, the regulatory landscape for digital lending is still evolving, but the State Bank of Vietnam's Fintech Sandbox Draft Decree (2021) has emphasized that platforms must provide clear disclosure of loan terms and decision criteria. However, there are no standardized guidelines yet for how credit risk scores should be calculated, especially when ML models are involved. This creates growing pressure on lenders to ensure their models are not only accurate but also explainable. To address these challenges, researchers have increasingly focused on interpretable machine learning. Traditional models such as logistic regression and decision trees are naturally interpretable, but may lack the flexibility to capture complex relationships in the data. In contrast, ensemble methods like random forests offer improved performance, but are considered black-box models. To bridge this gap, post hoc interpretability methods have been developed—most notably, LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations). LIME approximates a complex model locally using a linear surrogate [14], while SHAP attributes a model's prediction to individual features using cooperative game theory principles [15]. These methods have been applied to credit risk modeling, including work on Lending Club data [16], [17], [18].

In this study, we focus specifically on the lender's perspective, recognizing their need for both high-performing models and clear justifications for credit decisions. Using loan-level data from Lending Club, we evaluate the performance of logistic regression (with L1 and L2 regularization) and tree-based models (decision trees and random forests) under two different preprocessing strategies: (i) weight of evidence (WoE) and (ii) one-hot encoding with min-max scaling. Our results identify two models of interest: a logistic regression model using WoE, which is inherently interpretable, and a random forest model trained on one-hot encoded, scaled data, which achieves high accuracy but requires additional explanation tools. Therefore, we analyze the logistic model using standard coefficient interpretation, and apply LIME and SHAP exclusively to the random forest model to uncover the drivers behind its predictions.

By balancing predictive power with interpretability tailored for lenders, this work contributes to the responsible deployment of machine learning in P2P credit scoring, helping loan providers meet regulatory expectations while making informed, transparent lending decisions.

3. Methodology

3.1. Reviewing data

For this study, we utilized historical loan data from Lending Club, a leading U.S. peer-to-peer (P2P) lending platform, covering loans issued in 2018. The dataset is publicly available through Kaggle, specifically the "Lending Club Loan Data" dataset, which includes loans from 2007 to 2018 [23], licensed under CC0 1.0 Universal (Public Domain Dedication). This dataset contains hundreds of features per loan, including key financial attributes such as loan amount, interest rate, monthly installment, and borrower-related variables like homeownership type, annual income, monthly FICO score, debt-to-income ratio, and the number of open credit lines. The data represents loans actually funded through the platform, not loan applications, ensuring that the analysis reflects real lending outcomes.

To provide clarity on the dataset's structure, Table 1 presents an example of the training dataset, showcasing a subset of five loan records with selected features and their corresponding labels. This example illustrates the types of variables used and the binary classification labels derived for modeling.

Table 1. Example of Training Dataset from Lending Club 2018 Data

loan_amnt	**annual_inc**	**fico_range_low**	**dti**	**home_ownership**	**loan_status**
10000	60000	700	15.2	RENT	Fully Paid
15000	45000	665	22.5	MORTGAGE	Charged Off
20000	80000	720	18.7	OWN	Fully Paid
8000	35000	680	25.3	RENT	Default
12000	55000	695	20.1	MORTGAGE	Fully Paid

Note: loan_amnt (loan amount in USD), annual_inc (annual income in USD), fico_range_low (lower bound of FICO score), dti (debt-to-income ratio in %), home_ownership (borrower's homeownership status), loan_status (loan outcome).

Loan status serves as the outcome variable and reflects the borrower's repayment behavior. A loan is marked as "Current" if it is being repaid on time, "Late" if payment is between 16 and 120 days overdue, and "Default" if the delay exceeds 121 days. If Lending Club determines that a loan will not be repaid, it is labeled as "Charged-Off." To streamline the classification task, we limited our analysis to loans that were either Fully Paid, Default, or Charged-Off. We categorized Fully Paid loans as creditworthy, and those labeled as Default or Charged-Off as non-creditworthy. After filtering, the dataset comprised 8,323 non-creditworthy records and 47,384 creditworthy ones.

Following the definition of the classification labels, we examined the features available in the dataset. These features fall into three broad categories: borrower characteristics (such as FICO score, employment status, and annual income), platform-driven decisions (such as loan grade and interest rate), and loan performance outcomes (such as total payment). Because our objective is to develop a model that can be applied in real-world settings, we prioritized features that would be available to an investor at the time of loan issuance. This approach ensures that the model's predictions are not only accurate but also practical and actionable.

In doing so, we addressed two major concerns: data leakage and the use of platform-derived variables. Data leakage arises when a model incorporates information that would not be accessible at the time a prediction is made, especially when such information is strongly correlated with the target variable. For example, the total payment feature is highly predictive of loan outcome-loans that default or are paid off early typically have lower total payments. While including this variable may boost model performance during training, it undermines the model's applicability in real-time investment decisions, where such information is unavailable in advance.

Another concern involves variables that are not direct borrower attributes but instead are generated by Lending Club's internal risk models. The loan grade variable is a clear example, and features such as interest rate and installment amount are closely tied to this grade. Since these variables reflect Lending Club's proprietary assessment mechanisms rather than fundamental borrower characteristics, we excluded them from our analysis to ensure the model remains independent of platform-specific decisions and can generalize to other lending contexts.

3.2. Prediction models

The fundamental objective of credit scoring is to assess the creditworthiness of individual applicants, which is essentially a binary classification problem. A creditworthy applicant is expected to fulfill their financial obligations, whereas a non-creditworthy applicant is likely to default. Accordingly, we frame the consumer credit risk prediction task as estimating the probability of default for a borrower, based on a set of observed features.

Let $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ik})$ denote the feature vector for borrower i, which captures information such as credit history, income, debt-to-income ratio, and prior delinquencies. The target variable $y_i \in \{0, 1\}$ represents whether the borrower defaulted, where $y_i = 1$ indicates default and $y_i = 0$ indicates no default. The modeling tasks is then to estimate:

$$\hat{y}_i = \Pr(y_i = 1 \mid \mathbf{x}_i). \tag{1}$$

We apply two categories of machine learning models to this task: linear models-specifically L1- and L2-regularized logistic regression and tree-based models including decision trees and random forests.

3.2.1. Logistic Regression

Logistic regression is a statistical classification method that models the probability of a binary outcome as a function of a linear combination of input features. It was formally introduced in the context of binary response modeling by Cox (1958). The method models the log-odds of the probability of default as follows:

$$\log\left(\frac{1 - \Pr(y_i = 1)}{\Pr(y_i = 1)}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}. \tag{2}$$

The parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)$ are estimated by maximizing the likelihood function. To improve generalization and prevent overfitting, regularization is commonly applied.

L1-regularized logistic regression, also known as Lasso logistic regression, introduces a penalty term proportional to the absolute value of the coefficients:

$$\mathcal{L}_L = -\sum_{i=1}^{n} \left[y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \right] + \lambda \sum_{j=1}^{k} |\beta_j|. \tag{3}$$

This regularization induces sparsity, effectively performing feature selection by shrinking some coefficients to zero.

In contrast, L2-regularized logistic regression, or Ridge logistic regression, penalizes the squared magnitudes of the coefficients:

$$\mathcal{L}_L = -\sum_{i=1}^{n} \left[y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \right] + \lambda \sum_{j=1}^{k} \beta_j^2. \tag{4}$$

This penalty shrinks coefficients toward zero without eliminating them, which can be beneficial in the presence of multicollinearity.

3.2.2. Decision Tree

A decision tree is a non-parametric, supervised learning algorithm that predicts outcomes by recursively partitioning the input space based on feature thresholds. The Classification and Regression Tree (CART) algorithm, introduced by [12], constructs the tree by selecting feature-value splits that minimize an impurity criterion, typically Gini impurity:

$$\text{Gini}(t) = 1 - \sum_{c=1}^{C} p(c \mid t)^2 \tag{5}$$

where $p(c \mid t)$ is the proportion of class c in node t. This recursive partitioning continues until a stopping criterion is met (e.g., maximum depth or minimum node size), resulting in a tree structure where each leaf node represents a final prediction.

3.2.3. Random Forests

Random forest is an ensemble learning technique that aggregates predictions from multiple decision trees to improve classification performance and robustness [20]. Each decision tree in the ensemble is trained on a different bootstrap sample of the training data, and feature selection at each node is randomized. This combination of bootstrap aggregation (bagging) and random feature selection helps ensure low correlation among trees, which in turn reduces model variance.

To classify a new observation, each decision tree provides a prediction, and the random forest outputs the majority vote across all trees. While individual trees are relatively interpretable, the ensemble nature of random forests makes the model difficult to interpret as a whole. As such, random forests are often considered black-box models, despite their strong predictive performance and robustness to overfitting.

3.3. Preprocessing data

To prepare the dataset for modeling, we experimented with two distinct preprocessing strategies: weight of evidence (WoE) encoding and min-max scaling. Each strategy was applied independently, as WoE produces features that are already normalized, thereby eliminating the need for further scaling, while min-max scaling operates directly on the original continuous variables and does not require binning or WoE transformation.

Weight of evidence encoding is a widely used technique in credit risk modeling [21], particularly suitable for datasets containing special values, missing data, or outliers. We began by discretizing continuous features into bins, which allows special values to be grouped into separate categories and helps mitigate the influence of extreme values. WoE assigns each bin a numerical value based on the distribution of creditworthy and non-creditworthy borrowers. For bin i, the WoE value is defined as:

$$\text{WoE}_i = \ln\left(\frac{\frac{N_{\text{good},i}}{N_{\text{good}}}}{\frac{N_{\text{bad},i}}{N_{\text{bad}}}}\right) \tag{6}$$

where $N_{good,i}$ and $N_{bad,i}$ are the numbers of creditworthy and non-creditworthy observations in bin i, and N_{good} and N_{bad} are the total numbers of creditworthy and non-creditworthy observations in the dataset.

One advantage of WoE is that it standardizes feature values on a log-odds scale, which is especially useful for linear models like logistic regression. It also handles missing values and outliers effectively by assigning them to dedicated bins. However, because WoE relies on binning, it may introduce some loss of granularity and is less interpretable outside of credit modeling contexts.

As an alternative, we apply min-max scaling to the original continuous features without discretization. This method normalizes each feature x to a value x' in the range $[0, 1]$, according to the formula:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}. \tag{7}$$

This transformation ensures that all features are on a comparable scale, which can help improve numerical stability and convergence in gradient-based models. While tree-based models such as decision trees and random forests are typically invariant to monotonic transformations, scaling can still be beneficial in controlling feature dominance and improving performance, especially when the features span very different numeric ranges.

Because we use both linear and non-linear models, our preprocessing strategy is designed to test which approach better supports each model type. In particular, we expect WoE to be more effective for linear models, where encoding categorical and binned variables in terms of log-odds enhances model interpretability and alignment with assumptions. For tree-based models, which naturally handle non-linear splits, we test whether simple min-max scaling provides sufficient normalization without the need for more domain-specific transformations like WoE.

3.4. Interpretability methods

In high-stakes domains such as credit scoring, interpretability is a key requirement for model adoption and trustworthiness. While complex models like random forests often deliver superior predictive performance, they are frequently criticized for their black-box nature. In this section, we explore three interpretability methods: coefficient analysis using Weight of Evidence (WoE), Local Interpretable Model-Agnostic Explanations (LIME), and SHapley Additive exPlanations (SHAP). Each method provides a different lens through which to understand model behavior and explain individual predictions.

3.4.1. Coefficient for WoE

When logistic regression is trained using features encoded with Weight of Evidence (WoE), model interpretability is naturally preserved. Since WoE transforms each variable into a continuous value representing the log-odds of creditworthiness, the estimated coefficients in the logistic regression model can be interpreted directly as the marginal effect of each feature on the log-odds of default. A positive coefficient indicates that an increase in the WoE-encoded feature increases the likelihood of default (i.e., reduces creditworthiness), while a negative coefficient implies the opposite.

This approach is particularly appealing for credit risk applications because it aligns with long-standing industry practices and produces additive, transparent risk contributions across features. Moreover, when features are pre-binned and monotonic WoE encodings are applied, the signs and magnitudes of the coefficients tend to be more stable and easier to interpret.

3.4.2. Local Interpretable Model-Agnostic Explanations (LIME)

LIME is a post hoc model-agnostic technique that provides local interpretability by approximating the decision boundary of any black-box model around a specific data point with a linear model [14]. This linear approximation is trained by sampling perturbed versions of the original input and fitting a locally weighted linear regression model. The weights are assigned based on the proximity of the perturbed samples to the original instance, typically measured using a kernel function.

The coefficients of the resulting local surrogate model serve as feature importance scores, highlighting how each input feature contributes to the model's prediction for that particular instance. The strength of LIME lies in its flexibility—it can be applied to any model and any type of data. However, its reliance on sampling introduces randomness, and explanations can vary slightly across different runs. Furthermore, the local linear approximation may not faithfully represent highly non-linear decision boundaries.

3.4.3. SHapley Additive exPlanations (SHAP)

SHAP is an explainable AI method based on cooperative game theory. It attributes a model's prediction for a specific data point to the contributions of each feature using the concept of Shapley values, originally developed to fairly distribute payouts among players in a coalition [22]. SHAP satisfies several desirable properties for interpretability, including local accuracy (the sum of the attributions equals the model output), missingness (features not present receive zero contribution), and consistency (if a feature's contribution increases in a model, its Shapley value will not decrease).

We employ the Kernel SHAP implementation introduced by [15], which approximates Shapley values using a weighted least squares regression. To explain a data point x_i , Kernel SHAP constructs a dataset of feature subsets sampled from x_i, with the remaining features replaced by background values from the training data. Each subset receives a weight based on the size of the subset, with smaller subsets (closer to the marginal contribution of a single feature) weighted more heavily. The regression solution yields the estimated Shapley values.

Despite its theoretical appeal, Kernel SHAP suffers from poor scalability: computing exact Shapley values requires evaluating all 2^k feature subsets, which becomes computationally infeasible for high-dimensional datasets. Approximate methods and sampling strategies are used in practice, but the method remains relatively expensive compared to alternatives like LIME.

3.4.4. Comparative Discussion

Both LIME and SHAP provide complementary perspectives for model interpretability. LIME excels in computational efficiency and model-agnostic flexibility, offering quick local approximations that are especially useful when working with large feature sets or real-time explanations. However, LIME may lack fidelity in capturing true feature interactions and does not guarantee consistency or local accuracy.

SHAP, on the other hand, offers theoretically grounded explanations that reflect both individual feature contributions and their interactions. It provides robust, additive attributions that sum to the model's prediction, but it is computationally more intensive and may be less suitable for real-time applications.

In practice, the choice between LIME and SHAP should be guided by the specific use case. For instance, when interpretability is paramount for auditability or regulatory compliance, SHAP may be more appropriate despite its computational cost. In contrast, when speed is essential and the model is used in a dynamic setting with frequent queries, LIME may offer a more practical solution. By applying both techniques judiciously, practitioners can better understand and validate complex machine learning models, particularly in sensitive domains like credit risk assessment.

4. Results

4.1. Model performance

In this section, we evaluate the performance of various machine learning models under two different preprocessing strategies: Weight of Evidence (WoE) encoding and one-hot encoding with min-max scaling. The models under consideration include L1- and L2-regularized logistic regression, decision tree, and random forest classifiers.

We employ 5-fold cross-validation to assess the generalization performance of each model. In this setup, the dataset is randomly partitioned into five equal subsets. In each fold, one subset is held out as the test set, while the remaining four subsets are used for training. Thus, in each iteration, the training set consists of 80% of the data and the test set comprises the remaining 20%. Performance metrics are computed on the test set and then averaged across all five folds to ensure robustness.

We utilized a dataset of 55,707 records (8,323 non-creditworthy and 47,384 creditworthy) and applied a 5-fold cross-validation approach, which inherently combines training and testing phases without a separate validation set. Specifically, in each fold, the dataset was split into a training set of approximately 44,566 records (80% of the data) and a test set of 11,141 records (20%), totaling 38,994 records for training and 16,713 for testing across all folds, as derived from the provided split (38,994 training + 16,713 test = 55,707). We did not use a distinct validation set because the 5-fold cross-validation process effectively validates the model by rotating the test set across folds, optimizing performance metrics like recall for the non-creditworthy class, as detailed in our hyperparameter tuning with GridSearchCV. This approach ensures that the model is evaluated on multiple subsets, providing a robust estimate of generalization performance without requiring a separate validation set.

To address the class imbalance in our dataset, we implemented a combination of Synthetic Minority Oversampling Technique (SMOTE), class weighting, and a recall-focused scoring metric to enhance the performance of our machine learning models, particularly for the non-creditworthy class. SMOTE was applied to the training data within each fold of our 5-fold cross-validation to generate synthetic non-creditworthy samples, balancing the class distribution while preserving the original test set for realistic evaluation. Additionally, we incorporated class weighting (e.g., balanced or a 1:5 ratio favoring non-creditworthy) in our logistic regression and random forest models to penalize misclassifications of the minority class more heavily. By using recall as the primary scoring metric in hyperparameter tuning via GridSearchCV, we prioritized the identification of non-creditworthy loans, minimizing false negatives critical to credit risk assessment. These strategies collectively mitigated the risk of overfitting and improved the models' ability to accurately classify non-creditworthy records, as evidenced by enhanced recall scores in our results.

To evaluate classification performance, we use accuracy, area under the receiver operating characteristic curve (AUC), and recall, with particular emphasis on recall due to the cost-sensitive nature of credit risk. Since our goal is to minimize the number of high-risk borrowers who are incorrectly classified as low-risk (i.e., false negatives), recall-defined as the proportion of true defaulters correctly identified is of primary importance. The AUC reflects the model's ability to distinguish between creditworthy and non-creditworthy applicants, while accuracy captures the overall proportion of correctly classified samples. Given that our dataset is relatively balanced, accuracy remains a meaningful metric alongside AUC and recall. For all models, the predicted probability of default is converted to a binary classification using a threshold of 0.5.

Tables 1 and 2 summarize the results of the 5-fold cross-validation for the two preprocessing pipelines. Among the models trained on WoE-encoded data, logistic regression models perform best in terms of both AUC and recall. Specifically, L2-penalized logistic regression achieves an accuracy of 0.66, an AUC of 0.71, and a recall of 0.67. Although the random forest achieves the highest accuracy (0.71), its recall is considerably lower (0.50), which limits its effectiveness for detecting defaulters. This supports the view that WoE encoding, when combined with interpretable linear models, offers strong predictive performance while maintaining transparency.

Table 2. 5-fold cross-validation performance of ML models using WoE encoding

Model	**Accuracy**	**AUC**	**Recall**
L1 Logistic Regression	0.65	0.70	0.65
L2 Logistic Regression	0.66	0.71	0.67
Decision Tree	0.61	0.67	0.61
Random Forest	0.71	0.69	0.50

In contrast, when models are trained on min-max scaled data with one-hot encoding, the random forest outperforms the other models across all metrics. It achieves an accuracy of 0.78, an AUC of 0.69, and a recall of 0.38. However, its recall remains relatively low, suggesting that even with higher accuracy, it may not be optimal for identifying risky applicants. Meanwhile, L1-penalized logistic regression achieves a competitive recall of 0.66, though its accuracy (0.59) and AUC (0.67) are lower than those of the random forest.

Table 3. 5-fold cross-validation performance of ML models using min-max scaling

Model	**Accuracy**	**AUC**	**Recall**
L1 Logistic Regression	0.59	0.67	0.66
L2 Logistic Regression	0.71	0.65	0.45
Decision Tree	0.57	0.66	0.67
Random Forest	0.78	0.69	0.38

From a practical standpoint, logistic regression trained with WoE features presents an attractive option for credit scoring applications. It offers interpretable coefficients that align with industry standards and regulatory requirements. However, implementing WoE encoding requires feature binning, monotonicity constraints, and careful calibration, which increases preprocessing complexity.

On the other hand, random forests, though superior in raw predictive power when trained on scaled features, suffer from limited interpretability. The ensemble nature of the model, which aggregates hundreds of decision paths, makes it difficult to explain individual predictions—an issue particularly relevant in regulated domains like consumer finance.

In the context of peer-to-peer lending, both types of misclassification-false negatives (defaulters misclassified as creditworthy) and false positives (creditworthy applicants denied loans)-have important business implications. Misclassifying defaulters results in financial losses, while rejecting potentially reliable borrowers leads to lost revenue. Given the scale of the lending industry, even small gains in recall or precision can translate into substantial economic impact. Despite this, regulatory constraints and the need for explainability often prevent lenders from adopting more complex but opaque models. This trade-off motivates our deeper investigation into model interpretability in the next section.

4.2. Explaining model results

In this section, we analyze the interpretability of the models to support their practical adoption in loan decision-making systems. Machine learning models are increasingly used by lending institutions to assess the creditworthiness of borrowers. However, regulatory frameworks such as the Equal Credit Opportunity Act (ECOA) and the Fair Credit Reporting Act (FCRA) in the United States require that lenders provide specific reasons for loan denial. This has created a strong demand for interpretable models and reliable post hoc explanation techniques.

Interpretability is not only essential for regulatory compliance but also for building trust with applicants and improving internal risk assessment procedures. In this context, lenders seek to identify the key factors driving creditworthiness and to generate understandable explanations for individual decisions made by the model. Transparent models also allow companies to identify representative historical borrowers whose profiles are similar to new applicants, thus supporting a case-based reasoning approach.

4.2.1. WoE Coefficients for Logistic Regression

One effective approach to achieving model interpretability is through the use of logistic regression trained on Weight of Evidence (WoE)-encoded features. In this setting, each feature represents the log-odds of being creditworthy within a given bin, and the coefficients of the logistic regression model quantify the contribution of each feature to the log-odds of default.

The intercept of the trained L2-penalized logistic regression model is –1.713, and the coefficients for each WoE-transformed feature are presented in Table 3. A positive coefficient implies that higher values of the corresponding WoE feature (i.e., riskier bins) increase the likelihood of default, while negative coefficients imply the opposite. Because WoE encoding aligns feature values with the probability of default, the resulting coefficients can be directly interpreted as directional indicators of credit risk.

As shown in the table, the loan amount (loan_amnt_woe) is the most influential variable in determining creditworthiness, with a coefficient of 1.247. This suggests that larger loan amounts are associated with a higher probability of default. Other important predictors include recent credit inquiries (inq_last_6mths_woe, 0.971), annual income (annual_inc_woe, 0.908), and FICO score range (fico_range_low_woe, 0.725). These features align well with common industry understanding of credit risk factors.

The simplicity and transparency of this model make it particularly suitable for lending environments where interpretability and regulatory reporting are as important as predictive accuracy. In contrast to complex models such as random forests, logistic regression with WoE encoding provides clear justifications for both individual and population-level decisions.

Table 4. Coefficients of L2-Penalized Logistic Regression (WoE Model)

Feature	**Coefficient**
loan_amnt_woe	1.247
inq_last_6mths_woe	0.971
annual_inc_woe	0.908
fico_range_low_woe	0.725
verification_status_woe	0.677
home_ownership_woe	0.637
num_il_tl_woe	0.487
revol_util_woe	0.473
mort_acc_woe	0.423
mths_since_rcnt_il_woe	0.315
mo_sin_old_rev_tl_op_woe	0.255

4.2.2. Local Interpretable Model-Agnostic Explanations (LIME) for Random Forest

LIME (Local Interpretable Model-Agnostic Explanations) is a post hoc interpretability technique that approximates complex models by training a local, interpretable surrogate model around a prediction of interest [14]. In our case, LIME is used to interpret predictions made by the random forest model trained on one-hot encoded features with min-max scaling.

For each individual data point, LIME perturbs the instance to generate a synthetic neighborhood and fits a locally weighted linear regression to approximate the black-box model's behavior in that region. The coefficients of this surrogate model represent the impact of each feature on the prediction and can be interpreted as the change in the predicted probability resulting from a unit change in the feature value, holding other features constant.

Figure 1 shows a LIME explanation for a single borrower. The model assigns a 91% predicted probability of default, indicating high credit risk. The bar charts in the figure break down this prediction by feature contribution. Features shown in red increase the probability of default (Class 1), while those in green support a prediction of no default (Class 0).

In this example, the most influential features increasing the likelihood of default are verification_status_Verified = 0.00, indicating unverified income, which contributes approximately +0.05 to the prediction; application_type_Joint App = 0.00, adding +0.04; and both home_ownership_MORTGAGE = 0.00 and home_ownership_RENT = 0.00, each contributing +0.03. Additionally, fico_range _low = 0.35 and mort_acc = 0.25 fall into intervals that the model associates with higher risk, each contributing approximately +0.02 to the probability of default.

These features push the prediction strongly toward the "Default" class. On the other hand, features such as loan_amnt = 0.49, num_actv_bc_tl = 0.17, and inq_last_12m = 0.01 slightly counterbalance the default risk, with negative contributions shown in green. However, the mitigating effect of these features is not sufficient to override the cumulative positive influence of the others, resulting in a high default prediction.

LIME's local explanation highlights which features the model relied on for this specific decision and allows decision-makers to trace the rationale behind a prediction. While LIME does not guarantee global consistency or faithfulness to the underlying model, it is computationally efficient and flexible across model types and data structures. In production environments where transparency is critical for instance, when rejecting a loan application LIME can help generate individualized explanations in real time. These explanations satisfy regulatory requirements and help build trust with customers by providing a human-understandable rationale for each prediction.

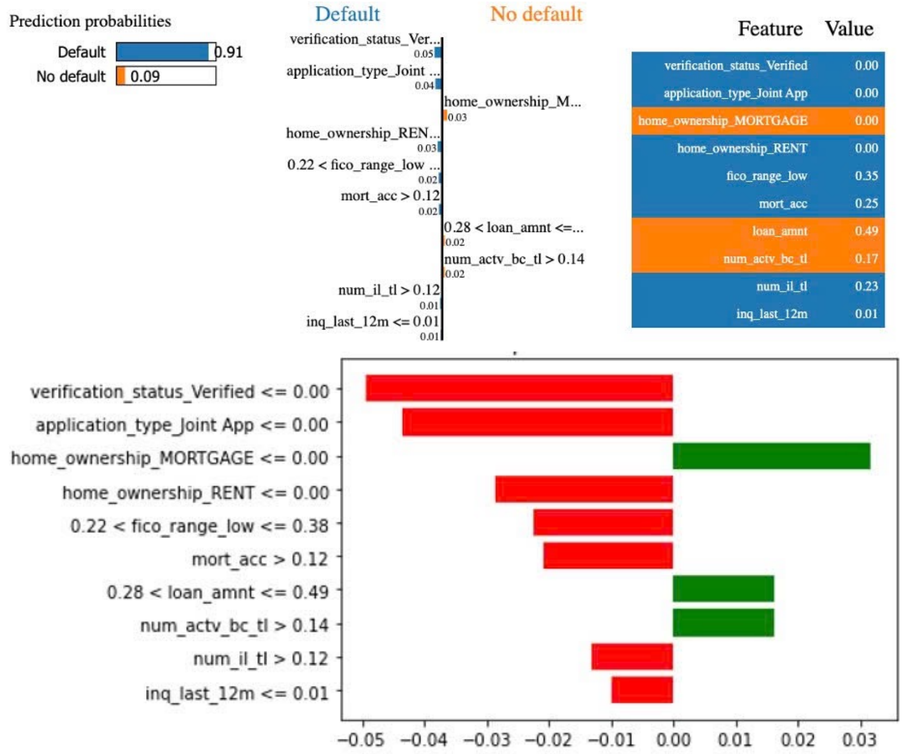
4.2.3. SHapley Additive exPlanations (SHAP) for Random Forest

To interpret the predictions of the random forest model trained on one-hot encoded features with min-max scaling, we apply SHapley Additive exPlanations (SHAP). SHAP is an explainability technique rooted in cooperative game theory that decomposes a model's prediction into the contributions of each input feature [22]. For this task, we use the Tree SHAP algorithm, which is optimized for ensemble models such as random forests [15].

SHAP provides global explanations by measuring the average magnitude of each feature's contribution across all instances in the dataset. These values represent how much each feature, on average, influences the model's prediction toward either class: "Default" (Class 1) or "No Default" (Class 0). Figure 2 presents the SHAP summary plot based on mean absolute SHAP values, with red bars representing contributions toward predicting default, and blue bars representing contributions toward predicting no default.

The results highlight loan_amnt, fico_range_low, and verification_status_Veri -fied as the most influential predictors in the model's decision-making. This aligns well with financial intuition: larger loan amounts and lower FICO scores are commonly associated with higher credit risk, while income verification status reflects the reliability of the reported income, which can strongly influence repayment behavior. Other important features include mort_acc (number of mortgage accounts), home_ownership_MORTGAGE, and inq_last_6mths (number of credit inquiries in the last six months), all of which are standard indicators in credit risk evaluation.

Unlike the WoE-based logistic regression model described in Section 4.2.1, which offers interpretability through linear coefficients aligned with log-odds,



Feature	Value
verification_status_Verified	0.00
application_type_Joint App	0.00
home_ownership_MORTGAGE	0.00
home_ownership_RENT	0.00
fico_range_low	0.35
mort_acc	0.25
loan_amnt	0.49
num_actv_bc_tl	0.17
num_il_tl	0.23
inq_last_12m	0.01

Figure 1. LIME explanation for one borrower classified as ”Default” by the random forest model (predicted probability = 0.91). Red bars indicate features contributing to the ”Default” prediction, while green bars indicate features supporting ”No Default.”

the random forest model requires post hoc interpretability tools like SHAP due to its non-linear and ensemble nature. While the random forest model achieves higher accuracy, its lack of transparency can be a barrier to deployment in regulated financial contexts. SHAP mitigates this by revealing how each feature contributes to predictions at both the global and individual levels.

Despite its advantages, SHAP also has limitations. While Tree SHAP is computationally efficient compared to the original Shapley value formulation, it can still be resource-intensive for very large models or datasets. Moreover, SHAP's explanations, while grounded in strong theory, may still be difficult to communicate to non-technical stakeholders, particularly when many features are involved.

Nonetheless, SHAP serves as a powerful bridge between predictive performance and interpretability. It allows stakeholders to audit the behavior of complex models like random forests and to gain trust in model predictions by understanding the most influential drivers of credit decisions.

We compare the interpretability results of our Random Forest model (min-max scaling) in Figures 1 and 2 with similar existing works to assess their quality. Our SHAP analysis (Figure 2) identifies loan amount, FICO score, and verification status as top predictors, aligning closely with [16] and [17], who also highlight FICO score and loan amount using SHAP on Lending Club data, and [18], who emphasize credit history on a Colombian P2P dataset. Similarly, our LIME explanation (Figure 1) provides detailed contributions (e.g., unverified income: +0.05, FICO score: +0.02) for a borrower predicted as ”Default” (probability 0.91), offering more granularity than LIME results in Hadji Misheva et al. and Ariza-Garzón et al., enhancing individual decision explanations. We argue that these results are ”good enough” for P2P lending credit risk assessment, as they provide actionable local (LIME) and global (SHAP) insights, meeting regulatory requirements for transparency and aligning with traditional risk factors like FICO score and loan amount. Despite the Random Forest's lower recall (0.38) compared to L2 Logistic Regression (0.67), the interpretability results are sufficiently detailed and relevant, supporting stakeholder trust and responsible lending practices though future work could improve predictive performance for non-creditworthy detection.

5. Conclusion

This study investigated the trade-off between predictive performance and model interpretability in the context of credit risk assessment for peer-to-peer (P2P) lending. Using historical data from Lending Club, we compared the effectiveness of logistic regression, decision tree, and random forest models under two preprocessing pipelines: Weight of Evidence (WoE) encoding and one-hot encoding with min-max scaling. Our results show that while random forest models trained on scaled one-hot features achieve the highest accuracy, logistic regression models using WoE encoding strike a more desirable balance between predictive power and interoperability.

From a practical perspective, the choice of model should reflect the priorities of the lending platform. When the primary objective is regulatory compliance and transparency, as is often the case in highly regulated environments, logistic regression with WoE encoding offers a clear advantage. This approach enables lenders to trace the contribution of each feature to the predicted probability of default, making the model easier to audit and justify. On the other hand, when prediction accuracy is paramount, especially in settings where regulatory

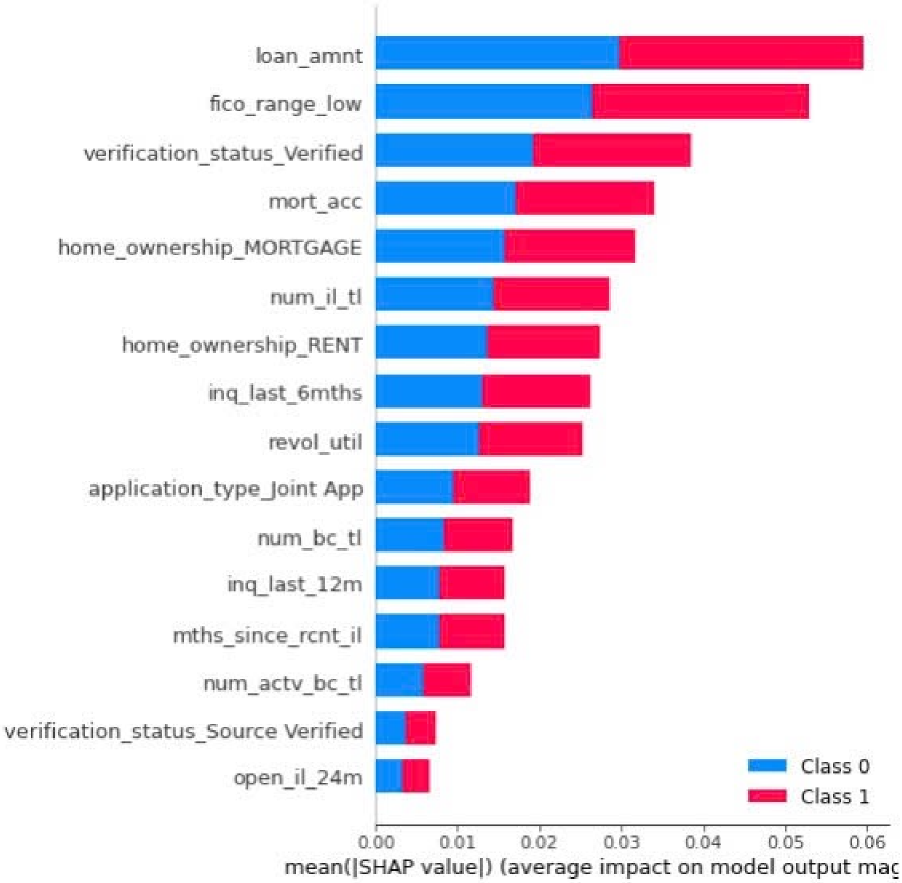


Figure 2. SHAP summary plot showing the global feature importance for the random forest model trained using one-hot encoding and min-max scaling. Red bars indicate contributions toward predicting ”Default” (Class 1), while blue bars indicate contributions toward predicting ”No Default” (Class 0).

constraints are less strict, random forests can provide superior performance.

To address the opacity of ensemble models, we employed two post hoc explainability techniques-LIME and SHAP-to interpret the predictions of the random forest. These tools revealed that key drivers of credit decisions include loan amount, FICO score, verification status, and the number of mortgage accounts factors that are consistent with traditional credit risk evaluation. The application of LIME enabled localized, instance-specific explanations, which are useful for generating individualized decision justifications. SHAP, in contrast, offered a global perspective on feature importance, contributing to broader model understanding and policy refinement.

Ultimately, our findings highlight that interpretability and accuracy need not be mutually exclusive. By selecting modeling techniques and explanation methods aligned with institutional goals and regulatory expectations, lenders can build trustworthy, effective credit scoring systems. As the P2P lending industry continues to evolve, integrating transparent machine learning models will be essential for promoting responsible lending practices, enhancing borrower trust, and maintaining regulatory compliance.

References

[1] **Lee, I., Shin, Y. J.** *Fintech: Ecosystem, business models, investment decisions, and challenges.* (Business Horizons, 2018), 61(1), 35–46.

[2] **Liu, H., Qiao, H., Wang, S., Li, Y.** *Platform competition in peer-to-peer lending considering risk control ability.* European Journal of Operational Research, 2018, 274(1), 280–290.

[3] **Turner, A.** *After the crisis, the banks are safer but the debt is a danger. Financial Times.* (2025, March 1). https://www.ft.com/content/9f481d3c-b4de-11e8-a1d8-15c2dd1280ff

[4] **Huang, R. H.** *Online P2P lending and regulatory responses in China: Opportunities and challenges.* European Business Organization Law Review (2018), 19(1), 78.

[5] **Havrylchyk, O.** *Regulatory framework for the loan-based crowdfunding platforms.* OECD Economics Department Working Papers. (2021)

[6] **Chapman, J. M.** *Factors affecting credit risk in personal lending.* In Commercial Banks and Consumer Installment Credit (1940), 109–139. NBER.

[7] **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. Electronic Commerce Research and Applications (2018), 31, 24–39.

[8] **Duan, J.** *Financial system modeling using deep neural networks (DNNs) for effective risk assessment and prediction.* Journal of the Franklin Institute (2019), 356, 4716–4731.

[9] **Sirignano, J., Sadhwani, A., Giesecke, K.** *Deep learning for mortgage risk* (2016) arXiv preprint arXiv:1607.02470.

[10] **Kvamme, H., Sellereite, N., Aas, K., Sjursen, S.** *Predicting mortgage default using convolutional neural networks.* Expert Systems with Applications (2018), 102, 207–217.

[11] **Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W., Siddique, A.** *Risk and risk management in the credit card industry.* (2016) Journal of Banking & Finance, 72, 218–239.

[12] **Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J.** Classification and regression trees. Monterey, CA: Wadsworth and Brooks. (1984).

[13] **Goodman, B., & Flaxman, S.** *European Union regulations on algorithmic decision-making and a "right to explanation".* AI Magazine (2017), 38(3), 50–57.

[14] **Ribeiro, M. T., Singh, S., & Guestrin, C.** *"Why should I trust you?": Explaining the predictions of any classifier.* In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016), pp. 1135–1144.

[15] **Lundberg, S. M., & Lee, S.-I.** *A unified approach to interpreting model predictions.* In Advances in Neural Information Processing Systems (2017), 30, 4765–4774.

[16] **Hadji Misheva, B., Hirsa, A., Osterrieder, J., Kulkarni, O., & Fung Lin, S.** *Explainable AI in credit risk management.* Credit Risk Management (2021). https://ssrn.com/abstract=3768437

[17] **Albanesi, S., & Vamossy, D. F.** *Predicting consumer default: A deep learning approach.* National Bureau of Economic Research Working Paper. (2019)

[18] **Ariza-Garzón, M. J., Arroyo, J., Caparrini, A., & Segovia-Vargas, M.-J.** *Explainability of a machine learning granting scoring model in peer-to-peer lending.* IEEE Access (2020), 8, 64873–64890.

[19] **Breiman, L.** *Random Forests* Machine Learning (2021), 45, 5-32. http://dx.doi.org/10.1023/A:1010933404324

[20] **Cox, D. R.** *The regression analysis of binary sequences.* Journal of the Royal Statistical Society: Series B (Methodological) (1958), 20(2), 215–242.

[21] **Siddiqi, N.** *Credit risk scorecards: Developing and implementing intelligent credit scoring (Vol. 3)* John Wiley & Sons. (2012)

[22] **Shapley, L.** *A Value for n-Person Games. In: Kuhn, H. and Tucker, A., Eds., Contributions to the Theory of Games II* Princeton University Press, Princeton (1958), 307-317. https://doi.org/10.1515/9781400881970-018

[23] Link dataset (https://www.kaggle.com/datasets/wordsforthewise/lending-club)

Dinh Thuy Tien
Thang Long University
Hanoi
Vietnam
dinhthuytien@thanglong.edu.vn

Nguyen Thi Tra My
Thang Long University
Hanoi
Vietnam
myntt@thanglong.edu.vn