sciendo

A Comparative Analysis of Credit Scoring Models and Generative AI Techniques

Andreea-Mădălina BOZAGIU*
Bucharest University of Economic Studies, Bucharest, Romania
**Corresponding author, andreea.bozagiu@fin.ase.ro*

Georgian-Dănuț MIHAI
Bucharest University of Economic Studies, Bucharest, Romania
danut.mihai@fin.ase.r

Andrei Costin NEACȘU
Bucharest University of Economic Studies, Bucharest, Romania
andrei.neacsu@fin.ase.ro

George Alexandru NEACȘU
Bucharest University of Economic Studies, Bucharest, Romania
neacsugeorge13@stud.ase.ro

Abstract. *This paper compares traditional credit scoring methods, deep learning models, and large language models (LLMs), using synthetic data to protect privacy and ensure consistency. Credit scoring has traditionally used methods like logistic regression and new AI models which may improve prediction accuracy. In this paper it was tested and evaluated these baseline methods (logistic regression), deep learning (Gradient Boosting Machine and Neural Networks), and LLM-based models for feature extraction and prediction looking at performance in areas like accuracy, precision, and recall. The results show that deep learning and LLM-based models perform better with complex data, while traditional models still work well with lower computational demands. This paper provides valuable insights into balancing accuracy, interpretability, and computational efficiency when developing credit scoring models.*

Keywords: Modelling Credit Risk, Credit Scoring, Machine Learning, Generative AI.

Introduction

The main purpose for choosing this topic is that selecting an effective credit risk modeling approach is crucial for both financial institutions and banks clients. Since the scorecard plays a key role in assessing creditworthiness and managing risk, it needs to be used as efficiently as possible to ensure accurate and fair decisions. Traditionally, statistical models such as logistic regression and rule-based scoring have been widely used to estimate default probability. With technological advances, more complex methods such as deep learning and boosting models have been adopted to improve the accuracy of predictions. More recently, Large Language Models (LLMs) have begun to be explored for financial tasks, including credit scoring, due to their ability to process large amounts of unstructured data and identify subtle patterns.

This study compares the performance of traditional credit scoring methods, advanced deep learning and boosting models, and LLMs, using synthetic data to ensure confidentiality and consistency of results. The main goal is to determine to what extent these modern techniques bring a real advantage in assessing credit risk and interpreting the importance of explanatory variables. The

DOI: 10.2478/picbe-2025-0098

results show that boosting models, Gradient Boosting Machine, provide the highest accuracy in predicting default probability, balancing performance and interpretability. Deep learning achieves competitive results but suffers from the need for a large amount of data for training and lack of transparency in decision making. As for LLMs, although they can provide insights based on textual data and can be useful in analyzing financial documentation, they still fail to match boosting models in terms of accuracy of scores and clarity in assigning importance to explanatory variables.

These findings emphasize the need for a hybrid approach, where traditional and machine learning-based models are strategically integrated to improve credit scoring while maintaining transparency and decision robustness, but also, should be considered the specific of each bank.

Literature

The literature review identifies the main research studies in the realm of credit scoring, outlining the evolution of methodologies from traditional statistical techniques to advanced artificial intelligence-based approaches. With the advent of the big data era and advances in information technologies, credit data have become increasingly large, complex, and nonlinear. Traditional statistical credit scoring models are unable to cope effectively with such complex datasets. In order to meet the evolving needs of lending organizations, machine learning-based models for credit scoring have been developed. These models include decision trees (DT) (Xia et al., 2018), random forests (RF) (Zhang, He et al., 2018), neural networks (NN) (Desai et al., 1996, West, 2000), support vector machines (SVM) (Pławiak et al., 2019), and gradient boosting decision trees (GBDT) (Xia et al., 2017).

West (2000) assessed the credit scoring performance of five neural network models: multilayer perceptron, mixture-of-experts, radial basis function, learning vector quantization, and fuzzy adaptive resonance. The models are compared to traditional methods, including logistic regressions, k nearest term and so on and so forth. The results show that the mixture-of-experts and radial basis function models overcome the multilayer perceptron in credit scoring tasks, while logistic regression is the most precise benchmark method.

Bensic et al. (2006) focused on credit scoring for small businesses under specific economic conditions, using a relatively limited dataset. The authors compare the accuracy of different models, including logistic regression, neural networks (NNs), and CART decision trees. While statistical tests reveal no significant differences between the models, the probabilistic NN model emerges as the most accurate, with the highest success rate and lowest Type I error. The study identified also key features for credit scoring for small businesses, highlighting the importance of the characteristics of the credit program and the entrepreneur's personal and business traits.

From another perspective, Luo et al. (2017) examined the performance of credit scoring models applied to CDS datasets in the context of post-2007–2008 financial crisis credit risk management. The authors evaluated the performance of deep learning algorithms, in particular deep belief networks (DBN) with Restricted Boltzmann Machines, and compared them with traditional credit scoring models such as logistic regression, multilayer perceptron, and support vector machine. The results show that DBN outperforms the other models, giving the best overall performance.

Dumitrescu et al. (2022) introduced a new credit scoring method called penalised logistic tree regression (PLTR), which combines the interpretability of logistic regression with the predictive power of machine learning decision trees. PLTR allows capturing nonlinear relationships in credit scoring analysis. The findings suggest that PLTR model outperforms logistic regression and performs competitively with random forests.

Regarding Gradient Boosting Decision Trees (GBDT), Liu et al. (2022) addressed its limitations for credit scoring, in particular the lack of diversity in individual classifiers and the performance-interpretability tradeoff. The authors propose two tree-based augmented GBDT models which introduce a step-wise feature augmentation mechanism to enhance classifier diversity while maintaining interpretability. Results on four large-scale credit scoring datasets show that the proposed model outperforms the GBDT accuracy and achieve comparable results to neural networks in feature augmentation efficiency.

Methodology

This study employs three different credit scoring approaches to assess their effectiveness in predicting default probability: Logistic Regression, Gradient Boosting Machine (GBM), and Neural Networks. The models are trained and evaluated using synthetic data to ensure privacy protection and consistency across methods.

The dataset used in this study was generated using Kaggle and consists of 3084 monthly observations. However, LLMs is used for key challenges in this context, particularly in quantifying variable importance and handling structured numerical data effectively.

Logistic regression is a traditional statistical approach used for binary classification problems such as credit default prediction. It estimates the probability of an event occurring using the logistic function:

$$P(Y = 1) = \frac{1}{1+e^{-(\beta 0+\beta 1X1+\beta 2X2+\cdots+\beta kXk)}} \quad (1)$$

where:

P= probability of the event, β_0= intercept, β_1,…..,β_k = the predictors for the independent variables X_0, X_1, …..,X_k

And the logistic regression is the following:

$$\text{TARGET}_i = \beta_0 + \beta_1 \text{PROFESSION\_RECODED} + \cdots + \beta_9 \text{CB\_FICO\_SCORE\_Recoded} + \varepsilon_i$$

The binary variable Y_i (TARGET) represents the status of all clients in the sample, taking a value of **0** if client *i* has defaulted and **1** if the client is non-default.

GBM is an ensemble learning method that builds multiple weak learners (decision trees) sequentially, improving performance at each step by correcting previous errors. The prediction at stage *t* is given by:

$$F_t(X)=F_{t-1}(X)+\eta \cdot h_t(X) \quad (2)$$

where:

$F_t(X)$ = the updated model prediction, $F_{t-1}(X)$ = the previous model prediction, h_t = a new weak learner trained on the residual errors, η = learning rate

Neural networks, particularly multilayer perceptrons (MLPs), use multiple layers of neurons to learn complex relationships between inputs and outputs. A typical feedforward neural network for credit scoring consists of:

1. Input Layer: Features such as credit history, income, and debt levels.
2. Hidden Layers: Neurons apply nonlinear transformations using activation functions (e.g., ReLU, Sigmoid).
3. Output Layer: Uses a sigmoid activation function to output default probability.

The network updates weights using backpropagation and the cross-entropy loss function:

$$L = -\frac{1}{N} + \sum_{i=1}^{N}(y_i \log(p_i) + (1 - y_i)\log(1 - p_i)) \quad (3)$$

where:

y_i= the actual default status, p_i = he predicted probability, N = the number of observations

The Multilayer Perceptron (MLP) network is the most frequently used neural network architecture in commercial applications, including credit scoring. Its ability to model complex, nonlinear relationships makes it highly effective in identifying credit risk patterns that traditional models may overlook.

The MLP architecture for a single hidden layer, with a network consisting of two input neurons and two output neurons, is illustrated in the figure below. In this representation, rectangles represent neurons, while arrows denote the weighted connections in the feedforward network. The four connections without a visible source originate from a bias unit, which has been omitted for simplicity. When an input data value is applied to the input neurons, the computed values at each neuron are propagated layer by layer through the network, ultimately activating the output neurons. This sequential processing allows MLP to learn complex patterns and refine credit risk predictions based on diverse input features.

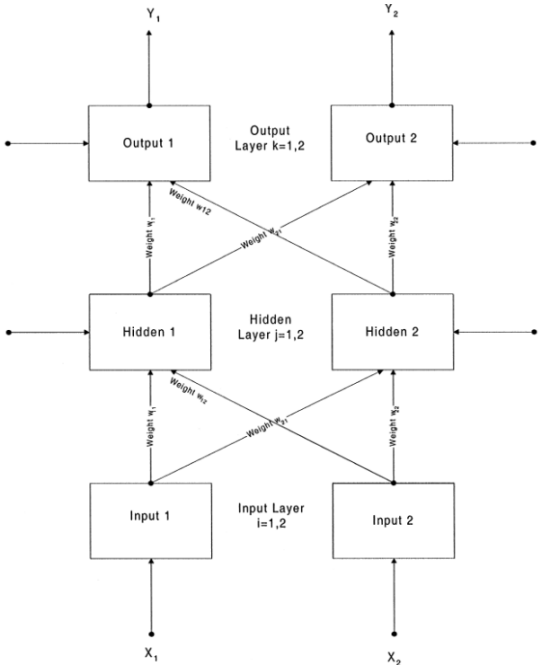


Figure 1. MLP architecture Neural Network

Souce: Authors' own research.

Results and discussions

The first step was the estimation of baseline model, logistic regression, which is represented in Table 1. This table highlights the importance and impact of each variable included in the scoring model. The Weight of Evidence (WoE) indicator measures the predictive power of an explanatory variable in relation to the dependent variable. This indicator is generally described as a measure of how well a variable separates good and bad customers. The term "bad customers" refers to individuals who have partially or fully defaulted on a loan and have been removed from the portfolio (typically, an

active customer is considered bad when the loan becomes non-performing, meaning it has more than 90 days of delay). On the other hand, "good customers" are those who have fully repaid their loan, including interest, or those who are still active with an ongoing loan that has less than 30 days of delay. The formula for the Weight of Evidence (WoE) indicator is given below:

$$WoE_{Gi} = ln\left(\frac{\%\ Clients\ Good_{Gi}}{\%\ Clients\ Bad_{Gi}}\right) \quad (4)$$

where *% Clients Good*$_{Gi}$ represents the percentage of good clients within the total number of clients in group GiG_iGi, and *% Clients Bad*$_{Gi}$ represents the percentage of bad clients within the total number of clients in the same group GiG_iGi.

From a statistical perspective, the Weight of Evidence (WoE) indicator represents the hypothetical scores that should be assigned to each characteristic of a variable, assuming that this is the only available information for developing the credit scoring system. Conventionally, the scoring points are calculated as: *Score=WoE×100*

Table 1. Logit Estimation and Score points

Variable	Category	Estimation	WoE	Same Sign	Score
INCERCEPT	-	0.70569	-	-	70
PROFESSION_Recoded	_1-Known Profession	0.6896***	0.5530	✓	55
	_0-Other Profession	0.56846**	0.1054	✓	11
INCOME_Recoded	_1-Above the minimum wage	0.1389***	0.7996	✓	80
	_0-Below the minimum wage	0.1624*	-0.2744	☒	-27
REVENUE_SOURCE_Recoded	_1-Employed	0.3501***	0.9698	✓	97
	_0-Retired	-0.7745*	-0.1137	✓	-11
AGE_Recoded	_1-Under 50 years old	0.0734***	0.3905	✓	39
	_0-Over 50 years old	0.03823***	-0.1369	✓	-14
MARITAL_STATUS_Recoded	_1-Married	0.22145*	0.1922	✓	19
	_0-Unmarried	-0.2296*	-0.0925	✓	-9
REGION_Recoded	_1-București-IF	0.59246**	0.5697	✓	27
	_0-Other regions	0.7559**	0.5818	✓	58
RELATIVE_INCOME_Recoded	_1-Above-average income	0.17445*	0.3406	✓	34
	_0-Below-average income	-0.1126**	-0.3341	✓	-33
TENOR_CURRENT_EMPLOYER_Recoded	_1-Information available	1.0298**	0.8862	✓	89
	_0-Information unavailable	-1.3566	-0.6307	✓	-63
CB_FICO_SCORE_Recoded	_1-Good score	0.72713**	1.4018	✓	140
	_0-Lower/bad score	0.0000	-0.2908	-	-29

*, **, *** Statistically significant at a **10%**, **5%**, and **1%** error level, respectively.
Source: Authors' own research.

Based on the results in Table 1, it can be concluded that potential clients with a lower or negative *WoE* for one of the two classes in which the variable is grouped will face a disadvantage when applying for a loan, as they exhibit a higher risk of default. This highlights the role of *WoE* in risk differentiation, ensuring that customers with stronger financial stability receive better credit conditions, while those with riskier profiles are flagged accordingly.

The Weight of Evidence values provide a crucial understanding of how different borrower characteristics influence the probability of default. A positive *WoE* indicates a lower likelihood of default, suggesting a safer credit profile, while a negative value indicates a higher risk. One of the most influential factors in reducing default probability is employment status. The data shows that being employed significantly decreases the risk of default (*WoE* = 0.9698), while being retired (WoE = -0.1137) slightly increases it. This finding aligns with the expectation that steady employment provides financial stability, whereas retirees may face more rigid income sources, leading to higher financial vulnerability.

Similarly, income level plays a decisive role in default prediction. Borrowers earning above the minimum wage (*WoE* = 0.7996) exhibit a considerably lower probability of default, while those earning below the minimum wage (*WoE* = -0.2744) face a heightened risk. This suggests that financial constraints make it more difficult for low-income borrowers to maintain consistent debt payments. Relative income, a variable measuring whether a borrower's earnings are above or below the average, reinforces this finding. Individuals with above-average incomes (*WoE* = 0.3404) show lower default rates, whereas those earning below the average (*WoE* = -0.3340) are at a higher risk.

Additionally, variables such as *CB_FICO_SCORE_Recoded* and *REVENUE_SOURCE_ Recoded* show significant differences in default probability between their categories, suggesting that these are highly important factors in credit decision-making. To further enhance the robustness of the scoring system, it is essential that the signs of the estimated coefficients align with those of the *WoE* indicator. This alignment ensures consistency in risk assessment, and it can be confirmed by checking the "Identical Sign" column.

The non-default probability for each client in the sample is:

$$p_i = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X_{i1}+\hat{\beta}_2 X_{i2}+\cdots+\hat{\beta}_n X_{in}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X_{i1}+\hat{\beta}_2 X_{i2}+\cdots+\hat{\beta}_n X_{in}}} \quad (5)$$

The probability of default (PD) is calculated as: PD=1−p_i

For example, a client who is 53 years old, married, and employed will have an estimated probability of non-default (*pi*), calculated using the logistic regression formula:

$$p_i = \frac{e^{0.7057-0.1369+0.1922+0.9698}}{1+e^{0.7057-0.1369+0.1922+0.9698}} = 84.95\% \quad (6)$$

And his PD=1-84.95%=15.05%

As can be seen, the logistic model offers an easy and straightforward method for interpreting data, being intuitive, suggestive, and simple to implement by economists.

The next step is the implementation of the deep learning models and using LLMs models. Before generating the ML models, it was used ChatGPT LLM to observe if this generative AI tool can predict the importance of variables to simplify the data analyzing process.

The output is represented in figure 1 and suggests that the most important variable from the scorecard model is FICO SCORE, meanwhile the less important variable is Relative Income Country.

Table 2. Classification of Variables Based on Predictive Power Using LLMs

Rank	Variable	Predictive Power
1	CB_FICO_SCORE_MAINDEBTOR	High predictive power
2	INCOME_MAINDEBTOR	Strong predictive indicator
3	AGE_MAINDEBTOR	Moderately strong predictor
4	MARITAL_STATUS_MAINDEBTOR	Moderate predictive impact
5	TENOR_CURRENT_EMPLOYER_MAINDEBTOR	Context-dependent
6	REVENUE_SOURCE_MAINDEBTOR	Informative, less direct
7	PROFESSION_MAINDEBTOR	Context-dependent
8	Region	Lower predictive power
9	RelativeIncomeCounty	Low predictive impact

Source: Authors' own research.

By using Large Language Models (LLMs) in scorecard development enhances the classification of variables by predictive importance, allowing for more automated, data-driven insights. While traditional methods rely on statistical techniques like Weight of Evidence (*WoE*) and feature importance from machine learning models, LLMs can complement these approaches by identifying hidden patterns, relationships, and potential interactions within the dataset. However, their effectiveness in structured numerical analysis remains limited compared to boosting models or logistic regression.

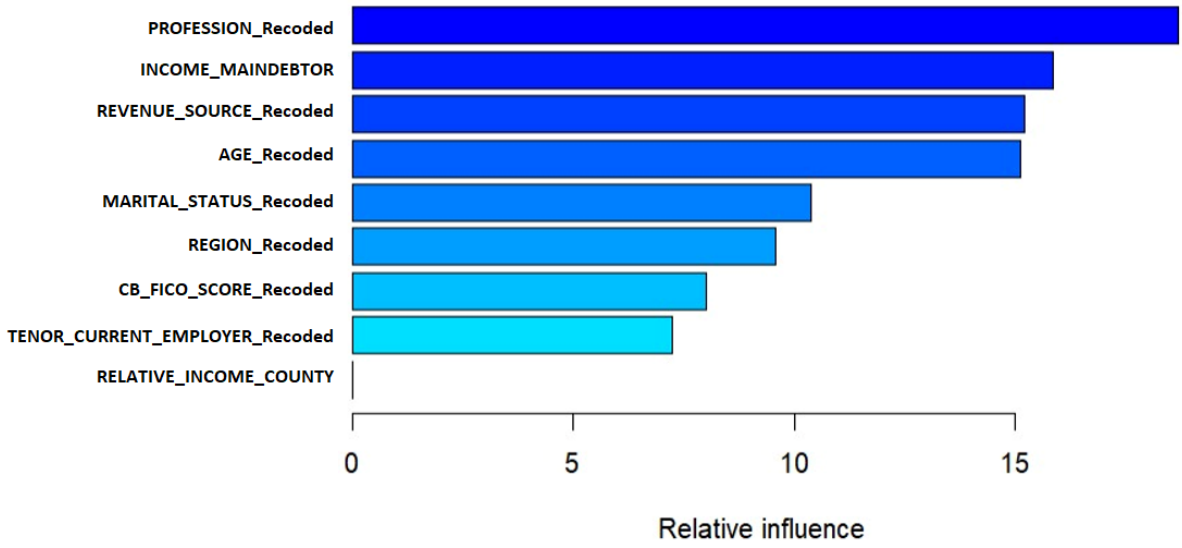
Moreover, LLMs can process unstructured financial data, such as transaction histories, customer interactions, or alternative credit signals, helping to refine risk assessment beyond structured numerical variables. However, despite their flexibility in textual and contextual analysis, their performance in structured numerical modeling remains inferior to boosting models or logistic regression, which are specifically optimized for tabular credit risk data. As a result, LLMs may serve as a supplementary tool rather than a standalone solution, contributing to enhanced feature engineering and qualitative risk assessment in credit scoring models.

Another advantage of LLMs in credit scoring is their ability to adapt and incorporate real-time economic and financial changes into risk assessment. Unlike traditional models that rely on historical data, LLMs can analyze news reports, market sentiment, and macroeconomic indicators to detect potential risks that may impact a borrower's creditworthiness. Additionally, they can assist in automating loan application processing by extracting relevant information from unstructured sources, such as financial statements or customer profiles.

However, due to their limited transparency in decision-making and challenges in quantifying variable importance in structured data, LLMs are best utilized as a complementary tool alongside well-established scoring models like GBM or logistic regression, ensuring a more comprehensive and robust credit risk evaluation.

In order to see if the LLMs provided a right estimation of the variables lets compare this method with Gradient Boosting Machine (GBM) model.

Figure 2. The Influence of Variables on the Dependent Variable TARGET



Source: Authors' own research.

In figure 2 the variable profession has the greatest influence on the score, followed by income level, source and age of the depositor. Based on this analysis, the most important variables to be considered in the credit decision making process can be determined.

Gradient Boosting Machines (GBM) are among the most effective techniques for credit scoring, offering a powerful method to capture nonlinear relationships and complex interactions between variables. Unlike traditional logistic regression, which assumes a linear relationship between predictors and the target variable, GBM iteratively builds decision trees, where each new tree corrects the errors of the previous one. This sequential learning process enhances prediction accuracy and helps identify subtle risk patterns that might be overlooked by simpler models.

One of GBM's key advantages is its feature importance ranking, which allows financial institutions to understand which variables most influence default probability. Unlike LLMs, which struggle with numerical feature importance, GBM provides a clear breakdown of how each variable contributes to risk assessment.

Additionally, hyperparameter tuning (e.g., learning rate, number of trees, and tree depth) enables fine-tuning the model to optimize predictive power while preventing overfitting. However, GBM models require significant computational resources and can be less interpretable than logistic regression, making them less ideal in highly regulated environments where transparency is critical.

The last step in the analyze is representation of Neural Network to compare the deep learning models.

The Multilayer Perceptron (MLP) network is the most used neural network architecture in commercial applications, including credit scoring. Its ability to model complex, nonlinear relationships makes it a valuable tool for assessing credit risk, especially when dealing with large datasets containing intricate patterns.

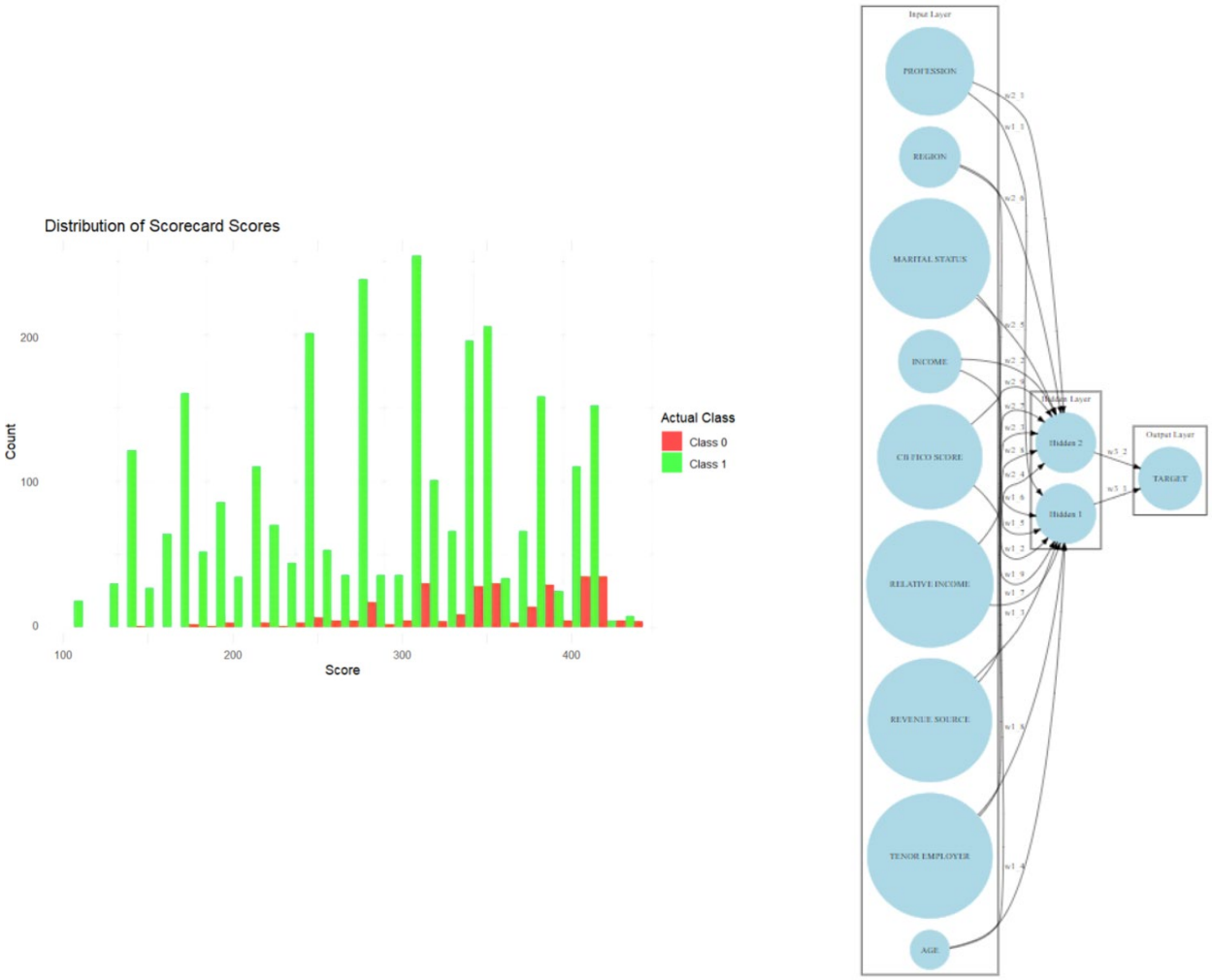


Figure 3. Neural Network outputs

Source: Authors' own research.

The above figures illustrate key aspects of the Neural Network-based scorecard model used for credit scoring. The left-side chart represents the distribution of scorecard scores, where the x-axis denotes the assigned credit scores, and the y-axis represents the count of observations. The colors indicate actual default classes: green for non-default clients (Class 1) and red for default clients (Class 0). The histogram shows that higher credit scores correlate with a lower proportion of defaulting clients, confirming the model's ability to distinguish between risk levels effectively.

On the right side, the diagram represents the Neural Network architecture used in the scorecard model. The input layer consists of key financial and demographic variables such as profession, income, credit score (CB_FICO_SCORE), relative income, revenue source, employment history (tenor employer), region, marital status, and age. These inputs are processed through two hidden layers, where weights are assigned and adjusted during training to optimize classification accuracy. The final output layer predicts the TARGET variable, which indicates whether a client will default or not. The structure suggests a fully connected feedforward network, where each input contributes to the decision-making process, reinforcing the importance of non-linear relationships in predicting credit risk.

Neural networks (NN), particularly Multilayer Perceptron (MLP) models, offer significant advantages in credit scoring by capturing nonlinear relationships and complex interactions between borrower characteristics. Unlike traditional logistic regression models, which assume a linear dependency between input features and default probability, neural networks adapt dynamically to patterns in the data, improving prediction accuracy. This is particularly valuable when dealing with high-dimensional and correlated variables, such as financial behavior, income levels, and credit history. One of the key benefits of NN models in credit scoring is their ability to learn from vast amounts of historical data, refining risk assessment based on complex patterns that may not be apparent in conventional methods. The hidden layers in a neural network enable it to detect intricate dependencies between financial and demographic attributes, making it an effective tool for identifying risky borrowers. Additionally, NN models can be continuously retrained on new financial and economic conditions, allowing for real-time adaptation to market trends and borrower behaviors.

Another major advantage of neural networks is their resilience to multicollinearity—a common issue in credit risk modeling where predictor variables are highly correlated. While traditional methods like logistic regression may suffer from unstable coefficients due to multicollinearity, NN models naturally distribute importance across multiple hidden layers, mitigating this problem. Furthermore, NN models can integrate alternative data sources, such as transactional history, online spending behavior, or even sentiment analysis from financial documents, enhancing the overall predictive power of the credit scoring system. Despite these advantages, interpretability remains a challenge, as NN models function as a black box, making it difficult for financial institutions to justify credit decisions in regulatory environments.

To evaluate the performance of these models, the most used techniques are the ROC (Receiver Operating Characteristic) curve, represented in figure 4, and AUC (Area Under Curve) indicator, table 3.

The ROC curve is a fundamental tool for evaluating model performance in classification problems, as it plots the true positive rate (Sensitivity) against the false positive rate (1 - Specificity) for various decision thresholds. A model with superior discriminative power will have a curve that is closer to the top-left corner of the graph, indicating higher sensitivity at lower false positive rates.

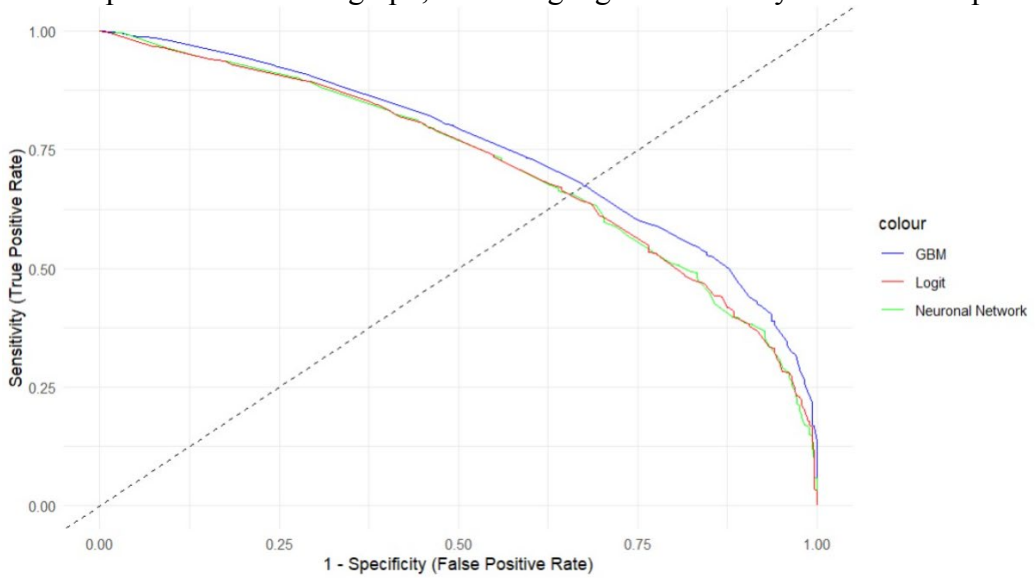


Figure 4. ROC curve

Source: Authors' own research.

Table 3. AUC indicator

Model	AUC
Logit	0.82
GBM	0.94
Neural Network	0.63

Source: Authors' own research.

When comparing Logistic Regression (Logit), Gradient Boosting Machine (GBM), and Neural Network (NN) based on their ROC curves, the objective is to determine which model provides the best ability to differentiate between default and non-default cases. The analysis shows that GBM demonstrates slightly better performance than the other two models, as its ROC curve is positioned closer to the top-left corner, suggesting improved sensitivity for a given false positive rate. This result aligns with the well-known advantage of GBM in capturing nonlinear relationships and complex interactions between features, leading to higher predictive accuracy.

Meanwhile, Logistic Regression and Neural Networks exhibit similar performance, as their ROC curves largely overlap throughout most of the plot. This suggests that, despite the increased complexity of Neural Networks, their classification power remains on par with traditional logistic regression. This may indicate that, in this dataset, the linear decision boundary of Logit is sufficient, and the additional complexity introduced by NN does not necessarily improve predictive ability.

Overall, GBM appears to be the most effective model, but the relatively close performance of the three methods highlights the importance of considering factors beyond ROC, such as interpretability, computational efficiency, and regulatory compliance, when selecting the most appropriate model for credit scoring applications.

As mentioned above, another way to compare the efficiency of the three scorecard models is by using the AUC (Area Under the Curve) indicator. This metric quantifies the percentage of the dataset that is correctly classified between the two outcome categories (0 and 1). A higher AUC value indicates a model with better discriminative power, meaning it can more accurately distinguish between default and non-default cases.

According to the results presented in the previous table, the GBM model outperforms the others, achieving an AUC of 94%, meaning that 94% of the dataset is correctly classified into binary categories. This suggests that GBM provides the most precise separation between default and non-default clients, leveraging its ability to model nonlinear interactions and complex decision boundaries more effectively than traditional methods.

Therefore, based on performance analysis, GBM proves to be the best model among the three considered, as it demonstrates both a slightly superior ROC curve and a higher AUC score. These results confirm GBM's strength in credit scoring applications, making it a preferred choice when predictive accuracy is a primary objective.

Conclusion

This paper provides a comprehensive comparison of three major credit scoring models—Logistic Regression (Logit), Gradient Boosting Machine (GBM), and Neural Networks (NN)—analyzing their strengths, weaknesses, opportunities, and limitations through a SWOT analysis. The results highlight the trade-offs between interpretability, predictive power, computational efficiency, and flexibility in risk assessment, ultimately reinforcing the importance of model selection based on specific business needs and data characteristics.

Logistic Regression remains a widely used approach due to its simplicity, interpretability, and stability. Its clear decision structure makes it easy to implement and explain to regulators and stakeholders. However, its inability to capture complex nonlinear relationships and relatively lower predictive power compared to advanced models limit its effectiveness in more complex datasets. It is best suited as a baseline model or as part of an ensemble approach where transparency is a key requirement.

Gradient Boosting Machine (GBM) demonstrates superior performance in predicting default probability, with higher accuracy and better feature importance evaluation. Its ability to model nonlinear interactions makes it a powerful tool for credit risk assessment. However, longer training times, increased computational demands, and reduced interpretability can pose challenges. GBM is particularly effective in cases where maximizing predictive accuracy is prioritized over transparency, making it an ideal choice for data-driven, high-volume lending decisions.

Neural Networks (NN) offer the highest capacity for capturing complex patterns in data, excelling in large datasets with intricate relationships. Their ability to model nonlinear dependencies makes them promise for advanced credit scoring applications. However, they require extensive computational resources, tuning, and suffer from low interpretability, often being referred to as a "black box" in decision-making. While they are the least restrictive model among the three, their tendency to overfit and the difficulty in explaining predictions pose challenges in financial and regulatory settings. Beyond performance differences, model restrictiveness plays a crucial role in determining which approach is most suitable for credit scoring. Logistic Regression imposes the strictest structure, potentially disadvantaging applicants who do not fit predefined risk patterns. GBM, while more flexible, can still be restrictive depending on the calibration of risk tolerance. Neural Networks offer the greatest flexibility but may overfit the training data, leading to biased credit decisions.

Overall, GBM emerges as the most effective model in terms of predictive accuracy, Logit remains the best choice for transparency and regulatory compliance, while Neural Networks provide the most advanced pattern recognition capabilities. The final model selection should be based on the specific requirements of the credit scoring process, considering interpretability, computational feasibility, and business risk tolerance. Future research could explore hybrid approaches, integrating the strengths of multiple models to enhance both accuracy and transparency in credit risk evaluation.

The analysis confirms that credit scoring models play a crucial role in reducing credit risk, enabling financial institutions to optimize decision-making processes and accurately identify eligible clients for loan approval, thereby maximizing profitability. By leveraging these models, banks can enhance their risk management strategies, improving portfolio quality while ensuring sustainable growth.

Selecting the most appropriate scoring model requires aligning the choice with business objectives, considering factors such as model restrictiveness and ease of implementation. A highly restrictive model may lead to a decline in loan approvals, potentially reducing sales volumes by excluding clients who could successfully meet their financial obligations. Therefore, financial institutions must strike a balance between risk mitigation and business expansion, ensuring that the chosen scoring model effectively supports both credit risk assessment and profitability goals.

Although the results confirm the efficiency of machine learning-based models in credit scoring, limitations related to interpretability, computational requirements, and decision transparency remain significant challenges. Future research could explore the hybrid integration of these models

with traditional approaches, as well as the use of explainable AI (XAI) techniques to enhance the comprehensibility and acceptability of decisions in the regulated financial environment.

References

Bensic, M., Sarlija, N., and Zekic-Susac, M. (2006). Modelling small-business credit scoring by using logistic regression, neural networks and decision trees. *Intelligent Systems in Accounting, Finance and Management*, Vol. 13(3), 133-150

Desai, V.S., Crook, J.N., and Overstreet Jr, G.A. (1996). A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research*, Vol. 95(1), 24-37

Dumitrescu, E., Hue, S., Hurlin, C., and Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, Vol. 297(3), 1178-1192

Liu, W., Fan, H., and Xia, M. (2022). Credit scoring based on tree-enhanced gradient boosting decision trees. *Expert Systems with Applications*, Vol. 189, 116034

Luo, C., Wu, D., and Wu, D. (2017). A deep learning approach for credit scoring using credit default swaps. *Engineering Applications of Artificial Intelligence*, Vol. 65, 465-470

Plawiak, P., Abdar, M., and Acharya, U.R. (2019). Application of new deep genetic cascade ensemble of SVM classifiers to predict the Australian credit scoring. *Applied Soft Computing*, Vol. 84, 105740

West, D. (2000). Neural network credit scoring models. *Computers & Operations Research*, Vol. 27(11-12), 1131-1152

Xia, Y., Liu, C., Da, B. and Li, Y. and Liu, N. (2017). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications*, Vol. 78, 225-241

Xia, Y., Liu, C., Da, B. and Xie, F. (2018). A novel heterogeneous ensemble credit scoring model based on stacking approach. *Expert Systems with Applications*, Vol. 93, 182-199

Zhang, H., He, H., and Zhang, W. (2018). Classifier selection and clustering with fuzzy assignment in ensemble model for credit scoring. *Neurocomputing*, Vol. 316, 210-221