Explainable AI in Credit Risk Assessment for External Customers

ALEXANDRU MATCOV, University of Twente, The Netherlands

The increasing use of AI in credit risk assessment has brought significant advancements to the financial industry. However, the complex nature of AI models often results in a lack of transparency, making it challenging for customers to understand and trust these systems. This paper will investigate how interpretability methodologies such as LIME and SHAP can improve customer comprehension of AI-driven credit risk evaluations. Through a rigorous literature review and analysis of a public credit dataset, this research will explore effective visualization strategies, evaluate the clarity and transparency that LIME and SHAP offer, and address the challenges of applying these methodologies to enhance interpretability in both public and private datasets.

Additional Key Words and Phrases: Artificial Intelligence (AI), Credit Risk Assessment, Interpretability, Explainability, Local Interpretable Model-agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP), Transparency

1 INTRODUCTION

In the rapidly evolving landscape of financial services, the adoption of Artificial Intelligence (AI) in credit risk assessment has been a pivotal factor, significantly advancing the precision and efficiency of decision-making processes. Simultaneously with the increase of quantity of data, the area of credit risk assessment has evolved from simple statistical models to more complex Machine Learning (ML) models [16]. Despite the widespread integration of sophisticated AI models in industries [9], they often operate as "black boxes" — their decision-making processes are not transparent to users [27].

Problematically, though they appear powerful in terms of results and predictions, AI algorithms suffer from opacity, and it is difficult to get insight into their internal mechanism of work, especially ML algorithms. This further compounds the problem, because entrusting important decisions to a system that cannot explain itself presents obvious dangers. To address this issue, model explainability has recently regained attention with the emerging area of eXplainable AI (XAI), a concept that focuses on opening black-box models in order to improve the understanding of the logic behind the prediction, making a shift towards more transparent AI [17]. It aims to create a suite of techniques that produce more explainable models whilst maintaining high-performance levels [1, 14, 33].

This lack of clarity diminishes user trust and does not meet regulatory requirements for transparency and accountability, highlighted by frameworks such as the European Commission's Artificial Intelligence Act (AIA) regulation which calls for explainability in high-risk AI systems like credit scoring [15]. Moreover, the topic of explainability holds particular significance for financial service providers within Europe due to the enforcement of the General Data Protection Regulation (GDPR). This regulation spans the entirety of the European Union and its provisions, particularly Articles 13-15, establish a user's right to receive an explanation regarding automated decisions that significantly impact them [26].

The challenge to match the high predictive accuracy of the complex ML models with user understanding has led to the development of post-hoc explainability techniques, notably Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP). These methodologies seek to elucidate the complexity behind AI model predictions, adopting trust and comprehension among users—fundamental elements for widespread adoption and responsible use of AI in financial services [26, 29].

Implementing these methodologies, however, presents its own set of challenges. The complexity found in publicly available financial datasets calls for a careful application of LIME and SHAP to ensure that interpretability remains both accurate and actionable. Further, it's crucial that efforts to improve explainability do not compromise the robust predictive capabilities of AI models, maintaining a balance between simplicity in explanation and model performance [24].

This study will undertake a detailed literature review to investigate how LIME and SHAP methodologies can enhance the interpretability and explainability of AI models in credit risk assessment, particularly for users without specialized expertise. In addition, this research intends to apply these cutting-edge XAI methods to ML-based credit scoring models using publicly available credit datasets. The primary aim is to discover and implement effective visualization techniques to improve the clarity and user-friendliness of interpretability tools like LIME and SHAP for the everyday customer.

While significant research has been undertaken in the realm of XAI, particularly concerning SHAP and LIME methodologies for credit risk assessment [6, 7, 12, 13, 15, 25, 26, 32], there remains a notable gap in making these interpretations accessible and easily comprehensible to external customers. There is an imperative need to enhance the explainability of these AI models without sacrificing their high accuracy while maintaining computational efficiency. This thesis aims to refine visualization techniques that break down LIME and SHAP outputs into clear, user-focused features, aligning with legal requirements for transparency. Thus, the main research question is:

How can LIME and SHAP methodologies enhance the explainability of AI models in credit risk assessment for external customer comprehension?

The primary research question can be answered by answering the following sub-research questions:

(1) To what extent can LIME and SHAP methodologies provide clarity and transparency of complex ML models for external customers?
(2) How can customers effectively visualize credit risk assessment information, and what insights can customers derive from the available information?

TScIT 40, February 2, 2024, Enschede, The Netherlands

2 RELATED WORK

In their 2016 paper, Ribeiro et al. [29] introduced LIME, a novel technique that elucidates the reasoning behind ML model predictions. Their approach was distinctive in its model-agnostic nature, capable of making opaque, complex models interpretable and trustworthy. The introduction of LIME provided a tool for bridging the gap between model accuracy and user trust, a pivotal step in the field of explainable AI.

In a subsequent advancement, Lundberg et al. [24] in 2017 proposed the SHAP framework, which presents a unified measure of feature importance in model predictions. By assigning importance values to each feature, SHAP enables an understanding of their impact on the model's output, marking an essential contribution to the interpretability of complex models. The SHAP framework has been instrumental in resolving the accuracy-interpretability trade-off in the usage of large datasets.

Building upon the foundation of LIME and SHAP in the domain of financial services, the paper by Misheva et al. in 2021 [26] addresses the pressing need for explainability in AI-powered credit risk management. The research evaluates the effectiveness of LIME and SHAP in interpreting complex ML models, crucial for establishing trust and transparency in automated financial decision-making processes. This exploration into post-hoc model explanations is pivotal in navigating the trade-offs between model complexity and the demand for clear, interpretable AI within the finance sector.

Another notable work in the field of finance is the study by Gramegna et al. [18] which assesses the discriminative power of LIME and SHAP, revealing SHAP's superiority in clarity and predictive accuracy for unsupervised learning models. While they're all implementing LIME, SHAP, or a combination of both, none of these papers explore the explainability aspect from an external customer perspective.

Expanding upon this, Davis et al. [12] apply both SHAP and LIME to analyze the explainability of credit risk models. Their findings highlight a significant trade-off: while LIME shows potential instability, KernelSHAP's computation time makes it less practical for large datasets. This points to the ongoing challenge of balancing model explainability and efficiency, especially in complex financial applications and large datasets.

In one of the latest advancements within the financial sector, the work by Fritz et al. [15] showcases a practical application of SHAP values through a method known as 'SHAP clustering'. This approach simplifies the understanding of AI/ML models used in credit risk management by grouping similar data points, enabling a more transparent view into the model's decision-making processes. Notably, this technique also enhances consumer interactions by allowing for quick, understandable explanations of automated financial decisions, thanks to its efficient GPU-accelerated computation. This innovation holds promise for widespread application in various financial technologies, from traditional banking to fintech platforms.

Although the literature focusing on the use of XAI for credit risk assessment is limited, there is still some other highly relevant research done in this field. For example, researchers have explored the integration of XAI into credit scoring models for peer-to-peer (P2P) leading datasets [6, 7, 13, 25, 32]. Most of these papers used the popular open-access dataset offered by the US-based P2P Lending Platform, Lending Club[1], a real P2P lending platform, focusing on financial data about loans grants. Subsequent studies, such as those by Davis et al. [12] on home equity credit risk and Benhamou et al. [3] on predicting market crashes in the S&P 500, have broadened the application of XAI in credit risk, showcasing its utility in diverse financial sectors. While using different ML models in combination with LIME, SHAP, or both, the above-mentioned papers conclude that the utilization of AI for enhanced predictive performance, paired with XAI for explainability, can significantly improve the current credit scoring models.

Nevertheless, a significant gap remains in existing research: while most studies focus on enhancing the technical explainability of machine learning models to aid finance professionals in improving credit risk prediction systems, they largely overlook the aspect of explainability from the perspective of external customers.

3 METHODOLOGY

3.1 RandomForest

The RandomForest (RF) classifier is an ensemble method that generates multiple decision trees during training and outputs the mode of their classifications or the mean prediction for regression [5]. Unlike gradient boosting models, RF builds each tree independently, using a different subset of the data. This process enhances generalization, reducing the risk of overfitting [2]. RF performs well on a wide range of tasks with minimal data preprocessing required. It can manage categorical features effectively and is less prone to dimensionality. The algorithm offers insights into feature importance, as it evaluates the impact of each feature on the model's accuracy. While RF typically requires careful tuning of hyperparameters, such as the number of trees and maximum depth, to prevent overfitting and underfitting, it is favored for its interpretability and reliability in predictions[23]. However, its computational cost increases with the size of the dataset, which can impact training time and memory requirements [34].

3.2 LIME

LIME seeks to approximate any black-box ML model with a local, understandable model to articulate individual prediction justifications [29]. It operates on a local level, meaning that explanations are specific to individual observations. LIME works by attempting to construct a simpler, local model using data points that resemble the specific instance it aims to explain. This local model can be derived from a class of interpretable models, which includes, but is not limited to, linear models and decision trees.

For a given observation x, LIME determines the explanation $\Phi(x)$ as:

$$\Phi(x) = \text{argmin}_{g \in \mathcal{G}} L(f, g, \pi_x) + \Omega(g) \tag{1}$$

In this context, $\mathcal{G}$ is a set of potential models that are interpretable, such as linear models and decision trees. Here, $g \in \mathcal{G}$ represents an explanation modeled as a simpler interpretable model. The function $f : \mathbb{R}^d \to \mathbb{R}$ is the complex model being approximated. The term

[1] https://www.lendingclub.com/

$\pi_x(z)$ is a proximity measure of an instance z from the original instance x, and $\Omega(g)$ quantifies the complexity of the explanation $g \in \mathcal{G}$.

The primary aim is to minimize the locality-aware loss L without presuming any characteristics about f, signifying the model-agnostic property of LIME. Here, L evaluates the fidelity of g in approximating f within the locality delineated by $\pi(x)$.

3.3 SHAP

Introduced by Lundberg et al. [24], the SHAP (SHapley Additive exPlanations) framework adapts concepts from cooperative game theory to analyze the distribution of contributions among input features in predictive models. The framework's strength lies in its versatility, allowing for the evaluation of feature impact across diverse covariates. SHAP facilitates the assessment of individual feature contributions to predictions, independent of the model's complexity [21].

The SHAP method, leveraging the Shapley value concept [30], provides a quantification of feature importance that remains consistent whether a feature is included in the model or not. In essence, SHAP simplifies the model's predictions into a linear combination of feature contributions. Formally, the SHAP framework models a prediction $f(x)$ using a function $g(z')$, which is a sum of feature attributions:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \tag{2}$$

where M denotes the total number of features. Here, ϕ_i represents the contribution of feature i, and z' is a binary vector indicating the presence of features.

Lundberg et al. [24] established that the unique additive approach satisfying local accuracy, missingness, and consistency is the Shapley value. This value is calculated for each feature x_i and is given by:

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!|}{M!} [f_x(z') - f_x(z' \setminus i)] \tag{3}$$

In this equation, f is the original predictive model, x' and z' denote the subsets of features, and $|z'|$ is the cardinality of z'. The term $f_x(z') - f_x(z' \setminus i)$ signifies the change in the prediction when feature set z' is included versus when it is absent, attributing this change to the feature x_i.

The Shapley value model provides an intuitive means of decomposing a model's predictions, offering insights that are faithful to the original model's local outputs. It accounts for the effect of having a particular feature present or absent, thus maintaining a truthful representation of feature contributions, and ensures that if a feature does not change the prediction, its attributed importance will reflect this [24].

4 EXPERIMENTAL SET-UP

The step-by-step organization of the experimental set-up is showcased in Figure 1. We start by introducing the chosen dataset, describing its features and characteristics, and justifying its selection. Preprocessing consists of detailed steps preparing the data, encompassing data cleaning, feature selection and feature engineering, handling missing values, and balancing the dataset. Subsequently, we describe the ML model of choice - RF, justifying its selection and describing its implementation. Finally, we delve into hyperparameter tuning, explaining our parameter choices and the optimization of the chosen ML model.

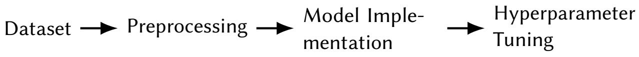


Fig. 1. Experimental Set-up Process

4.1 Dataset

The dataset used in this research is sourced from Kaggle [10], is divided into two subdatasets: *cs-training.csv* and *cs-test.csv*, encompassing a total of 12 features and 251, 503 records of borrower's historical financial and demographic data. The *cs-training.csv* contains 150, 000 records, while *cs-test.csv* comprises the remainder. For this paper, only the *cs-training.csv* dataset was utilized, maintaining all 12 features. This decision was made to reduce the computational time required for model training and for applying explainable methods, particularly SHAP, which, due to its detailed output, requires significant computational resources, making it less practical for larger datasets [12].

The data is highly imbalanced as the defaulting customers constitute a minority class, with only $\sim 6.7\%$ (10, 026 instances) of the records showing serious delinquency against 139, 974 that do not. This imbalance presents a significant challenge for predictive modeling, as it can skew the model's performance towards the majority class. At the core of the dataset is the 'SeriousDlqin2yrs' target variable, which is aligned with the industry standard and definition of default [28]. This variable identifies whether a borrower has been delinquent by 90 days or more on any credit line over a two-year period, making it an essential element for developing risk prediction models.

The dataset was chosen for its strong industry relevance, reflected in the inclusion of the target variable 'SeriousDlqin2yrs', which aligns with key financial benchmarks, and its realistic representation of data imbalance, mirroring actual credit risk scenarios. Additionally, its 12 clear and comprehensible features, coupled with the dataset's overall cleanliness and minimal instances of missing values, make it suitable for realistic financial risk assessment. Special attention may be drawn to the 'count' column in the descriptive statistics, as shown in Table 1, which highlights the completeness of each feature by indicating any instances of missing values.

Feature	count	mean	std	min	25%	50%	75%	max
Unnamed: 0	150,000	75,000.5	43,301.4	1	37,500.8	75,000.5	112,500.3	150,000
SeriousDlqin2yrs	150,000	0.1	0.2	0	0.0	0.0	0.0	1
RevolvingUtilizationOfUnsecuredLines	150,000	6.0	249.8	0	0.0	0.2	0.6	50,708
age	150,000	52.3	14.8	0	41.0	52.0	63.0	109
NumberOfTime30-59DaysPastDueNotWorse	150,000	0.4	4.2	0	0.0	0.0	0.0	98
DebtRatio	150,000	353.0	2,037.8	0	0.2	0.4	0.9	329,664
MonthlyIncome	120,269	6,670.2	14,384.7	0	3,400.0	5,400.0	8,249.0	3,008,750
NumberOfOpenCreditLinesAndLoans	150,000	8.5	5.1	0	5.0	8.0	11.0	58
NumberOfTimes90DaysLate	150,000	0.3	4.2	0	0.0	0.0	0.0	98
NumberRealEstateLoansOrLines	150,000	1.0	1.1	0	0.0	1.0	2.0	54
NumberOfTime60-89DaysPastDueNotWorse	150,000	0.2	4.2	0	0.0	0.0	0.0	98
NumberOfDependents	146,076	0.8	1.1	0	0.0	0.0	1.0	20

Table 1. Descriptive statistics of the dataset.

4.2 Preprocessing

4.2.1 Inputation of missing values. During the initial data analysis, it was observed that two features, namely ‘MonthlyIncome‘ and ‘NumberOfDependents‘, contained missing values. To handle these missing entries and preserve the integrity of the dataset for accurate model training, an imputation technique was employed.

The k-Nearest Neighbors (kNN) algorithm was selected for this purpose, with the number of neighbors set to 5. The model choice is based on fact that it is non-parametric, implying no assumption about the underlying data distribution is made. This approach is particularly effective for our dataset as it considers the similarity of entries and provides a statistically reasonable estimate for the missing data.

The choice of five neighbors was determined through preliminary testing, which suggested that it balances the bias-variance tradeoff effectively while also being computationally efficient. After applying kNN imputation, all missing values in ‘MonthlyIncome‘ and ‘NumberOfDependents‘ were successfully estimated, resulting in a complete dataset ready for the subsequent stages of preprocessing and model training.

4.2.2 Correlation Analysis and Feature Engineering. The preprocessing phase included a critical evaluation of predictors through correlation analysis to identify and address collinearity, which can adversely affect model performance and stability [22]. Initially, a high degree of collinearity was observed among the features ’NumberOfTime30-59DaysPastDueNotWorse’, ’NumberOfTime60-89DaysPastDueNotWorse’, and ’NumberOfTimes90DaysLate’, with correlation coefficients between 0.98 and 0.99. To resolve this, these features were consolidated into a single feature named ’NumberOfTimesPastDue’, effectively capturing the essence of the data while reducing redundancy. The original three features were then removed from the dataset.

This consolidation was followed by a reassessment of the dataset’s correlation structure. The subsequent analysis revealed a moderate collinearity of 0.43 between ’NumberOfOpenCreditLinesAndLoans’ and ’NumberRealEstateLoansOrLines’. Given that the RF model was selected for its robustness to collinear predictors and the focus on minimal feature engineering to preserve data integrity and high interpretability, it was determined that further feature manipulation was unnecessary. Additionally, the feature ’Unnamed: 0’ was removed from the dataset as it only represented the index of each entry, deeming it irrelevant for our analysis. Thus, with the high collinearity addressed and the moderate collinearity deemed acceptable, the dataset was finalized for the next stages of modeling (See Figure 2).

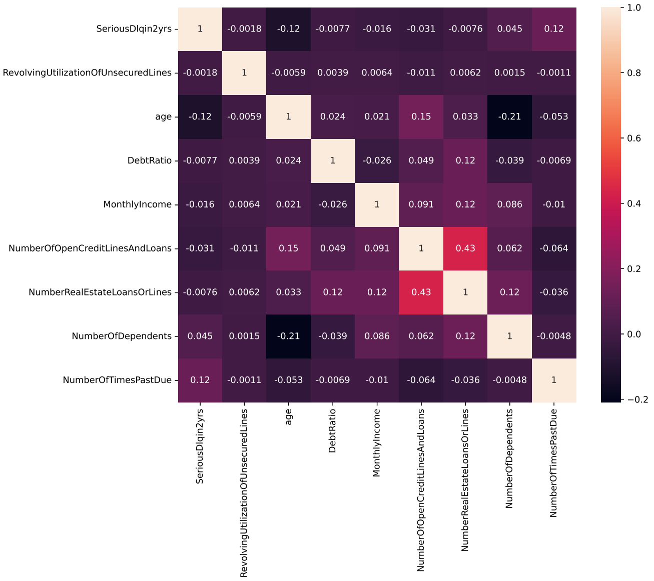


Fig. 2. Correlation Matrix

4.2.3 Balancing the dataset. The initial assessment of the dataset revealed a significant class imbalance, which could lead to biased model training and impact the generalizability of the predictive model. To address this, rather than undersampling the majority class or utilizing penalized models, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. This technique generates synthetic samples for the minority class, thereby enhancing the balance of the dataset [8]. By applying SMOTE, we achieved a balanced dataset consisting of 278, 764 entries with an equal representation of classes, set at a 50/50 ratio. This strategic approach to balance ensures a more robust and fair training process for the RF model.

4.3 Model Implementation

The chosen methodology for the classification task involved the implementation of the RF algorithm, utilizing the `RandomForestClassifier`. The selection of RF was justified by its inherent robustness against overfitting, due to its ensemble approach. Additionally, RF's capacity to handle high collinearity among features renders the presence of moderately collinear variables a non-issue, aligning with our earlier correlation analysis findings.

During the training phase, the RF model was rigorously trained and its performance was quantitatively measured using the Area Under the Curve (AUC) metric [20], on both the training and validation datasets to ensure predictive accuracy.

4.4 Hyperparameter Tuning

The hyperparameter tuning of the RF classifier in the experiment was carefully designed to balance model complexity and predictive accuracy, while also considering computational efficiency. The number of estimators was set to 500 to provide a comprehensive ensemble of trees without excessive computational demand. The maximum depth was set to 10 to ensure that the trees are deep enough to capture relevant data without becoming overly complex. The hyperparameters were carefully tested across a spectrum of values, vigilantly monitoring for any signs of overfitting to preserve the model's generalizability. Remaining hyperparameters were left at their default settings.

The model was trained on a designated training set, and its performance was evaluated on both the training and a separate validation set. The use of a validation set provides a more accurate representation of the model's expected performance on unseen data.

5 RESULTS

The results will be discussed in 2 sections. The first section focuses on evaluating the performance of the RF model and its explainability using LIME and SHAP, emphasizing their role in enhancing transparency in banking decisions. The subsequent section aims to translate the complex results of the RF model into easily understandable formats for customers, emphasizing the educational aspect and aiding in better financial decision-making.

5.1 Model Analysis and Financial XAI

5.1.1 Model Evaluation. The performances of the models were evaluated using both Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves, along with their corresponding Area Under the Curve (AUC) values for each. The ROC AUC metric gives an overview of the model's discrimination ability across all thresholds [4], while the PR AUC focuses on precision and recall [31]. Collectively, they offer a direct and insightful evaluation of the model's classification efficacy, particularly relevant in the context of credit risk assessment datasets where imbalanced data and the accurate identification of the less prevalent class (defaults) are key concerns [11].

Performance metrics, as summarized in Table 2, show a high level of accuracy, with 91.26% of the predictions being correct. It also shows a precision of 90.72%, denoting a high rate of correct predictions and strong precision in identifying true positives. The recall of the model is 91.81%, reflecting its ability to identify the majority of actual positive instances. Lastly, the F1-score, which balances precision and recall, stands at a solid 91.62%, suggesting a well-rounded performance of the model across various aspects of classification success.

Model	Parameter	Performance on Test Data
RF	n_estimators: 500, max_depth: 10	Accuracy: 0.9126 Precision: 0.9072 Recall: 0.9181 F1-score: 0.9126 ROC AUC: 0.97 PR AUC: 0.97

Table 2. Performance metrics of the RF model.

The evaluation of the RF classifier using ROC and PR curves, along with their corresponding AUC values, demonstrates the model's effectiveness in credit risk assessment. The ROC AUC plot (Figure 3) reveals a high score of 0.97, indicating excellent discriminative ability across different thresholds, while the PR AUC plot (Figure 4), with a score of 0.97, suggests a strong precision-recall balance. These curves collectively highlight the classifier's capability in accurately identifying less prevalent but critical positive cases (defaults).

The confusion matrix, visualized in Table 3, illustrates the RF classifier's performance with a high true negative rate of 90.7% (25419), signifying its strong ability to correctly identify the majority of non-default cases. Conversely, the model's true positive rate stands at 91.9% (25460), reflecting its effectiveness in detecting defaults. The model also maintains a low false positive rate of 9.3% (2604), indicative of a modest number of non-defaults incorrectly identified as defaults. The false negative rate is 8.1% (2270), representing the defaults that the model failed to catch.

	Predicted Negative	Predicted Positive
Actual Negative	90.7% (25419)	9.3% (2604)
Actual Positive	8.1% (2270)	91.9% (25460)

Table 3. Confusion matrix for the RF model.

The evaluation of the model demonstrates a good performance, with a high ROC AUC score reflecting strong discriminative abilities and a balanced Precision-Recall indicating effective identification of both classes. Additionally, the F1-score suggests a good balance of precision and recall, underlining the model's robustness in handling the varied demands of the classification task.

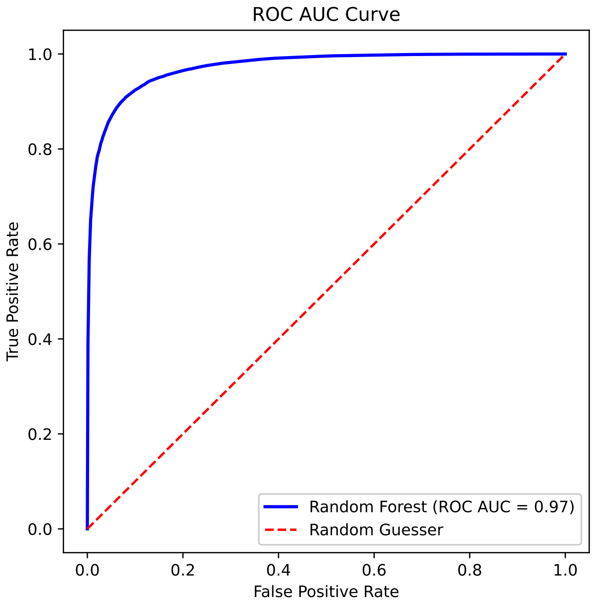


Fig. 3. ROC AUC of the RF model

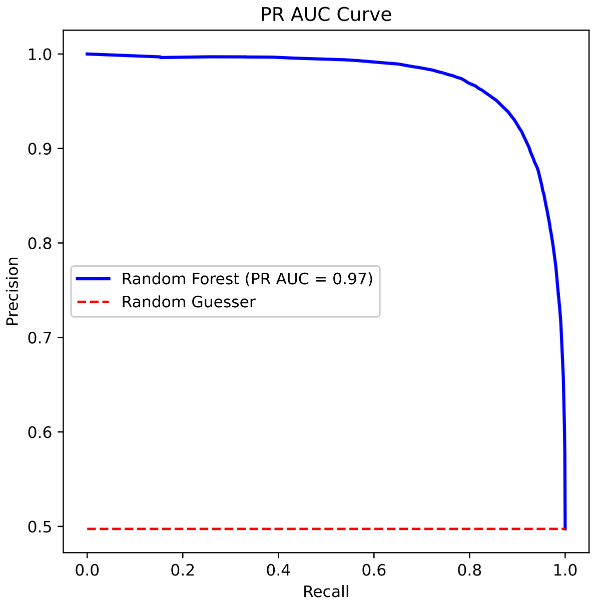


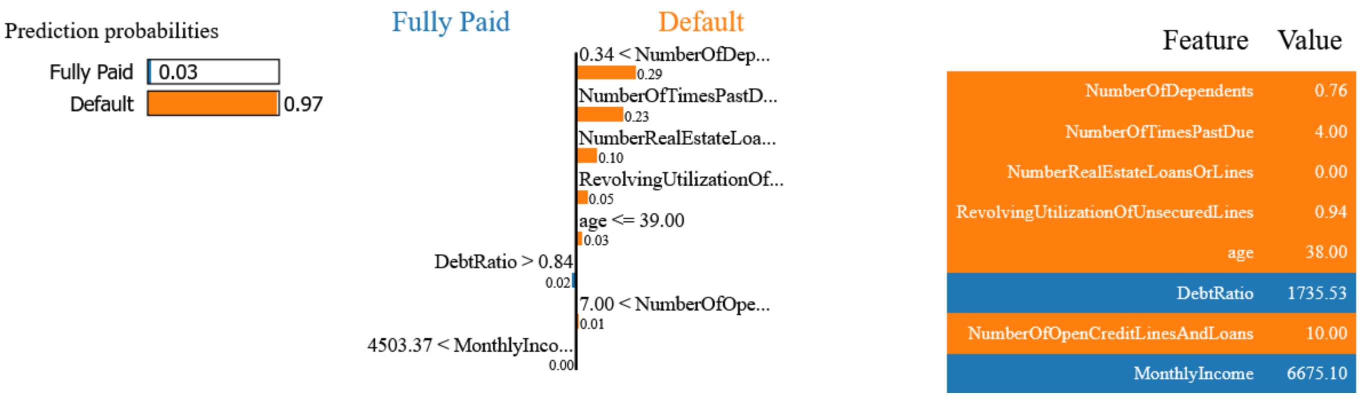
Fig. 4. PR AUC of the RF model

5.1.2 LIME Explainability. In this section, to expand further on the explanations provided by the LIME, we interpret 2 distinct cases as predicted by the model: one leading to a prediction of 'Default' and the other to 'Fully Paid'.

In the case leading to a 'Default' prediction, as illustrated in Figure 5, the model underscores several features with significant impact. Notably, 'NumberOfTimesPastDue' stands out with a high value, which is indicative of frequent late payments—a strong predictor of credit risk. Similarly, a high 'RevolvingUtilizationOfUnsecuredLines' value points to substantial credit usage, which is often associated with increased risk of default. These factors are paramount as they relate directly to the borrower's past behavior and current financial leverage.

Additionally, the model considers 'DebtRatio' and 'NumberOfOpenCreditLinesAndLoans', albeit to a lesser extent. A high debt ratio can signify that a large portion of the borrower's income is dedicated to servicing debt, which could potentially hinder their ability to fulfill new financial obligations. Conversely, numerous open credit lines might reflect a borrower's established credit history but could also imply extended credit obligations.

The LIME analysis also shows the lower impact of 'age', 'MonthlyIncome', and 'NumberOfDependents' in this default scenario. These factors, although less influential in this particular prediction, usually



Feature	Value
NumberOfDependents	0.76
NumberOfTimesPastDue	4.00
NumberRealEstateLoansOrLines	0.00
RevolvingUtilizationOfUnsecuredLines	0.94
age	38.00
DebtRatio	1735.53
NumberOfOpenCreditLinesAndLoans	10.00
MonthlyIncome	6675.10

Fig. 5. LIME explanation for a customer classified as a "Default" Loan type by RF Model

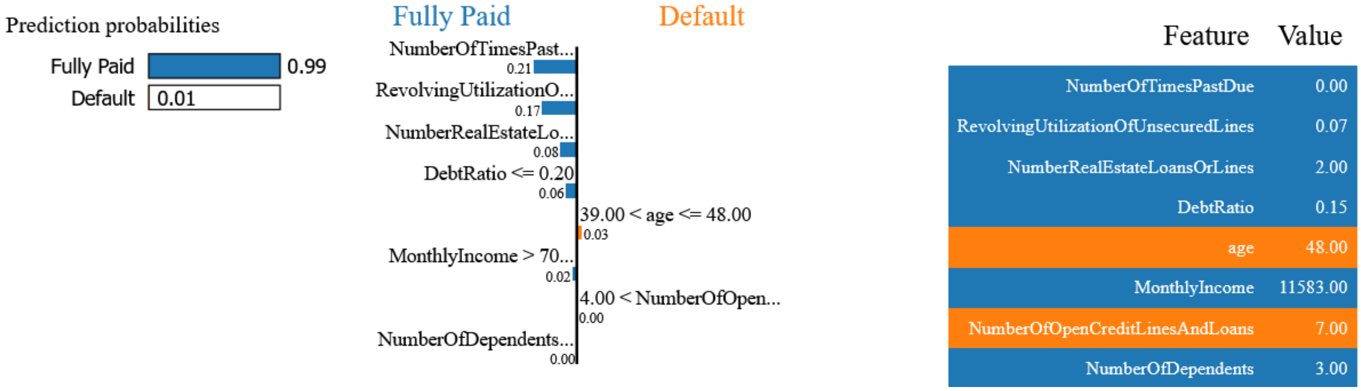


Fig. 6. LIME explanation for a customer classified as a "Fully Paid" Loan type by RF Model

play a critical role in comprehensive credit evaluations. For instance, age may represent financial experience, monthly income indicates repayment capacity, and dependents can reflect additional financial responsibilities.

Conversely, the ‘Fully Paid’ scenario represented in Figure 6 is swayed by positive indicators such as a lower ‘NumberOfTimesPastDue’, which signals a solid payment history, and a more moderate ‘RevolvingUtilizationOfUnsecuredLines’, implying responsible credit usage. Other variables like ‘age’, ‘MonthlyIncome’, and ‘NumberOfDependents’ also contribute to this outcome, painting a picture of financial stability and lower credit risk.

These features’ influence on the prediction outcomes aligns with practical credit assessment principles. A borrower’s age can be indicative of financial maturity, while monthly income and the number of dependents factor into their overall financial obligations. High income and fewer dependents generally suggest a greater capacity to service debts, justifying the model’s prediction of ‘Fully Paid’ in this instance.

Through LIME, we can substantiate the model’s predictions with realistic financial behavior patterns, affirming the interpretability and reliability of our predictive model in assessing credit risk.

5.1.3 SHAP Explainability. SHAP values offer a comprehensive view of feature influence across the entire dataset, highlighting the impact of each feature on the model’s output. For our analysis, we calculated SHAP values on the validation subset of our dataset to ensure an unbiased evaluation of feature importance. The SHAP summary plot derived from validation data, as illustrated in Figure 7, reveals that the top five most impactful features are:

1) ’RevolvingUtilizationOfUnsecuredLines’
2) ’NumberOfTimesPastDue’
3) ’NumberOfDependents’
4) ’NumberRealEstateLoansOrLines’
5) ’age’

The most influential feature, ’RevolvingUtilizationOfUnsecuredLines’, shows a clear pattern where higher values (indicated by red dots) are associated with an increased risk of default. Conversely, lower values (blue dots) correlate with a decreased likelihood of default, underlining the feature’s critical role in financial risk assessments.

’NumberOfTimesPastDue’ is another significant predictor where more instances of past due payments contribute to a higher probability of default. This aligns with typical credit risk models where payment history is a strong indicator of future credit behavior.

Other features such as ’NumberOfDependents’, ’NumberRealEstateLoansOrLines’, and ’age’ also demonstrate important, yet varying, degrees of impact. For instance, a higher ’NumberOfDependents’ may reflect increased financial responsibility, potentially affecting the ability to repay. Meanwhile, ’NumberRealEstateLoansOrLines’ and ’age’ provide additional context to a borrower’s financial situation and risk profile.

In essence, the SHAP summary plot not only quantifies the strength of each feature’s influence on the predictive model but also validates the intuitive understanding of risk factors in credit evaluation.

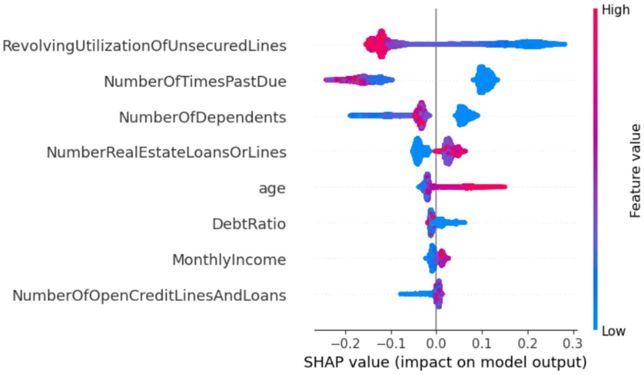


Fig. 7. SHAP explanation visualizing the contributions of features to the predictive outcome.

5.2 Customer Interpretation of Model Insights

For non-technical stakeholders in the finance sector, understanding the implications of ML models on credit risk assessments is crucial but often challenging due to the complex nature of ML algorithms.

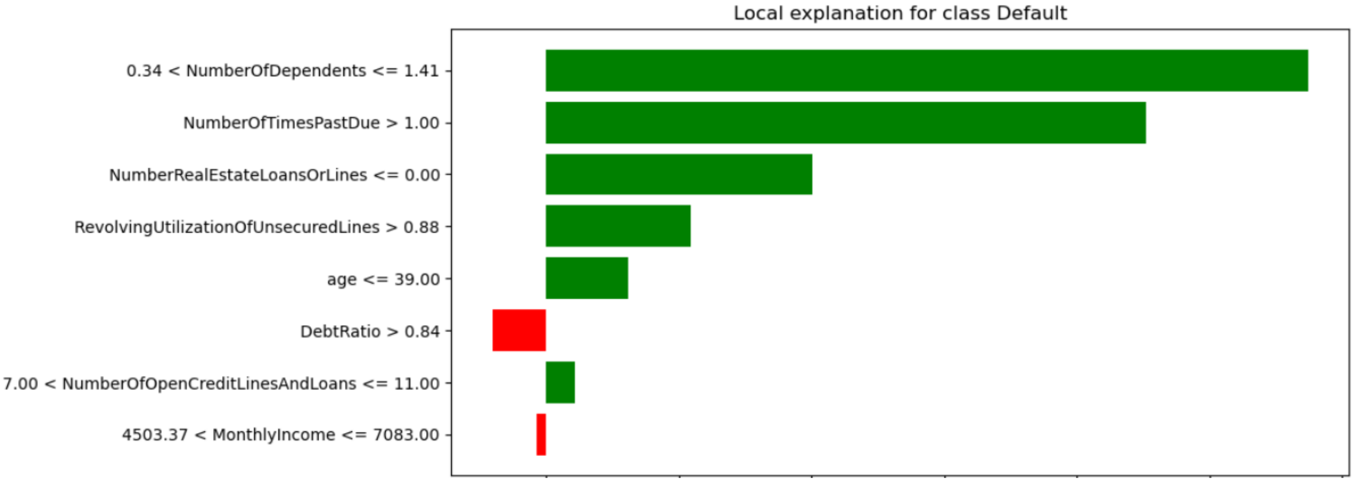


Fig. 8. Local explanation for class Default depicted through a simplified bar chart visualization.

These stakeholders, including business executives, legal and risk audit teams, regulators, and end users, require clear and accessible explanations to build trust in the model's predictions [19].

Based on the work done by Misheva et al. [19], this paper proposes to address some of the varying needs for explainability for external customers in 2 steps:

1) Simplified Visualizations: Employ graphical representations such as bar graphs that rank features by their importance, with color coding to indicate the direction of their impact.
2) Feature Importance Tables: Provide tabular data that ranks features from most to least influential, contextualizing their role in the model's decision-making process.

5.2.1 Simplified Visualizations. For customers who may not be well-versed in the complexities of credit risk analysis and machine learning models, it is essential to provide explanations in a form that is easily digestible and actionable. Simplified visual aids, such as the bar chart shown in Figure 8, can play a pivotal role in this context.

The bar chart offers a clear and concise representation of the factors that the model considers significant in predicting default in this case. Each bar corresponds to a feature, with its length representing the strength of the feature's impact and its color indicating the nature of its influence—positive or negative towards the likelihood of default.

For instance, a longer green bar suggests a strong positive influence increasing the risk of default, as seen with features like 'NumberOfTimesPastDue' and 'RevolvingUtilizationOfUnsecuredLines'. In contrast, a red bar, such as the one for 'MonthlyIncome', indicates a negative or mitigating effect on the risk of default, implying that higher values of this feature could reduce the likelihood of a default. By observing these bars, customers can quickly grasp which aspects of their financial profile have the most significant effect on the model's prediction. Such insights empower them to understand the reasoning behind their credit assessments and potentially take steps to improve their financial health.

5.2.2 Feature Importance Tables. To assist in the interpretation of predictive models, feature importance tables rank the attributes of a customer's financial profile by their influence on the credit decision. These rankings are derived from model explanations, such as the LIME visualization depicted earlier, and are presented straightforwardly for ease of understanding. Below, Table 4 ranks features from the LIME plot by their relevance, along with simple recommendations for improving each feature to potentially enhance credit standing.

Feature	**Relevance**	**Explanation**
NumberOfTimesPastDue	High	Reduce late payments
RevolvingUtilizationOfUnsecuredLines	High	Lower credit utilization
age	Medium	Longer credit history
DebtRatio	Medium	Lower debt-to-income ratio
NumberOfOpenCreditLinesAndLoans	Low	Manageable number of credit lines
MonthlyIncome	Low	Increase income stability
NumberOfDependents	Low	Dependent-to-income balance

Table 4. Ranked feature relevance with simple credit improvement suggestions.

6 CONCLUSION

Answering the first research question, we find that LIME and SHAP methodologies significantly enhance the clarity and transparency of complex ML models for external customers. By breaking down predictions into individual feature contributions, these methodologies demystify the model's decision-making process and provide tangible insights. As demonstrated through the visualizations and discussions in this paper, LIME and SHAP allow customers to see exactly which factors have influenced a credit decision, thereby making the opaque workings of sophisticated ML models more accessible.

Regarding the second research question, effective visualization of credit risk assessment information can be achieved through simplified graphical representations such as bar charts. These visualizations directly map the influence of various features on the model's output, as evidenced by our discussions of Figure 8. Customers can derive insights about their financial behavior patterns and understand the potential impact of each financial indicator on their creditworthiness. This understanding enables them to identify areas of their financial profile that could be improved to potentially enhance their credit ratings.

In conclusion, the application of LIME and SHAP methodologies has proven to be a valuable asset in promoting the transparency and understandability of AI-driven credit risk assessments. By employing these techniques, we can ensure meaningful engagement with external users, empowering them with knowledge and fostering trust in automated decision systems.

6.1 Future Work

Future enhancements can focus on the development of an interactive tool, like a dashboard that will allow customers to engage with the model's predictions more directly. This dashboard will provide a hands-on experience, enabling users to modify feature values and immediately see the impact on their credit risk assessment.

Concurrently, efforts can be made to fine-tune the delivery of explanations and personalized financial advice. This shall involve identifying the most effective ways to communicate complex credit information to customers. Improving these communication strategies, we can help customers make more informed financial decisions and enhance their trust in AI-driven credit evaluations.

ACKNOWLEDGMENTS

The author of the paper would like to thank dr. Marcos Machado and dr. rer. nat. Daniel Braun for their assistance and support during the time of the research project.

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.

[2] Mariana Belgiu and Lucian Drăguţ. 2016. Random forest in remote sensing: A review of applications and future directions. *ISPRS journal of photogrammetry and remote sensing* 114 (2016), 24–31.

[3] Eric Benhamou, Jean-Jacques Ohana, David Saltiel, and Beatrice Guez. 2021. Explainable AI (XAI) models applied to planning in financial markets. (2021).

[4] Andrew P Bradley. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition* 30, 7 (1997), 1145–1159.

[5] Leo Breiman. 2001. Random forests. *Machine learning* 45 (2001), 5–32.

[6] Niklas Bussmann, Paolo Giudici, Dimitri Marinelli, and Jochen Papenbrock. 2020. Explainable AI in fintech risk management. *Frontiers in Artificial Intelligence* 3 (2020), 26.

[7] Niklas Bussmann, Paolo Giudici, Dimitri Marinelli, and Jochen Papenbrock. 2021. Explainable machine learning in credit risk management. *Computational Economics* 57 (2021), 203–216.

[8] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[9] Michael Chui, Bryce Hall, Helen Mayhew, Alex Singla, and Alex Sukharevsky. 2022. The state of AI in 2022—and a half decade in review, McKinsey.

[10] Will Cukierski Credit Fusion. 2011. Give Me Some Credit. https://kaggle.com/competitions/GiveMeSomeCredit

[11] Jesse Davis and Mark Goadrich. 2006. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*. 233–240.

[12] Randall Davis, Andrew W Lo, Sudhanshu Mishra, Arash Nourian, Manish Singh, Nicholas Wu, and Ruixun Zhang. 2022. Explainable machine learning models of consumer credit risk. *Available at SSRN 4006840* (2022).

[13] Murat Dikmen and Catherine Burns. 2022. The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending. *International Journal of Human-Computer Studies* 162 (2022), 102792.

[14] Alberto Fernandez, Francisco Herrera, Oscar Cordon, Maria Jose del Jesus, and Francesco Marcelloni. 2019. Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to? *IEEE Computational intelligence magazine* 14, 1 (2019), 69–81.

[15] Sebastian Fritz-Morgenthal, Bernhard Hein, and Jochen Papenbrock. 2022. Financial risk management and explainable, trustworthy, responsible AI. *Frontiers in artificial intelligence* 5 (2022), 779799.

[16] Jorge Galindo and Pablo Tamayo. 2000. Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications. *Computational economics* 15 (2000), 107–143.

[17] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An approach to evaluating interpretability of machine learning. *arXiv preprint arXiv:1806.00069* (2018), 118.

[18] Alex Gramegna and Paolo Giudici. 2021. SHAP and LIME: an evaluation of discriminative power in credit risk. *Frontiers in Artificial Intelligence* 4 (2021), 752558.

[19] Branka Hadji Misheva, David Jaggi, Jan-Alexander Posth, Thomas Gramespacher, and Joerg Osterrieder. 2021. Audience-Dependent Explanations for AI-Based Risk Management Tools: A Survey. *Frontiers in Artificial Intelligence* 4 (2021), 794996.

[20] Jin Huang and Charles X Ling. 2005. Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering* 17, 3 (2005), 299–310.

[21] Andreas Joseph. 2019. Shapley regressions: A framework for statistical inference on machine learning models. (2019).

[22] Max Kuhn and Kjell Johnson. 2013. Data Pre-processing. In *Applied Predictive Modeling*. Springer New York, New York, NY, 27–59. https://doi.org/10.1007/978-1-4614-6849-3{_}3

[23] Gilles Louppe. 2014. Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502* (2014).

[24] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[25] Praveena Manikonda, Kai Kwan Poon, Courtney Nguyen, and Ming-Hwa Wang. 2020. Explainable machine learning for credit lending. *CMPE257* (2020).

[26] Branka Hadji Misheva, Joerg Osterrieder, Ali Hirsa, Onkar Kulkarni, and Stephen Fung Lin. 2021. Explainable AI in Credit Risk Management. arXiv:2103.00949 [q-fin.RM]

[27] Giuseppe Paleologo, André Elisseeff, and Gianluca Antonini. 2010. Subagging for credit scoring models. *European journal of operational research* 201, 2 (2010), 490–499.

[28] Uday Rajan, Amit Seru, and Vikrant Vig. 2015. The failure of models that predict failure: Distance, incentives, and defaults. *Journal of financial economics* 115, 2 (2015), 237–260.

[29] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " 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*. 1135–1144.

[30] Lloyd S Shapley. 1953. Stochastic games. *Proceedings of the national academy of sciences* 39, 10 (1953), 1095–1100.

[31] Helen R Sofaer, Jennifer A Hoeting, and Catherine S Jarnevich. 2019. The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution* 10, 4 (2019), 565–577.

[32] Swati Tyagi. 2022. Analyzing Machine Learning Models for Credit Scoring with Explainable AI and Optimizing Investment Decisions. *arXiv preprint arXiv:2209.09362* (2022).

[33] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. 2019. Explainable AI: A brief survey on history, research areas, approaches and challenges. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*. Springer, 563–574.

[34] Zhi-Hua Zhou. 2012. *Ensemble methods: foundations and algorithms.* CRC press.