ORIGINAL SCIENTIFIC PAPER
UDK: 005.22:336.717
005.334:336.77
004.89
DOI: 10.5937/EKOPRE2506255R
Date of Receipt: June 6, 2025

Milovan Rankov
Lilly021
University of Belgrade
Faculty of Economics

APPLICATION OF EXPLAINABLE AI METHODS IN THE CREDIT SCORING MODELING

Primena metoda objašnjavajuće veštačke inteligencije na modele kreditnog skoring

Abstract

This paper explores the potential application of explainable artificial intelligence (XAI) techniques to complex deep learning models (DNN) and gradient boosting algorithms (XGB) in the context of credit decision-making. The primary focus is on SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), which contribute to understanding complex models. For empirical analysis, a dataset from Taiwan was used, tracking credit card user behavior and containing information on clients' demographic and financial characteristics, as well as their debt repayment status. The results show that complex machine and deep learning models outperform logistic regression and, when combined with XAI techniques, provide better insights into model operations and enhance transparency. This could, in perspective, lead to broader practical applications, as it facilitates understanding for model users, loan applicants, and regulatory bodies.

Keywords: *Explainable AI, SHAP, LIME, credit scoring, machine learning, deep learning, transparency, interpretability, credit risk, XGBoost*

Sažetak

Ovaj rad istražuje potencijal primene tehnika objašnjavajuće veštačke inteligencije (XAI) na kompleksne modele dubokog učenja (DNN) i algoritme povećanja gradient (XGB) u kontekstu kreditnog odlučivanja. Prevashodno su u fokusu Šeplijeva aditivna objašnjenja (SHAP) i Lokalna interpretabilna model-agnostička objašnjenja (LIME) koji doprinoste razumevanju složenih modela. Za empirijsku analizu korišćena je baza podataka iz Tajvana koja prati ponašanje korisnika kreditnih kartica i sadrži informacije o demografskim i finansijskim karakteristikama klijenata, kao i status otplate duga. Rezultati pokazuju da su kompleksni modeli mašinskog i dubokog učenja dominantniji od logističke regresije i da u kombinaciji sa XAI tehnikama doprinose boljem uvidu u rad modela i povećavaju transparentnost. To može u perspektivi uticati na široku primenu u praksi jer olakšava razumevanje korisnicima modela, tražiocima kredita i regulatornom telu.

Ključne reči: *Objašnjiva AI, SHAP, LIME, kreditni skoring, mašinsko učenje, duboko učenje, transparentnost, interpretabilnost, kreditni rizik, XGBoost*

Introduction

The assessment of creditworthiness, i.e., the evaluation of credit risk for individuals and businesses, represents one of the key operations in the risk management process within any financial institution. Therefore, it is extremely important to use models that ensure high accuracy on one hand, and transparency on the other. Traditionally, financial institutions have relied on statistical models such as logistic regression and decision trees. However, in recent years, with the emergence of artificial intelligence (AI), machine learning (ML), and deep learning (DL), predictive performance has significantly improved, allowing financial organizations to process large volumes of data and uncover complex patterns.

Despite these advancements, a major limitation of AI-based credit scoring models remains their lack of transparency. Many state-of-the-art models—such as deep neural networks (DNNs) and gradient boosting (XGB)—are seen as black boxes by both users and other stakeholders. This creates challenges on multiple levels, especially from the perspective of regulatory compliance (e.g., ethical AI adoption, model validation requirements), user understanding, and ultimately, the trust of loan applicants.

The concept of XAI has been established as a solution to this problem, providing methodologies that enhance model transparency while preserving predictive performance. XAI techniques allow financial institutions to understand the intuition behind ML models, ensuring fairness, accountability, and regulatory compliance. By integrating XAI into credit risk management, institutions can mitigate risks associated with biased decision-making, improve client trust, and promote regulatory confidence.

Several methods are mentioned in the literature that can improve the interpretability of credit scoring models. Techniques such as SHAP, LIME and others offer ways to explain model behavior. These methods help financial analysts and policymakers understand which factors influence credit risk assessment, thereby supporting more informed decision-making.

The main purpose and motivation of this paper is to present the results of applying XAI to credit scoring modeling, comparing various interpretability techniques. The rest of the paper has the following structure: Chapter 2 gives an overview of the existing literature on credit scoring models and XAI methods; Chapter 3 outlines the theoretical foundations of logistic regression, DNN, and XGB models, as well as XAI techniques. Chapter 4 describes the data used in the empirical analysis. Chapter 5 shows a comparative analysis of the different models, and Chapter 6 offers concluding remarks and potential directions for future research.

Literature overview

The increasing application of artificial intelligence (AI) and complex modeling approaches in credit risk management—especially in credit scoring has led to a significant rise in the development and application of explainable artificial intelligence (XAI) techniques. These methods aim to enhance model interpretability while preserving predictive accuracy, which is particularly important in regulated financial environments. XAI plays a vital role in ensuring regulatory compliance (e.g., GDPR, ECB model validation guidelines), improving customer trust, and increasing oversight of automated decision-making processes [8,9, 25].

Credit risk modeling is defined as the process by which financial institutions estimate the probability that a borrower will default on their financial obligations [15]. Traditionally, models such as logistic regression and decision trees have been popular due to their simplicity and interpretability [14, 26]. However, these conventional approaches often fall short in predictive power when compared to more advanced machine learning (ML) models [1]. In contrast, complex ML and deep learning (DL) models demonstrate superior predictive performance, although their opaque structure poses significant interpretability challenges [21].

Among ML techniques, XGBoost (eXtreme Gradient Boosting) has emerged as a leading method in credit scoring due to its accuracy, speed, and robustness to overfitting. It iteratively builds an ensemble of decision trees to minimize the prediction error, making it particularly suitable for structured financial data [6, 16]. Deep neural networks (DNNs) are also widely used because they can

capture highly nonlinear relationships between features and model complex behavioral patterns [17]. However, the lack of transparency in both approaches limits their practical applicability in risk-sensitive environments, hence the growing demand for XAI techniques such as SHAP and LIME.

Two of the most widely used XAI methods in financial modeling are SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). SHAP is based on cooperative game theory and assigns each feature a contribution score that reflects its importance in a prediction, offering global and local interpretability [20]. In contrast, LIME approximates complex models using locally interpretable linear models, enabling analysts to better understand specific predictions [25].

Recent studies confirm the relevance and applicability of these techniques. For instance, SHAP has been successfully applied to interpret credit risk models in retail banking, as shown in the work of Hajelkrem, Olea, and Lange [13], while LIME has proven effective in generating intuitive, case-specific insights in credit evaluation contexts [5]. Moreover, hybrid frameworks that combine multiple XAI techniques have been proposed to improve both interpretability and predictive performance, such as those developed by Moskato, Zhang, and Kumar [23].

In summary, while complex ML models such as XGBoost and DNNs provide substantial improvements in accuracy, their integration into financial decision-making processes requires explainability. XAI techniques such as SHAP and LIME offer a practical solution to bridge this gap, ensuring transparency, regulatory compliance, and stakeholder trust.

Methodology

Credit Scoring models

Logistic regression

Logistic regression (LR) is a statistical method mostly used for classification problems where output variable has binary structure. Unlike linear regression, where continuous values are predicted, LR estimates the probability that a given instance belongs to one of two categories. Its popularity stems from its simplicity, interpretability, and efficiency across a variety of applications.

At the core of LR is the logistic (sigmoid) function, which transforms real-valued inputs into the range [0, 1], enabling interpretation of outputs as probabilities:

$$p = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+ \beta_2 x_2+...+\beta_n x_n)}}$$

Where:

β_0 – the intercept, representing the baseline probability when all independent variables are zero

β_i – coefficients that indicate the influence of individual predictors on the log-odds of the dependent variable. Positive values increase the log-odds, while negative values decrease it

x_j – independent variables (predictors). In credit scoring, these may include income, credit history, debt-to-income ratio, employment status, etc.

The model parameters (β_0, β_1, ..., β_n) are estimated using the method of maximum likelihood estimation (MLE), which calculates equation for coefficients that maximize the likelihood of getting the actual data [7].

The main strength of logistic regression is its interpretability. Coefficients can be directly linked to the odds (the logged one) of the output values, allowing for clear explanation of individual predictors' impact. However, LR also has notable limitations. Its performance relies highly on features of the input data (primarily on their quality) and the proper specification of variables. Moreover, LR struggles to capture complex, nonlinear interactions among features, which can limit its effectiveness in modeling sophisticated financial behaviors.

XGBoost

XGBoost (eXtreme Gradient Boosting) extends the traditional gradient boosting framework by integrating advanced methods for optimization like regularization, weighted quantile sketching, and parallelized computation. These improvements make XGBoost highly efficient and suitable for a various practical use cases. In simplified terms, XGBoost creates an ensemble of decision trees, where each subsequent tree improving previous one by correcting errors.

$$\hat{y}i = \sum\ k = 1^{K} f_k\,(x_i), \quad f_k \in F$$

In this equation, **K** represents the total number of trees, and each f_k predicts an output based on the input x_i. The final estimation $\hat{y}i$ is the sum of all tree outputs. The optimization process of XGBoost aims to maximize a regularized objective function:

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

Here, $l(y_i, \hat{y}_i)$ is the loss function that quantifies the error between the predicted and actual value, while $\Omega(f_k)$ is the regularization term that prevents overfitting and is defined as:

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

In this formulation, **T** shows the number of leaves in a tree, whereas w_j represent the weight of each leaf *j*, and *γ*, *λ* are regularization parameters controlling the model's complexity [6].

From a practical perspective, XGBoost is valued for its exceptionally high predictive accuracy and adaptability across various data types. On the technical side, it offers numerous advantages, including built-in regularization to combat overfitting and robustness to outliers and extreme values.

However, XGBoost is not without limitations. The most prominent drawback is its complexity and lack of transparency. Furthermore, tuning the model such as determining the optimal number of trees, their depth, and the learning rate—requires significant technical expertise and understanding of the algorithm's internal mechanisms.

DNN

Deep Neural Networks (DNNs) are advanced models, which used human brain as inspiration for building the logic behind. The model is made of plenty of different connected layers which are interactively and collaboratively working on input data while generating output results. Each layer introduces an additional level of abstraction, enabling the network to understand and capture complex patterns and relationships in the data. Due to their ability to model high-dimensional feature interactions and process large-scale datasets, DNNs have proven particularly useful in the context of credit scoring [11].

The architecture of a DNN includes an input layer, one or more hidden layers, and an output layer. Each layer contains neurons (or nodes) that apply transformations to the incoming data. Each connection between neurons in the network is assigned with weighs, that are adapted during training process in order to minimize error between predicted and actual outputs [11].

The output of neuron *i* in layer *l* is defined as:

$$h_i^{(l)} = f\left(\sum_j w_{ij}^{(l-1)}\, h_j^{(l-1)} + b_i^{(l)}\right)$$

$w_{ij}^{(l-1)}$ represents the weight of the connection between neuron *j* in the previous layer and neuron iii in the current layer,

$h_j^{(l-1)}$ is the output of neuron *j* in the previous layer,

$b_i^{(l)}$ denotes the bias term of neuron,

f is the activation function applied to the weighted sum, typically a nonlinear function such as the sigmoid or ReLU

DNNs offer several key advantages. They are capable of modeling highly nonlinear and hierarchical relationships in data, automatically extracting relevant features without the need for manual feature engineering. This makes them well-suited for tasks such as credit scoring, where behavioral patterns can be subtle and multidimensional.

However, the use of DNNs presents certain challenges. Training deep architectures requires substantial computational resources and time, especially for large networks. Moreover, the decisions made by DNNs are often difficult to interpret compared to simpler models like logistic regression. Additionally, overfitting can occur—particularly when the model architecture is overly complex or when the training dataset lacks sufficient diversity to generalize effectively.

Model performance evaluation

Evaluating the performance of credit scoring models is essential to assess the accuracy and reliability of predictions. Performance is measured using a variety of metrics that provide insights into a model's classification power, discrimination capability, and generalization to real-world

data. In this study, the following evaluation metrics are used: accuracy, precision, recall, F1 score, area under the ROC curve (AUC-ROC), and the Gini coefficient.

Accuracy

Accuracy indicates the percentage of correctly classified cases relative to the total number of instances:

In the equation below, TP (True Positive) are correctly predicted cases of default, TN (True Negative) correctly predicted cases of regular repayment, FP (False Positive) wrongly predicted bankruptcy and finally FN (False Negative) are wrong predictions that bankruptcy will not occur:

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

Although accuracy gives a general measure of model success, it may be misleading in the presence of imbalanced datasets. For instance, if the majority of clients repay loans on time, a model that always predicts non-default could achieve high accuracy without actually identifying risky borrowers.

Precision and recall

Precision measures the proportion of true defaults among those predicted as risky:

$$\text{Precision} = \frac{TP}{TP + FP}$$

High precision means the model rarely misclassifies solvent clients as defaulters, which is important in scenarios where false positives may lead to unjustified credit denials.

Recall assesses how well the model identifies actual default cases:

$$\text{Recall} = \frac{TP}{TP + FN}$$

A high recall value indicates the model successfully detects risky clients, which is crucial when the cost of missing a default is high—such as issuing loans to non-creditworthy applicants, potentially resulting in financial losses.

F1 Score

The F1 score represents the harmonic mean of precision and recall:

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

It is particularly useful in scenarios requiring a balance between precision and recall, especially when both false positives and false negatives carry significant consequences.

Explainable Artificial Intelligence (XAI)

The term XAI refers to a set of methods which has aim to machine learning and especially complicated deep learning models more understandable, while preserving their advantages, most notably, predictive accuracy. These methods are especially relevant in the context of credit risk management, as they help address challenges related to interpreting, understanding, validating, and verifying model outcomes, particularly in communication with regulatory authorities.

Interpretability, in this context, refers to the extent to which a model's logic and outputs are easy to understand. Models can be categorized into two broad groups:

1. *Intrinsically interpretable models* – where the decision-making logic is transparent by design (e.g., decision trees, linear regression).
2. *Black-box models* – where post hoc interpretation techniques are needed to explain the outputs (e.g., neural networks, ensemble methods).

XAI techniques can also be grouped based on their *scope of explanation*:

- *Global interpretability methods* aim to explain the overall logic and structure of the model, providing insights into how it generally functions across the entire dataset.
- *Local interpretability methods* focus on explaining individual predictions made by the model, answering questions such as "Why was this loan application rejected?"

This distinction is important because different stakeholders require different types of explanations.

- A *credit analyst* may want to understand how the model functions overall to ensure it aligns with risk policies.
- *A bank client* is typically only interested in knowing why their application was rejected and what factors they can improve to qualify in the future.
- *Regulatory bodies* assess models at the macro level and are primarily concerned with transparency, fairness, and compliance, for which global interpretability is more appropriate. Some XAI methods support both global and local interpretability, thereby meeting the needs of all stakeholders.

Another way to classify XAI techniques is by the *type of explanation they provide*:

1. *Feature-based explanations* – which identify the variables that contributed most to a specific prediction.
2. *Example-based explanations* – which use representative samples or similar cases to justify the decision.
3. *Rule-based explanations* – which rely on simplified logical rules to approximate how the model makes decisions.
4. *Visualization-based explanations* – which present model behavior graphically to facilitate intuitive understanding.

Together, these approaches help the most complex deep learning models to become more transparent and reliable—an essential requirement in financial decision-making and credit risk assessment.

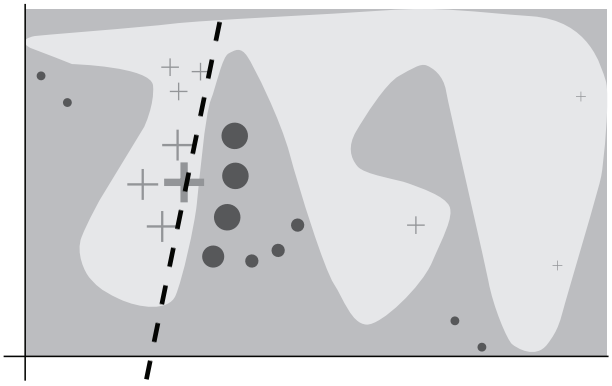
LIME

LIME is an interpretability method used to approximate the behavior of a complex model locally—around a specific prediction. The central idea behind LIME is that any complex model can be approximated by a simpler, more interpretable and easily understandable model (e.g., linear regression) in the local neighborhood of an instance. The underlying assumption is that two nearby input points should produce similar predictions.

LIME works through the following steps:

1. *Perturbation of Input Features*: For a given data instance, LIME generates a set of perturbed samples by randomly altering feature values. These perturbations are drawn from the training data distribution to ensure representativeness.
2. *Prediction on Perturbed Instances*: The complex model (e.g., a DNN) is used to generate predictions for each perturbed instance.
3. *Weighting Based on Proximity:* Weights are assigned to each perturbed instance according to its closeness to the original instance. Instances closer to the original point are considered more relevant and therefore receive higher weights.
4. *Fitting a Simple Model:* A simple and interpretable model (typically linear regression) is trained using the weighted, perturbed instances to approximate the complex model's behavior in the local region.
5. *Generating the Explanation:* The coefficients of the simple model are interpreted as indicators of feature importance, showing which input variables contributed most to the prediction in question[25].

Figure 1: LIME method



From a mathematical perspective, let *f* represent the complex model, and let *x* be the instance for which an explanation is sought. LIME aims to find a simple model *g* that approximates *f* locally around *x*. The optimization is defined as follows [25]:

$$\xi(x) = \arg\min_{g \in G} L\ (f, g, \pi_x) + \Omega(g)$$

Where:

G is the family of interpretable models

$L(f, g, \pi_x)$ is a loss function measuring ability of the interpretable model g approximates the complex model f in the locality defined by πx.

π_x defines the local neighborhood around the instance x.

$\Omega(g)$ is a regularization term that penalizes the complexity of the model g.

LIME is widely adopted in practice because it is model-agnostic—it can be applied to any machine learning or AI model—and provides *local interpretability*, which is highly valuable in domains such as credit scoring. For example, it allows both credit analysts and borrowers to understand the specific reasons behind a rejected loan application.

However, LIME also has notable limitations. Since it relies on approximation, the quality of its explanations depends heavily on the accuracy of the local surrogate model. The perturbation process can be problematic for categorical or high-dimensional data, potentially resulting in misleading or unstable explanations if not managed carefully.

SHAP

SHAP explanations are derived from Shapley values, a concept rooted in cooperative game theory that aims to fairly distribute contributions among players in a coalition. In the context of SHAP, model features are treated as cooperative players, and their contribution to the model's prediction is measured by assessing how the inclusion of each feature changes the output in combination with other features.

The Shapley value for a given feature *i* is computed using the following formula:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f(S \cup \{i\}) - f(S)]$$

Where:

F is the set of all features

S is a subset of features excluding feature *i*

f(S) is the model prediction using only features in subset *S*

ϕ_i represents the marginal contribution of feature *i* to the final prediction, averaged across all possible combinations of features

This formulation ensures fairness, as the contribution of each feature is weighted by its influence on the model output across all potential scenarios.

Several SHAP variants have been developed to optimize computations for specific model types. In this study, we utilize *Deep SHAP*, tailored for deep learning models and combining SHAP with DeepLIFT, and *Tree SHAP*, optimized for tree-based models such as LightGBM and Random Forest.

SHAP has three primary advantages:

1. *Theoretical soundness* – grounded in game theory, providing a rigorous mathematical foundation for fairness
2. *Model agnosticism* – applicable across a broad spectrum of machine learning models
3. *Versatility in explanation* – capable of generating both local (individual prediction) and global (model-wide) interpretability [20]

However, SHAP is not without limitations. One key drawback is its computational intensity, especially with complex or high-dimensional models. Another issue stems from approximation errors, which can occur when estimating marginal contributions. Lastly, although SHAP values explain how a feature influences a prediction, they do not inherently explain why that feature is important, which may hinder intuitive understanding for some stakeholders.

Data

This study utilizes a dataset on the behavior of credit card holders in Taiwan retrieved from Kaggle [16]. The sample consists of 30,000 observations, making it highly suitable for various types of statistical and machine learning analyses. Moreover, the fact that the data is based on real customer records makes it particularly appropriate for bankruptcy prediction—especially in the context of credit card defaults.

The dataset includes a total of 24 variables (or 23 if the ID field—used solely as a unique customer identifier—is excluded). The explanatory variables encompass demographic characteristics, payment behavior, and bill amounts over the previous six months. These features are well-aligned with the binary target variable, which indicates whether the customer defaulted on their next payment or continued to meet their obligations.

Descriptive statistics (see Table 2) show that key variables such as credit limit (LIMIT_BAL), billing amounts (BILL_AMT1–6), and payments (PAY_AMT1–6) exhibit wide value ranges and high standard deviations [16]. This

Table 1: Variable overview

Variable Name	Description	Data Type
ID	Unique client identifier	Numerical
LIMIT_BAL	Approved credit limit (NT dollars)	Numerical
SEX	Client's gender (1 = male, 2 = female)	Categorical
EDUCATION	Education level (1 = graduate, 2 = university, 3 = high school, 4 = other)	Categorical
MARRIAGE	Marital status (1 = married, 2 = single, 3 = other)	Categorical
AGE	Client's age (years)	Numerical
PAY_0	Compensation status 09.2005	Categorical
PAY_2	Compensation status 08.2005	Categorical
PAY_3	Compensation status 07.2005	Categorical
PAY_4	Compensation status 06.2005	Categorical
PAY_5	Compensation status 05.2005	Categorical
PAY_6	Compensation status 04.2005	Categorical
BILL_AMT1	Bill amount in 09.2005	Numerical
BILL_AMT2	Bill amount in 08.2005	Numerical
BILL_AMT3	Bill amount in 07.2005	Numerical
BILL_AMT4	Bill amount in 06.2005	Numerical
BILL_AMT5	Bill amount in 05.2005	Numerical
BILL_AMT6	Bill amount in 04.2005	Numerical
PAY_AMT1	Past settlement 09.2005	Numerical
PAY_AMT2	Past settlement 08.2005	Numerical
PAY_AMT3	Past settlement 07.2005	Numerical
PAY_AMT4	Past settlement 06.2005	Numerical
PAY_AMT5	Past settlement 05.2005	Numerical
PAY_AMT6	Past settlement 04.2005	Numerical
default_payment_next_month	Default status next month (1 = yes, 0 = no)	Categorical

Tabel 2: Descriptive statistics

Variable Name	Observations	Mean	Std. Deviation	Min	Max
ID	30000	15000.5	8660.40	1	30000
LIMIT_BAL	30000	167484.32	129747.66	10000	1000000
SEX	30000	1.60	0.49	1	2
EDUCATION	30000	1.85	0.79	0	6
MARRIAGE	30000	1.55	0.52	0	3
AGE	30000	35.49	9.22	21	79
PAY_0	30000	-0.017	1.12	-2	8
PAY_2	30000	-0.134	1.20	-2	8
PAY_3	30000	-0.166	1.20	-2	8
PAY_4	30000	-0.221	1.17	-2	8
PAY_5	30000	-0.266	1.13	-2	8
PAY_6	30000	-0.291	1.15	-2	8
BILL_AMT1	30000	51223.33	73635.86	-165580	964511
BILL_AMT2	30000	49179.08	71173.77	-69777	983931
BILL_AMT3	30000	47013.15	69349.39	-157264	1664089
BILL_AMT4	30000	43262.95	64332.86	-170000	891586
BILL_AMT5	30000	40311.40	60797.16	-81334	927171
BILL_AMT6	30000	38871.76	59554.11	-339603	961664
PAY_AMT1	30000	5663.58	16563.28	0	873552
PAY_AMT2	30000	5921.16	23040.87	0	1684259
PAY_AMT3	30000	5225.68	17606.96	0	896040
PAY_AMT4	30000	4826.08	15666.16	0	621000
PAY_AMT5	30000	4799.39	15278.31	0	426529
PAY_AMT6	30000	5215.50	17777.47	0	528666
default_payment_next_month	30000	0.2212	0.4151	0	1

indicates the presence of highly diverse financial profiles among clients, from low-risk individuals to heavily indebted customers, and suggests the existence of significant outliers. In many cases, customers show high monthly spending and repayment values, which may reflect active card usage as well as potentially risky financial behavior.

The correlation matrix (Figure 2) highlights the degree of linear association between variables. As expected, strong correlations are observed between monthly payments and corresponding bill amounts, which aligns with the structure of the dataset. However, most other variables exhibit weak linear relationships, suggesting that linear models may not be sufficiently effective, and that more advanced, nonlinear modeling techniques should be considered.

As for the dependent variable, the distribution histogram (Figure 3) reveals a class imbalance: approximately 78% of clients paid their credit card bills on time, while 22% defaulted. This imbalance can significantly affect the performance of classification models, making it necessary to apply balancing techniques such as SMOTE (Synthetic Minority Over-sampling Technique) to improve predictive accuracy.

Results

The comparative analysis of the logistic regression (LR), XGBoost (XGB), and deep neural network (DNN) models is presented using the previously defined performance metrics: accuracy, F1 score, precision, recall, and Gini coefficient

Figure 2: Correlation matrix among variables in the data set.

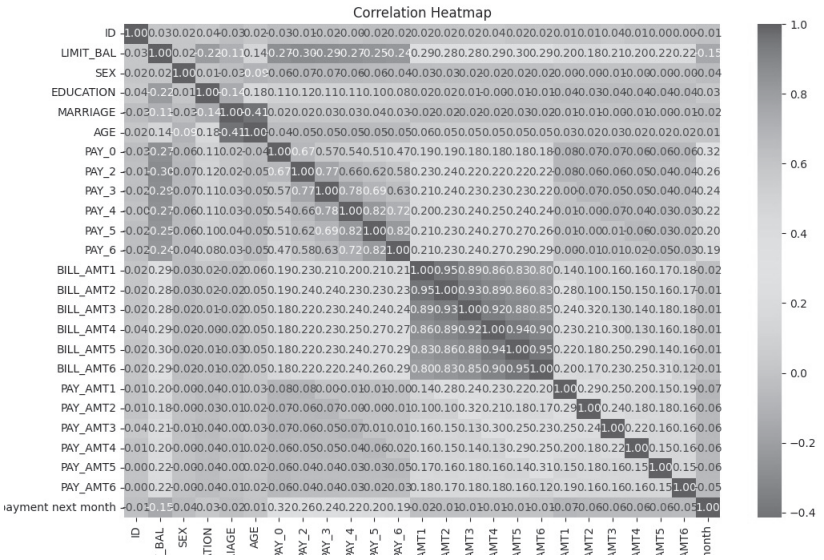
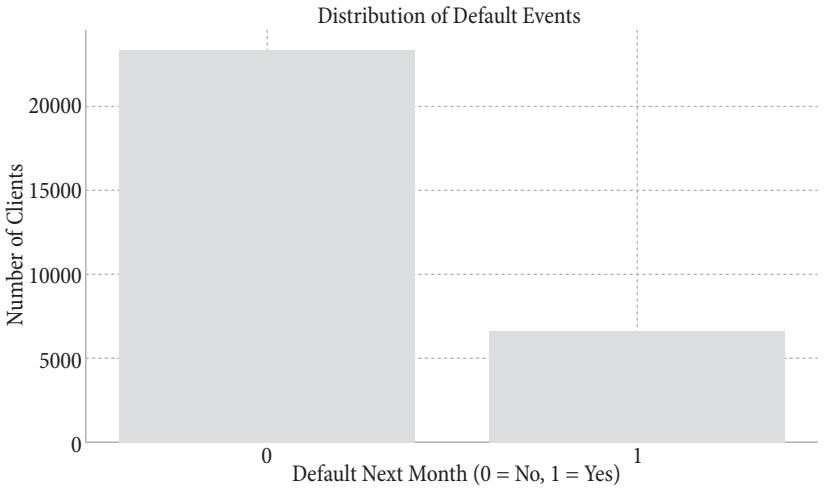


Figure 3: Distribution of defaults



(see Table 3). Based on the results, it is immediately evident that both XGB and DNN models significantly outperform logistic regression across all evaluation criteria. However, their complexity and lack of transparency often limit their adoption in practice. This challenge can be addressed by incorporating explainable AI (XAI) techniques. In this study, we apply LIME for local interpretability and SHAP for global interpretability.

Figure 4 illustrates the results obtained using the LIME method for both the XGB model (left) and the DNN model (right). The purpose of these visualizations is to provide a simplified explanation of the effect of individual features on the final model decision. The length of the bars represents the magnitude of each variable's impact, while the color indicates the direction of influence. Orange bars denote variables contributing to the prediction that a client will default, while blue bars indicate variables associated with a lower probability of default i.e., regular repayment behavior.

In both models, a strong influence is observed from a combination of factors such as PAY_0, PAY_AMT1 (as well as their corresponding values from previous periods),

Table 3: Models comparison

Model	Accuracy	F1 score	Precision	Recall	Gini coefficient
Logistic regression	0.713	0.614	0.728	0.538	0.477
XGBoost	**0.736**	**0.644**	0.752	**0.562**	0.558
Deep neural networks	0.735	0.643	**0.755**	0.559	**0.564**

Figure 4: LIME method-XGB left and DNN right

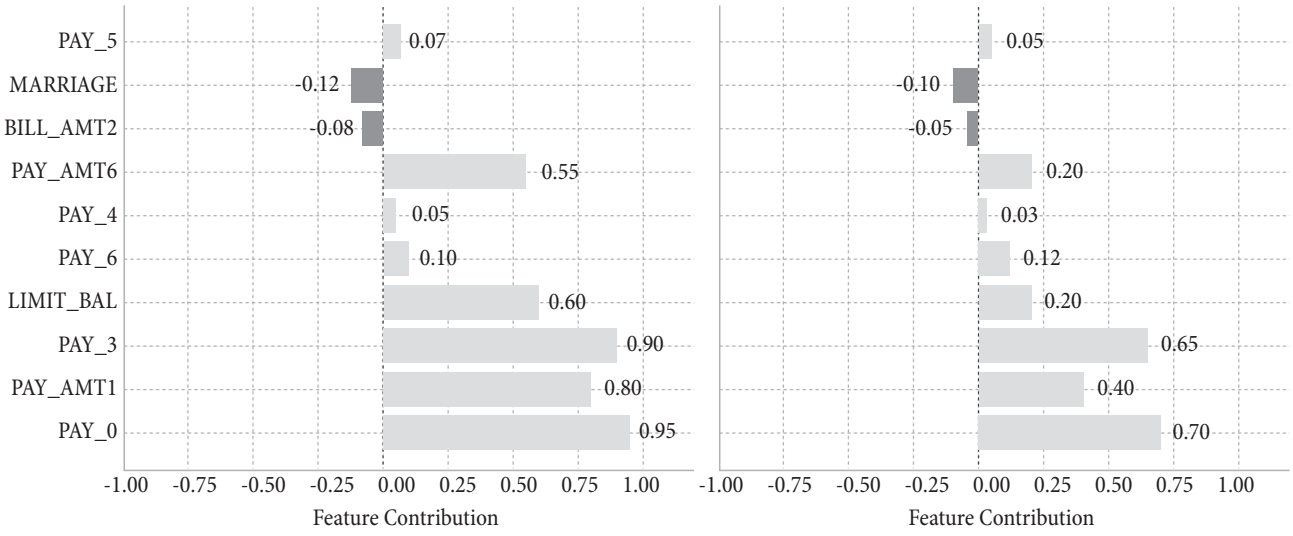


Figure 5: SHAP on XGB

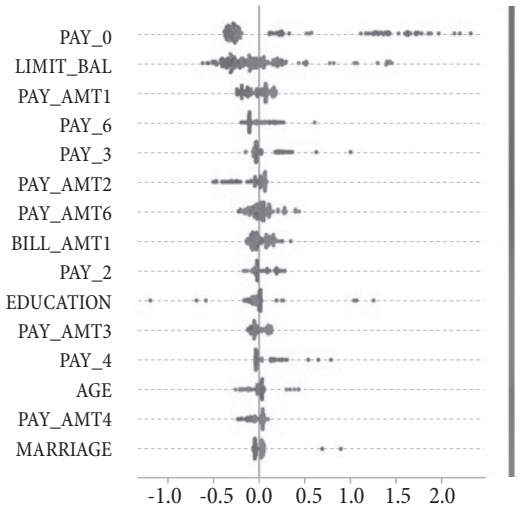
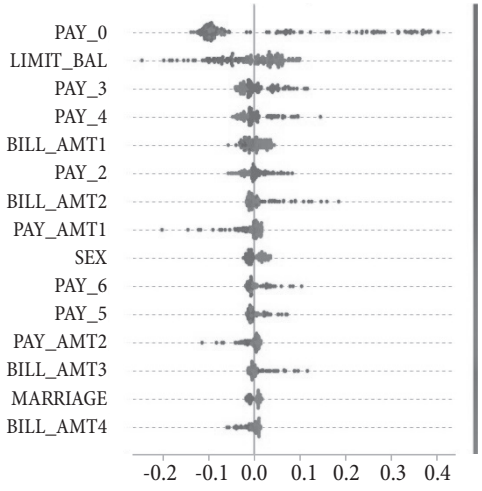


Figure 6: SHAP on DNN



and LIMIT_BAL. All three variables show a negative effect, meaning that higher values are associated with a greater likelihood of default.

As previously noted, LIME's primary strength lies in explaining individual predictions, which is especially valuable for validating specific credit decisions. However, to obtain a broader view and understand the model's behavior globally, SHAP analysis must also be employed.

Figures 5 and 6 display the SHAP value distributions for the XGB and DNN models, respectively. The interpretation of these visualizations is straightforward. Features are ranked on the Y-axis in descending order of importance, while the X-axis indicates the direction of impact. Positive SHAP values suggest the feature contributes to predicting default, whereas negative values suggest the feature supports regular repayment. The color gradient shows the actual value of the feature: red for high values and blue for low values.

Conclusion

This research combines, on the one hand, a comparative analysis of models used for credit risk assessment of credit card holders, and on the other hand, the application of explainable AI (XAI) techniques—specifically LIME and SHAP—on complex models such as XGBoost (XGB) and deep neural networks (DNN), thereby enhancing their transparency and interpretability.

Using standard performance evaluation metrics (F1 score, accuracy, precision, Gini coefficient, among others), XGB and DNN models have, as expected, demonstrated superior predictive power compared to logistic regression. Nevertheless, their complexity presents a barrier to broader adoption. In contrast, logistic regression, although the weakest performer in terms of predictive accuracy, remains a preferred model from the perspective of end users (e.g., banks) and regulatory or supervisory authorities due to its simplicity and interpretability.

A key contribution of this study lies in the implementation of XAI methods, providing a transparent view into model behavior. LIME and SHAP offer local and global interpretability, respectively, and reveal which variables have the most significant influence on model predictions. Based on the analysis, the most influential predictors of individual default are: repayment status in the most recent month (PAY_0), the maximum approved credit limit (LIMIT_BAL), and the billing amount from the previous month (BILL_AMT1).

The combination of advanced modeling techniques and explainable methods confirms that it is possible to achieve both high predictive accuracy and model transparency. This means that machine learning and deep learning models need not remain black boxes—rather, their internal behavior can be decomposed and meaningfully interpreted.

However, there are several potential improvements that could further enhance the robustness, interpretability, and predictive performance of these models. First, incorporating generative AI models such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) could improve data quality, reveal latent patterns, and detect anomalies, all of which can lead to better overall results. Additionally, integrating multiple datasets from different geographic regions, as well as unstructured data, could further increase model robustness and generalizability.

References

1. Agrawal, R., Goyal, N., & Sharma, M. (2023). Advancements in credit scoring using machine learning techniques. *Journal of Financial Analytics, 12*(1), 45–63.
2. Baesens, B. (2014). *Analytics in a big data world: The essential guide to data science and its applications*. Wiley.
3. Baesens, B., Setiono, R., Mues, C., & Vanthienen, J. (2003). Using neural network rule extraction and decision tables for credit-risk evaluation. *Management Science, 49*(3), 312–329.
4. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
5. Busmann, A., Flemming, M., & Rehbein, C. (2021). Local explanation of credit scoring with LIME in practice. *Expert Systems with Applications, 168*, 114236.
6. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).
7. Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological), 20*(2), 215–242.
8. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
9. Đuričin, D. (2009). Uočavanje i upravljanje rizicima u uslovima globalne ekonomske krize. *Ekonomika preduzeća*, *57*(1-2), 1-15.

10. Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.
11. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
12. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182.
13. Hajelkrem, K., Olea, M., & Lange, S. (2023). Explaining credit risk models with SHAP values: A case study in retail banking. *Finance and Technology Review, 6*(2), 89–102.
14. Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A, 160*(3), 523–541.
15. Hull, J. C. (2015). *Risk management and financial institutions* (4th ed.). Wiley Finance.
16. Kaggle. (2022). *Default of credit card clients dataset*. Retrieved from https://www.kaggle.com/datasets/nickzografos/default-of-credit-card-clients/code
17. Lazarević, M. (2017). The treatment of credit risk in post-crisis reform of Basel rules: Modification of the standardised and foundation IRB approach. *Ekonomika preduzeća*, *65*(5-6), 365-380.
18. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.
19. Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM, 61*(10), 36–43.
20. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774.
21. Mester, L. J., Nakamura, L. I., & Renault, M. (1997). Transactions accounts and loan commitments as a source of contagion. *Journal of Money, Credit and Banking, 29*(4), 579–601.
22. Molnar, C. (2022). *Interpretable machine learning – A guide for making black box models explainable*. https://christophm.github.io/interpretable-ml-book/
23. Moskato, A., Zhang, J., & Kumar, R. (2021). Hybrid XAI for credit risk models: Combining SHAP and LIME. In *Proceedings of the International Conference on Artificial Intelligence in Finance*.
24. Rankov, M. (2025). Komparativna analiza modela kreditnog skoringa: konvencijalni vs modeli bazirani na mašinskom i dubokom učenju. *Ekonomske ideje i praksa, 2025*(1), 1–15.
25. Ribeiro, M. T., Singh, S., & Guestrin, C. (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* (pp. 1135–1144).
26. Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002). *Credit scoring and its applications*. SIAM Monographs on Mathematical Modeling and Computation.

Milovan Rankov

is a consultant, certified Financial Risk Manager (FRM), and data scientist with over 12 years of professional experience in the financial industry. He specializes in credit risk modeling, regulatory compliance, and the application of machine learning in banking. He holds an academic degrees from Goethe University Frankfurt, Belgrade University and University of Novi Sad where he focused on quantitative finance and risk management. Milovan has published several papers on explainable AI and advanced modeling techniques, contributing to the academic discourse on transparency in AI-driven credit scoring. His work bridges academic research and real-world application, with a strong focus on aligning innovative data science solutions with regulatory frameworks such as BCBS 239 and GDPR.