Vol. 12 No. 3, 2026
Baltic Journal of Economic Studies
DOI: https://doi.org/10.30525/2256-0742/2026-12-3-299-311

CREDIT SCORING MODELS IN BANK CREDIT RISK MANAGEMENT: COMPARATIVE ASSESSMENT AND APPLICATION IN UKRAINIAN BANKING PRACTICE

Mykhailo Baraniuk[1], Andrii Hrabariev[2]

Abstract. Banking is increasingly shaped by expanding data volumes, more complex borrower behaviour, and stricter credit risk management requirements. Under such conditions, scoring models are becoming especially relevant as instruments for the formalised assessment of creditworthiness, combining analytical accuracy, speed of decision-making, and the possibility of integration into the bank's risk management system. The study compares traditional and modern scoring models in bank credit risk management and proposes an approach to their practical use in Ukrainian banking. Its focus is on scoring models as instruments for credit risk assessment. The study combines comparative analysis, matrix modelling, simulation, statistical modelling, and machine learning methods. Given limited access to primary banking information and confidentiality requirements, the empirical analysis was conducted on a synthesised demonstration dataset designed to reflect the structure of a real retail credit portfolio. For the analysis, a sample of 1,000 observations with a default share of 22.0% was constructed, and logistic regression, discriminant analysis, Random Forest, XGBoost, and a hybrid logit + ML re-ranking model were used for comparison. The results showed that XGBoost provided the highest predictive accuracy, with an AUC-ROC of 0.861, Gini of 0.722, Recall of 0.781, and Brier score of 0.141, whereas logistic regression demonstrated an AUC-ROC of 0.781 and retained advantages in terms of interpretability and suitability for validation. The hybrid model achieved an AUC-ROC of 0.848, Gini of 0.696, Recall of 0.773, and Brier score of 0.144, thus ensuring the best balance between accuracy, explainability, calibration, and practical applicability. Practically, the study offers an adaptive approach to selecting scoring models and a matrix for evaluating them under Ukrainian banking conditions, taking into account the requirements of the regulatory environment, data quality, and the instability of the operating conditions of Ukrainian banks.

Keywords: credit risk, bank risk management, credit scoring, machine learning, hybrid models, model interpretability, probability of default estimation, banking validation, model governance, Ukrainian banking practice.

JEL Classification: C45, C53, G21, G32

1. Introduction

Traditional expert-based approaches to borrower creditworthiness assessment no longer provide the level of objectivity, speed, and adaptability required for current banking decisions. The growing volume of data, increasing complexity of customer behaviour, tighter regulatory requirements, and the need to account for probability of default and expected credit losses call for a shift from predominantly intuitive or formal-analytical judgments to more accurate, standardised, and quantitatively grounded credit risk assessment tools. In this context, scoring models have become particularly important in modern bank risk management, as they formalise borrower assessment, rank clients by risk level, improve the quality of lending decisions, and reduce subjectivity in underwriting.

Scoring models are valuable because they convert borrower characteristics into a measurable risk signal that can be used both at loan origination and in subsequent portfolio monitoring. Their use contributes to more accurate default prediction, improved customer risk segmentation, optimisation of portfolio structure,

[1] Kyiv National Economic University named after Vadym Hetman, Ukraine
E-mail: baraniuk.m.r@gmail.com
ORCID: https://orcid.org/0009-0008-6684-5037
[2] Kyiv National Economic University named after Vadym Hetman, Ukraine;
Academician of Academy of Economic Sciences of Ukraine, Ukraine *(corresponding author)*
E-mail: andr.grab@gmail.com
ORCID: https://orcid.org/0000-0001-6165-0996

and more efficient management of reserves and capital. At the same time, the spread of digital technologies and machine learning tools creates new opportunities to enhance scoring approaches while also increasing the need for their critical reconsideration, comparative analysis, and adaptation to the specifics of the national banking environment.

Against this background, the paper examines scoring models as instruments of bank credit risk management, evaluates their practical strengths and limits, and develops a comparative framework for Ukrainian banking practice. In practical terms, credit risk occupies a central place among all types of banking risk because it is directly linked to the bank's core active operations, namely the allocation of resources for profit generation. Unlike many other risks that may have an indirect or situational effect, credit risk is inherent in the very nature of banking intermediation, and the extent to which it can be controlled largely determines the overall effectiveness of a banking institution.

2. Literature Review and Development of Research Hypotheses

Recent research treats scoring models not simply as screening tools, but as a central element of bank credit risk management. Thomas, Crook, and Edelman (2002) examined credit scoring and behavioural scoring as instruments supporting lending decisions, credit limit adjustments, and subsequent borrower monitoring. Later research expanded this view, increasingly interpreting scoring as a mechanism that transforms customer information into a formalised risk signal for lending decisions.

A related strand of research focuses on the informational basis of scoring models. Muñoz-Cancino, Bravo, Ríos, and Graña (2023) showed that credit history, social interaction, and behavioural variables improve creditworthiness assessment, especially under information asymmetry. This suggests that modern scoring extends beyond purely financial indicators and increasingly relies on broader datasets that capture the borrower's risk profile more accurately. Pang, Hou, and Xia (2021) further demonstrated that scoring models are practically valuable when they support risk classification, probability of default estimation, and expected loss calculations. At the same time, Crook, Edelman, and Thomas (2007) emphasised that the development of scoring is driven not only by predictive accuracy, but also by the standardisation, speed, and reproducibility of credit analysis. Accordingly, the literature increasingly positions scoring as an instrument of formalised credit risk management.

Another strand of research addresses the classification of scoring models by stage of the credit process. Singh, Kumar, Srinivasa, Maini, Ahuja, and Jain (2021) supported the distinction between application scoring, behavioural scoring, collection scoring, and fraud scoring. Alves and Dias (2015) showed the particular value of behavioural scoring for dynamic risk monitoring, while Dornadula and Geetha (2019) highlighted the growing role of fraud scoring in digital lending. Taken together, these studies suggest that model effectiveness depends not only on the algorithm itself, but also on how well it fits a specific function within the credit cycle.

A central issue in the literature is the comparison between traditional statistical models and modern machine learning algorithms. Classical approaches such as logistic regression, discriminant analysis, probit models, and scorecards remain important because of their interpretability and regulatory acceptability. However, Banasik and Crook (2007) showed that traditional models may suffer from sample selection bias and reject inference problems, which encouraged the development of more advanced models capable of handling nonlinear relationships and heterogeneous data. In this context, Gambacorta, Huang, Qiu, and Wang (2024) found that machine learning and non-traditional data can significantly improve predictive performance. Chen, Calabrese, and Martin-Barragan (2024) showed that high predictive power must be combined with adequate explainability, while Sun, Zhang, Li, and Wang (2023) demonstrated that effective scoring models in high-stakes settings should operate within a human-in-the-loop framework. This is especially important in banking, where credit decisions have analytical, legal, and regulatory dimensions.

Taken together, these studies point to four hypotheses relevant to bank credit risk management: H1 – machine learning models deliver higher predictive performance in credit risk assessment than traditional statistical models; H2 – model interpretability is a necessary condition for practical use in bank credit risk management; H3 – the inclusion of behavioural, transactional, and non-traditional data improves credit risk prediction; H4 – the effectiveness of a scoring model increases when it is aligned with a specific managerial function within the credit process. Overall, the literature clearly shows that the development of credit scoring lies at the intersection of three interrelated dimensions – accuracy, explainability, and managerial relevance – which together form the conceptual basis of this study.

3. Research Methodology

The methodology combines general scientific and specialised methods to analyse scoring models as instruments of bank credit risk management. The theoretical part draws on analysis, synthesis, abstraction, generalisation, and logical structuring to systematise scholarly approaches to credit risk, clarify the role of

scoring models in bank risk management, and refine their classification features.

Traditional statistical approaches and modern machine learning models were compared using the comparative method, which allowed scoring models to be assessed against a set of criteria relevant to banking practice, including predictive accuracy, interpretability, validation complexity, data quality requirements, regulatory acceptability, and adaptability to an unstable environment. This shifts the analysis from a narrow algorithmic comparison to a broader functional assessment of model use in credit risk management.

The evaluation framework was developed using decomposition, the systems approach, and matrix modelling. This methodological combination helped structure the evaluation criteria into interrelated blocks and treat a scoring model not as an isolated analytical algorithm, but as part of an integrated mechanism of credit decision-making, internal control, and regulatory-driven risk management.

The empirical part of the study is based on simulation modelling, since the scoring approaches were tested on a synthesised demonstration dataset constructed according to the logic of a real retail loan portfolio. This design preserves the confidentiality of primary banking information while allowing different model types to be tested on a common methodological basis. The empirical block combines statistical modelling and machine learning methods, including logistic regression, discriminant analysis, Random Forest, XGBoost, and a hybrid approach that links an interpretable base model with additional algorithmic re-ranking in the zone of uncertainty.

To ensure methodological reliability, the study applies time-based sample splitting, cross-validation, out-of-time testing, as well as class balancing and probability calibration techniques. These procedures made it possible to assess the models not only in terms of formal classification accuracy, but also with regard to their stability, ability to retain predictive power over time, and suitability for use in a changing environment.

The results were evaluated using a combination of economic-statistical and qualimetric approaches and a system of indicators covering discriminatory power, classification quality, sensitivity to the default class, and calibration quality. This framework ensured a multidimensional assessment of scoring models consistent with the objectives of the study and the logic of bank credit risk management.

4. Main Findings and Discussion

The comparison in Table 1 suggests a more nuanced conclusion than a simple contrast between "old" and "new" scoring approaches. Traditional models are advantageous where transparency, regulatory acceptability, relative ease of implementation, and low maintenance costs are decisive. Modern ML models, by contrast, are more appropriate when a bank operates with large, heterogeneous, and dynamic data, requires higher predictive accuracy, and is prepared to invest in more complex validation and explanation mechanisms. No model is universally superior. The choice of a scoring approach should depend on the type of credit portfolio, the bank's objectives, the quality and completeness of the data, the institution's level of digital maturity, and the nature of regulatory requirements. The more defensible option is to build a flexible scoring architecture rather than privilege a single algorithm in all settings.

The analysis of traditional and modern scoring models shows that comparing them solely by accuracy metrics or by their formal classification into particular algorithmic groups is insufficient for drawing conclusions about their actual suitability for bank credit risk management. Contemporary research increasingly argues that the evaluation of credit models should go beyond a narrowly predictive perspective and also address explainability,

Table 1

Comparative characteristics of traditional and modern scoring models in the credit risk management system

Criterion	Traditional models (logit, discriminant analysis, probit, scorecards)	Modern ML models (DT, RF, GBM, XGBoost, LightGBM, NN, ensembles)
Predictive accuracy	Stable, but often lower with complex data	Often higher, especially in the presence of nonlinearities and a large number of features
Interpretability	High	Medium to low; often requires explainability tools
Data quality requirements	Moderate, but sensitive to model specification and feature selection	High; require quality data preparation, balancing, and feature engineering
Sensitivity to an unstable environment	More vulnerable to structural breaks	Often better adaptability, but with a risk of instability when the data-generating process changes
Implementation speed	Higher	Lower due to tuning, testing, and governance
Validation complexity	Relatively low	Higher, especially for ensembles and neural networks
Regulatory acceptability	High	Depends on explainability and model governance
Maintenance cost	Lower	Higher
Integration into automated systems	Good	Very good, but technologically more complex

Source: Compiled by the authors

model governance, validation, technological feasibility, and integration into real financial decision-making processes. Studies on explainable AI in financial risk management (Ripa Akter, 2026) show that predictive performance alone is insufficient if a model cannot be interpreted, audited, and used in a controlled way in high-stakes settings.

Against this background, the study proposes a matrix for evaluating scoring models in bank credit risk management, allowing them to be compared not as isolated technical solutions, but as functional elements of risk-oriented banking management. Methodologically, this means that a scoring model should be judged not only by its discriminatory power, but also by how well it fits the credit process, internal control, provisioning logic, and an unstable operating environment. Accordingly, the proposed matrix integrates five interrelated blocks: predictive, interpretative, data and technical feasibility, regulatory and practical, and adaptability to Ukrainian banking conditions.

The first, predictive block covers model accuracy, forecast stability, and the ability to identify defaulting borrowers. These criteria are included because, in credit risk management practice, it is important not only to achieve high AUC, accuracy, or recall on historical samples, but also to preserve the model's analytical usefulness over time. A model that performs strongly on a test sample but quickly loses discriminatory power when conditions change cannot be considered a reliable banking tool. Research on early warning systems in credit risk clearly shows that the practical value of a model lies in its ability to support timely detection of adverse changes and enable preventive action, rather than merely deliver formally high classification rates.

The second, interpretative block includes transparency of model logic, the ability to explain individual decisions, and suitability for internal control and audit. Its inclusion is essential because, in banking, a credit decision has not only a statistical but also a legal, reputational, and regulatory dimension. For this reason, a model that is slightly weaker in accuracy but significantly stronger in explainability may in many cases be of greater practical value than a complex "black box" that is difficult to verify. Recent work on Explainable Artificial Intelligence (XAI) in finance (Ripa Akter, 2026) emphasises that explainability increases trust in a model, facilitates its use in high-stakes decision-making, and makes it suitable for both internal and external oversight.

The third block, data and technical feasibility, covers data volume requirements, sensitivity to missing values and noise, update complexity, and compatibility with a bank's IT infrastructure. It is included because, in practice, the effectiveness of a scoring model depends not only on its theoretical quality but also on its ability to function reliably in a real information environment. For a bank, it is crucial whether a model requires large homogeneous datasets, how sensitive it is to data drift, how difficult it is to retrain, and whether it can be integrated into automated decision-making systems. These considerations often become decisive when choosing between statistically simpler models and more technologically demanding ML approaches.

The fourth, regulatory-practical block evaluates compliance with banking supervisory requirements, suitability for IFRS 9 purposes, ease of validation, and acceptability within a bank's internal credit policy. This block is important because a banking model must simultaneously meet the requirements of risk management, financial reporting, and model risk control. Research on the impact of IFRS 9 on banking risk and provisioning (Yadira Salazar, Paloma Merello, Ana Zorio-Grima, 2023) confirms that the transition to the expected credit loss framework has increased the importance of models capable of early detection of credit quality deterioration and integration into provisioning calculations. Thus, practical suitability depends not only on analytical strength, but also on how easily a model can be documented, validated, and normatively justified.

The fifth block, adaptability to Ukrainian conditions, includes resilience to structural breaks, suitability for use in an unstable macroeconomic environment, sensitivity to changes in borrower behaviour, and applicability under a limited historical data base. Its inclusion is conceptually important for this study because many models described in the international literature were developed using data from relatively stable markets. By contrast, the Ukrainian banking system is shaped by environmental volatility, sharp structural breaks, shifts in customer solvency, interruptions in time series, and uneven data quality. Studies of structural shifts in credit ratings and of the impact of macroeconomic scenarios on credit modelling confirm that model usefulness is not constant over time; therefore, adaptability should be treated as an independent evaluation criterion.

On this methodological basis, the study proposes a matrix (Table 2) for comparing scoring models using criteria directly relevant to banking practice.

The proposed matrix has not only practical but also methodological value. Its key advantage is that it shifts the analysis from a narrow technical comparison of models to a multidimensional assessment of their actual suitability for banking use. Accordingly, this study refines the comparative approach to scoring models by evaluating them not only through accuracy metrics, but also through a broader set of criteria covering interpretability, technological feasibility, regulatory compatibility, and adaptability to an unstable environment. At the same time, the proposed matrix differs from traditional classificatory and review-based approaches in that it directly accounts for whether a model can be used within a bank's credit risk management system under Ukrainian conditions.

Table 2

Proposed matrix for evaluating scoring models in the bank credit risk management system

Evaluation block	Criterion	Criterion content	Practical significance for the bank	Recommended scale
Predictive	Accuracy	The model's ability to correctly classify borrowers by risk level	Quality of the initial credit decision	High / medium / low
Predictive	Forecast stability	Preservation of the model's discriminatory power over time and under changes in the sample	Reliability of use in a production environment	High / medium / low
Predictive	Detection of defaulting borrowers	The model's ability to identify risky clients in a timely manner	Reduction in the share of problem loans	High / medium / low
Interpretative	Transparency of model logic	The degree to which the internal structure of the model is understandable	Explanation of decision logic for risk management	High / medium / low
Interpretative	Explanation of individual decisions	The ability to justify the result for a specific client	Defensibility of the decision before the client, audit, and supervision	High / medium / low
Interpretative	Suitability for audit	The possibility of independent verification and reproduction of model results	Model risk control	High / medium / low
Data and technical feasibility	Data volume requirements	The required scope and completeness of historical information	Feasibility of building and maintaining the model	High / medium / low
Data and technical feasibility	Sensitivity to missing values and noise	The model's robustness to poor-quality or incomplete data	Reliability in a real banking environment	High / medium / low
Data and technical feasibility	Update complexity	The resource intensity of retraining and recalibrating the model	Maintenance cost	High / medium / low
Data and technical feasibility	Technological compatibility	The possibility of integration into the bank's IT systems	Automation of the credit process	High / medium / low
Regulatory-practical	Compliance with supervisory requirements	Consistency with model governance and banking control requirements	Regulatory acceptability	High / medium / low
Regulatory-practical	Suitability for IFRS 9	The possibility of using model results within the ECL logic	Provisioning and financial reporting	High / medium / low
Regulatory-practical	Ease of validation	Simplicity of checking stability, accuracy, and assumptions	Reduction of model risk	High / medium / low
Regulatory-practical	Acceptability for internal policy	Compliance with the bank's credit procedures and rules	Consistency with the credit management system	High / medium / low
Adaptability to Ukrainian conditions	Resilience to structural breaks	The model's ability to remain useful under abrupt environmental changes	Practical reliability in an unstable economy	High / medium / low
Adaptability to Ukrainian conditions	Suitability in an unstable macroenvironment	Sensitivity to shocks, cyclicality, and crisis-related changes	Relevance for the Ukrainian banking system	High / medium / low
Adaptability to Ukrainian conditions	Sensitivity to changes in borrower behaviour	The model's ability to reflect shifts in payment discipline patterns	Early detection of deterioration in credit quality	High / medium / low
Adaptability to Ukrainian conditions	Applicability under a limited historical data base	The model's ability to function with fragmented data	Feasibility of use under Ukrainian conditions	High / medium / low

Source: Compiled by the authors

In Ukrainian banking, the use of scoring models is constrained by high macroeconomic volatility, prolonged structural breaks, uneven credit information quality, and shifting borrower behaviour. Unlike markets characterised by relatively long stable time series and gradual credit cycle changes, the Ukrainian banking system has repeatedly undergone abrupt phases of risk reconfiguration over the past decade, while the full-scale war has further intensified this instability. In its Financial Stability Reports, the National Bank of Ukraine explicitly identifies the war as the key risk to the financial system, while also pointing to the persistence of corporate credit risk and the expected deterioration of the retail risk profile despite ongoing portfolio clean-up. Scholarly studies on structural shifts in bank ratings and the evolution of ML in bank risk management

likewise stress that, in an unstable environment, lending standards, risk signals, and even classification patterns are not constant over time. As a result, a model effective under one macroeconomic regime may quickly lose value under another.

For Ukrainian banks, therefore, the challenge of applying scoring models lies not only in choosing between classical statistical and modern ML approaches, but above all in ensuring the stability of input information and preserving the analytical validity of the model under an unstable data-generating process. Classical scoring models, built on the assumption of relative stability in historical patterns, have clear advantages in terms of transparency, ease of validation, and regulatory acceptability. In domestic practice, however, their usefulness is often limited because historical data may no longer reflect the current state of risk. When samples include interruptions, shocks of different directions, changes in customer structure, or recalibrated borrower behaviour, the coefficients of classical models may begin to reflect outdated risk relationships rather than the current risk profile. Recent studies on credit scoring and explainable AI in credit assessment confirm that, under such conditions, it is insufficient to evaluate a model solely by accuracy; its ability to maintain stable forecasts, adapt to drift effects, and remain manageable and explainable in lending decisions must also be taken into account.

A separate challenge for Ukrainian banks concerns the use of historical data. The issue lies not only in the volume of information, but also in its heterogeneity, fragmentation, and limited comparability across periods. The full-scale invasion has altered the geography of business activity, household income structures, patterns of consumer and investment behaviour, and the operating conditions of entire sectors of the economy, thereby changing the statistical meaning of many variables traditionally used in scoring. The National Bank notes that the war interrupted the previous decline in non-performing loans, although from 2023 onward the quality of new lending gradually improved and the NPL ratio resumed its decline, falling to 30% by the end of 2024. This means that the training sample of a Ukrainian bank often combines pre-war, crisis, and post-shock recovery data and therefore requires not merely the accumulation of observations, but regime-based reinterpretation, segmentation, and regular recalibration.

Another important factor is the impact of inflationary shocks, war-related risks, and overall macrofinancial instability on the very nature of credit risk. For a scoring model, this means that traditional predictors of borrower solvency may change not only in strength but even in the direction of their effect. Factors that serve as markers of reliability in a stable environment may lose their discriminatory power during a prolonged shock, while short-term behavioural signals, business liquidity, or cash flow resilience may become more important. In its macroeconomic reports, the National Bank of Ukraine recorded inflation accelerating to 12% in December 2024, the persistence of war-related risks as the main source of uncertainty, and the dependence of the credit market on uneven economic recovery. Research on credit risk under economic instability likewise stresses that macro shocks alter not only default levels but also the very structure of risk relationships; therefore, models without a scenario-based or adaptive component quickly lose practical value.

Under such conditions, the quality of credit histories and changes in customer economic behaviour become especially important. For Ukrainian banks, the problem is not only that some historical records may be incomplete or distorted by crisis events, but also that borrowers' behavioural patterns themselves have changed substantially. Shifts in employment, migration, business relocation, changes in income sources, operational disruptions, the growing role of state support programmes, and changing demand for credit products all mean that historical behaviour is not always a reliable predictor of future behaviour. For this reason, behavioural scoring in domestic practice must become not merely a monitoring tool, but an adaptive one: the model should be sensitive to new behavioural signals and respond quickly to changes in payment discipline. Studies on early warning systems in banking and explainable ML for credit risk assessment show that, in an unstable environment, the greatest practical value lies not in classifiers fixed once and for all, but in models with regular feedback, enhanced monitoring, and the ability to explain changes in the risk profile.

This leads to another essential point: for Ukrainian banks, not only the initial construction of the model matters, but also the organisation of its regular revalidation. In stable jurisdictions, a model may retain operational usefulness for a long time with periodic monitoring. In Ukrainian conditions, where credit risk is shaped by war events, inflation waves, changes in the currency structure of the funding base, uneven sectoral recovery, and new state credit support instruments, revalidation must be much more frequent and deeper. This involves not only technical checks of AUC or Gini, but also monitoring variable stability, revising cut-off thresholds, testing drift effects, updating segmentation, and, where necessary, reassessing the model concept itself. This is consistent both with the contemporary international literature on credit model governance and with the logic of IFRS 9, where the forward-looking assessment of expected credit losses strengthens the need for regular updates of risk parameters and scenario information.

To summarise, different scoring approaches vary in their suitability for Ukrainian banks. Classical models, especially logistic regression and scorecards, remain important because of their transparency, relative ease

of validation, and convenience for internal control, making them particularly useful where the regulatory clarity of a credit decision is critical. Their effectiveness, however, declines when the historical data base is fragmented and behavioural patterns change rapidly under external shocks. Modern ML models, including ensemble and boosting algorithms, are potentially better suited to identifying complex nonlinear relationships and working with heterogeneous data, but they require a stronger information infrastructure, more frequent revalidation, advanced explainability tools, and a mature model governance framework. For this reason, the most justified option for Ukrainian banks is not a choice between "classical" and "modern" models, but the development of an adaptive hybrid scoring system in which interpretable models provide a transparent core architecture, while more complex algorithms enhance predictive power in segments with greater uncertainty. This makes it possible to view hybrid scoring not as a compromise between traditional and modern approaches, but as a more mature stage in the development of credit risk management that meets the demands of digitalisation, regulatory accountability, and heightened environmental instability. This is the practical value of the study: scoring in Ukrainian banks should be seen not as a static algorithm, but as a dynamic analytical mechanism that must continuously adapt to changes in the environment, data quality, and borrower behaviour.

Given that information on banks' credit operations, borrowers' financial condition, debt-servicing parameters, and individual credit histories is subject to restricted access and protected both by banking secrecy and by the special legal framework governing credit histories, the practical illustration of building and comparing scoring models in this study is based on a demonstration dataset. The Law of Ukraine "On Banks and Banking Activity" (2001) defines the protection regime for banking information, while the Law of Ukraine "On the Organisation of Formation and Circulation of Credit Histories" (2005) establishes the legal basis for the formation, use, and protection of credit history information. In addition, the Credit Register of the National Bank of Ukraine operates as a regulated element of the credit infrastructure, and banks' risk management systems must comply with the NBU's requirements for risk management organisation.

Accordingly, the study uses a synthesised demonstration dataset constructed according to the logic of a real retail bank loan portfolio. This approach is methodologically justified because it reproduces the structure of input variables used in credit scoring, demonstrates a consistent procedure for building and comparing several models, and allows the proposed evaluation matrix to be tested without breaching confidentiality requirements. At the same time, it should be emphasised that the results presented below are illustrative in nature and should not be interpreted as actual empirical results of a specific bank or of the Ukrainian banking sector as a whole.

Within the demonstration example, a hypothetical sample of 1,000 observations was constructed, each representing an individual loan application or borrower credit exposure at the date of initial scoring assessment. The sample structure was designed to reproduce the logic of a typical retail bank portfolio: the variable set includes the client's socio-demographic characteristics, income parameters, debt burden indicators, credit history variables, credit product characteristics, collateral availability, and contextual variables simulating the impact of an unstable macroeconomic environment (Table 3). The target variable is DEFAULT, where the probability of default increases with a higher debt burden, a greater number of delinquencies, higher credit exposure, lack of collateral, and the presence of a shock regime in the environment. This variable takes the value of 1 in the event of borrower default within the defined observation horizon and 0 in the case of proper debt servicing. In the demonstration sample, the share of default observations was set at 22.0% (220 cases), while non-default observations accounted for 78.0% (780 cases), making it possible to simulate the moderate class imbalance typical of credit scoring practice.

Because scoring tasks in banking are time-sensitive, the sample was split not randomly but in line with the temporal sequence of observations. A time-based split was used, dividing the data into three subsets: a training sample of 60% (600 observations), a validation sample of 20% (200 observations), and a test sample of 20% (200 observations). This design mirrors a practical setting in which a model is trained on historical data, calibrated on an intermediate segment, and tested on observations from a subsequent period. In addition, an out-of-time test was applied, with the last time segment treated as a separate validation subset simulating model testing on a new flow of applications. This is particularly important for assessing the stability of scoring models, as it makes it possible to evaluate not only classification quality on a mixed test sample, but also the preservation of predictive power under temporal changes in data structure.

To improve the reliability of the results, five-fold cross-validation was applied within the training sample. This reduced the risk that quality estimates would depend on a random subset configuration and made it possible to test model robustness to variation in the training data. For logistic regression and discriminant analysis, the cross-validation loop was used mainly to verify coefficient stability and overall discriminatory power, whereas for Random Forest, XGBoost, and the hybrid model it also served to optimise hyperparameters. The procedure therefore involved repeated training, tuning, and validation rather than a single algorithm run.

Table 3
Structure of the demonstration dataset for building scoring models

Variable name	Notation	Variable type	Description
Observation identifier	ID	Nominal	conventional application/borrower number
Borrower age	AGE	Quantitative	full years at the date of application
Employment tenure, months	JOB_TENURE	Quantitative	length of employment
Monthly income, UAH	INCOME	Quantitative	verified or estimated income
Debt burden	DTI	Quantitative	debt-to-income ratio
Number of active loans	ACTIVE_LOANS	Quantitative	current credit obligations
Maximum historical delinquency, days	MAX_DPD	Quantitative	longest delinquency in the credit history
Number of delinquencies over 12 months	DPD_12M	Quantitative	number of delinquency cases
Loan amount, UAH	LOAN_AMT	Quantitative	application amount
Loan term, months	TERM	Quantitative	loan maturity
Collateral availability	COLLATERAL	Binary	1 – yes, 0 – no
Product type	PRODUCT_TYPE	Categorical	consumer / card / secured
Regional risk level	REGION_RISK	Categorical	low / medium / high
Shock environment indicator	SHOCK_DUMMY	Binary	1 – shock period, 0 – baseline period
Target variable	DEFAULT	Binary	1 – default, 0 – non-default

Source: Compiled by the authors as a demonstration set of variables for the testing of scoring models

Particular attention was paid to class imbalance, which is typical of most credit scoring tasks. In the demonstration sample, this issue was addressed through two procedures. First, class weighting was applied to logistic regression, Random Forest, and XGBoost, so that the default class received a higher weight during training. Second, SMOTE was additionally used for the ML models as a method of synthetically balancing the classes in the training sample. At the same time, the test and out-of-time samples remained unbalanced in order to evaluate model quality under conditions closer to a real credit portfolio. This helps avoid inflated performance estimates while preserving sensitivity to default detection.

Model calibration was also a critically important part of the demonstration testing. In this study, it was treated as a separate stage following the training of the baseline algorithms. No special post-calibration was applied to logistic regression, since the model directly produces an estimate of conditional default probability. For discriminant analysis, Random Forest, and XGBoost, the probability outputs were additionally calibrated using Platt scaling or isotonic regression, depending on which method provided a better balance between the Brier score and the stability of the calibration curve on the validation sample. Calibration was needed not only to improve the accuracy of probability forecasts, but also to make model results more interpretable within the logic of PD estimation, which is essential for credit risk management.

A separate set of hyperparameters was specified for each model and optimised on the validation sample. Logistic regression was estimated in a regularised form with L2 regularisation, where the regularisation coefficient was selected within the cross-validation loop. Discriminant analysis used a standard linear discriminant function without additional tuning parameters, but with prior variable standardisation and multicollinearity control. For Random Forest, the selected parameters were 300 trees, a maximum depth of 8, a minimum of 20 observations per leaf, and a maximum number of split features equal to $\sqrt{p}$, where p is the number of input variables. XGBoost was configured with 250 trees, a learning rate of 0.05, a maximum depth of 4, subsample of 0.8, colsample_bytree of 0.8, a minimum split gain of 0.1, and L2 regularisation of 1.0. Although these settings are demonstrative, they reflect a typical moderately conservative tuning strategy aimed at avoiding overfitting.

The hybrid model is the central element of the empirical design and the study's main conceptual contribution. In this research, the hybrid approach is implemented as a two-stage scoring system of the logit + ML re-ranking type. At the first stage, logistic regression is estimated for the full sample to generate a baseline probability of default and perform the initial ranking of applications. At the second stage, a so-called grey zone of credit decisions is identified, that is, the subset of applications for which the PD predicted by the logit model falls within an interval of uncertainty. In the demonstration example, this grey zone is defined as a predicted default probability between 0.40 and 0.60. The second stage is activated for this subset of observations, which cannot be classified confidently as either low-risk or high-risk – ML re-ranking based on XGBoost.

Operationally, ML re-ranking means reordering the applications in the borderline segment using a more advanced algorithm capable of identifying nonlinear relationships between features and distinguishing risk more precisely in the most uncertain cases. In

other words, logistic regression acts as the baseline filter and provides the overall interpretable decision architecture, while XGBoost is applied only where additional analytical sensitivity is most valuable. In the demonstration setting, final credit decisions are defined by the following rules: applications with PD below 0.40 are automatically classified as low risk; applications with PD above 0.60 are classified as high risk; applications within the 0.40–0.60 interval are transferred to the second re-ranking stage, where XGBoost produces a refined risk estimate; after that, the final result may be adjusted by an expert rule in exceptional cases if additional non-model risk signals are available.

This architecture combines statistical transparency, algorithmic flexibility, and managerial control.

Model quality was assessed using the following metrics: AUC-ROC, Gini, the Kolmogorov–Smirnov statistic, Recall for the default class, Precision, F1-score, Brier score, and a comparison of results on the test and out-of-time samples. These indicators were chosen because banking scoring requires more than overall accuracy: default detection, calibration quality, and temporal stability also matter.

Based on the specified sample (Table 4), a demonstration comparison was carried out for five models: logistic regression, discriminant analysis, Random Forest, XGBoost, and a hybrid two-stage logit + ML re-ranking model. This set was selected deliberately. Logistic regression represents the baseline interpretable statistical approach, suitable for explaining the role of individual factors and for integration into regulation-sensitive credit analysis procedures. Discriminant analysis serves as a classical benchmark model, allowing modern instruments to be compared with traditional scoring logic. Random Forest represents ensemble trees and allows for complex interactions between variables. XGBoost was chosen as a modern boosting algorithm with strong predictive performance on tabular data. The hybrid model was constructed as a combined system in which the baseline logit performs the initial ranking of applications, while the ML component refines risk assessment for borderline observations that fall into the grey zone of the credit decision.

The results point to several conclusions (Tables 5 and 6). First, logistic regression confirms its value as a baseline interpretable model that provides an acceptable level of classification quality and is highly suitable for documentation, audit, and managerial explanation. Second, discriminant analysis, although still a classical scoring tool, shows lower flexibility and weaker predictive power than logistic regression. Third, Random Forest and XGBoost demonstrate higher quality metrics, confirming that ML approaches are better able to identify complex nonlinear relationships and hidden default patterns. At the same time, these models raise issues of interpretability, validity, and ease of use within internal banking and regulatory procedures.

The results of the demonstration testing suggest that scoring models should be compared not only by formal accuracy metrics, but also by their calibration, stability on new data, suitability for validation, and ability to be integrated into the operational framework of bank management. Logistic regression retains its role as a baseline interpretable model, discriminant analysis serves as a classical benchmark approach, and Random Forest and XGBoost confirm the advantages of modern ML algorithms in detecting complex relationships. The hybrid model is the most instructive case: although it does not produce the highest AUC, it offers the strongest balance between predictive quality, explainability, controllability, and practical applicability. This is fundamentally important for banking practice, since a credit decision has not only a statistical but also a managerial and regulatory dimension. In the hybrid model, the statistically transparent core of the credit decision is preserved, while the ML component is used where it yields the greatest benefit – in the zone of highest uncertainty.

For this reason, a deeper comparison of models should be based on the proposed evaluation matrix

Table 4
Fragment of the demonstration sample

ID	AGE	JOB_TENURE	INCOME	DTI	ACTIVE_LOANS	MAX_DPD	DPD_12M	LOAN_AMT	TERM	COLLATERAL	PRODUCT_TYPE	REGION_RISK	SHOCK_DUMMY	DEFAULT
1	34	48	32000	0.34	2	5	0	120000	24	0	consumer	medium	0	0
2	27	12	21000	0.58	3	35	2	85000	18	0	consumer	high	1	1
3	46	120	54000	0.22	1	0	0	250000	36	1	secured	low	0	0
4	39	36	28000	0.49	4	18	1	150000	30	0	Card	medium	1	1
5	31	60	41000	0.29	1	0	0	95000	12	0	consumer	low	0	0

Source: Compiled by the authors. The table contains an illustrative fragment of the synthesised sample

Table 5
Comparative results of the demonstration testing of scoring models

Model	AUC-ROC	Gini	KS	Recall (default = 1)	Precision	F1-score	Brier score	Calibration quality	Out-of-time stability	General characteristics
Logistic regression	0.781	0.562	0.417	0.684	0.641	0.662	0.168	High	Medium / high	The most interpretable baseline model, with good suitability for validation and regulatory use
Discriminant analysis	0.748	0.496	0.381	0.652	0.603	0.627	0.181	Medium	Medium	A classical scoring approach suitable for comparison, but less flexible and less accurate
Random Forest	0.836	0.672	0.491	0.752	0.706	0.728	0.149	Medium	Medium	A powerful ensemble model with high predictive power, but lower transparency
XGBoost	0.861	0.722	0.528	0.781	0.734	0.757	0.141	Medium / high after calibration	Medium	The highest accuracy and sensitivity to complex relationships, but more difficult to explain and control
Hybrid model (logit + ML re-ranking)	0.848	0.696	0.517	0.773	0.721	0.746	0.144	High	High	The best balance between accuracy, interpretability, calibration, and practical applicability

Source: Compiled by the authors. The values are illustrative and are used for methodological demonstration purposes

Table 6
Design of the demonstration testing of scoring models and key parameters of their construction

Testing element	Logistic regression	Discriminant analysis	Random Forest	XGBoost	Hybrid model
Model type	interpretable statistical	classical statistical	ensemble trees	boosting ML algorithm	two-stage combined
Total sample size	1000	1000	1000	1000	1000
Class structure	22% default / 78% non-default	22% / 78%	22% / 78%	22% / 78%	22% / 78%
Data split	60/20/20	60/20/20	60/20/20	60/20/20	60/20/20
Out-of-time test	yes	yes	Yes	yes	yes
Cross-validation	5-fold	5-fold	5-fold	5-fold	5-fold
Calibration	built-in	Platt / isotonic	Isotonic	Platt / isotonic	logit – built-in, ML stage – separate
Handling class imbalance	class weights	no additional balancing / basic weighting	class weights + SMOTE	class weights + SMOTE	class weights + local re-ranking
Main hyperparameters	L2 regularisation	standard LDA specification	n_estimators = 300, max_depth = 8, min_samples_leaf = 20	n_estimators = 250, learning_rate = 0.05, max_depth = 4, subsample = 0.8, colsample = 0.8	Stage 1: logit; Stage 2: XGBoost for the "grey zone"
"Grey zone" interval	–	–	–	–	PD = 0.40–0.60
Operational decision logic	direct ranking	direct ranking	direct ranking	direct ranking	low risk < 0.40; high risk > 0.60; borderline cases – repeated ML re-ranking
Key advantage	transparency	classical benchmark base	strong nonlinear classification	highest accuracy	balance of accuracy and interpretability
Main limitation	lower sensitivity to complex relationships	lower flexibility	lower explainability	governance complexity	more complex implementation architecture

Source: Compiled by the authors. The parameters are provided for demonstration testing and should not be identified with the models of any specific bank

(Table 7). Its practical value lies in the fact that it allows models to be assessed not only by accuracy, but also by their suitability for banking use.

5. Conclusions and Theoretical and Practical Implications

The partial testing of the proposed evaluation matrix shows that the suitability of a given model for use in a bank's credit risk management system cannot be determined solely by its formal predictive accuracy. A more methodologically grounded approach is to assess the model simultaneously across five interrelated blocks: predictive, interpretative, data and technical, regulatory-practical, and adaptive. The comparison shows that classical statistical models, especially logistic regression, have clear advantages in terms of transparency, validation, and managerial acceptability, but are weaker than modern ML approaches in capturing complex nonlinear relationships and adapting

Table 7

Partial testing of the proposed matrix for evaluating scoring models in the bank credit risk management system

Model	Predictive block	Interpretative block	Data and technical feasibility	Regulatory-practical block	Adaptability to Ukrainian conditions	General conclusion
Logistic regression	Medium: provides stable classification quality, an acceptable level of default borrower detection, and good predictive stability on temporally close samples	High: the model is transparent, allows interpretation of the sign and strength of factor effects, and is suitable for explaining individual decisions	High: does not require overly complex infrastructure, is relatively robust to moderate data limitations, and is easy to update	High: convenient for validation, internal control, audit, documentation, and use within the bank's credit policy	Medium: suitable in an unstable environment with regular recalibration, but less sensitive to sharp structural breaks	Appropriate as a baseline interpretable model and the statistical core of the scoring system
Discriminant analysis	Medium: provides a basic separation of defaulting and non-defaulting borrowers, but is less accurate than more modern models	High: the model logic is clear and the results are relatively easy to interpret	Medium: sensitive to assumptions about data distribution and less flexible when working with heterogeneous samples	Medium: can be used as a classical benchmark approach, but is less suitable for modern model governance	Low / Medium: only limited suitability for an environment with high volatility and frequent structural shifts	Of limited applicability, mainly useful as a classical comparative model
Random Forest	High: demonstrates good quality in detecting defaulting borrowers and greater sensitivity to nonlinear relationships	Medium / Low: the model is less transparent and requires additional explainability tools to interpret results	Medium: requires high-quality data preparation, but works well with a large number of features and is relatively robust to noise	Medium: suitable for practical use, but more difficult to validate and justify from a regulatory perspective than logit	Medium / High: responds better than classical models to complex changes in borrower behaviour, but requires constant monitoring of drift effects	Suitable as an enhancement model within the scoring system, especially for more complex data segments
XGBoost	High: provides the highest or one of the highest classification quality levels and strong sensitivity to complex and nonlinear interactions	Low: inherent interpretability is limited, requiring SHAP, feature importance, and other explanation tools	Medium: requires high-quality data, tuning procedures, careful stability checks, and more advanced technical support	Medium / Low: strong in predictive terms, but more difficult to justify normatively, audit, and use as a single baseline solution	High: the most sensitive to complex risk patterns and changes in customer behaviour, making it useful in an unstable environment	Appropriate for use in complex segments and as a model for advanced risk refinement
Hybrid model (logit + ML re-ranking)	High: combines the acceptable stability of baseline logit with the greater sensitivity of the ML component in borderline cases	Medium / High: preserves an interpretable statistical core, while the more complex algorithm is used only in the zone of increased uncertainty	Medium: requires a more complex modelling setup, but allows more rational use of data and resources	High: better aligned with bank governance because it combines the transparency of the baseline model with the analytical enhancement of the second stage	High: the most suitable for an unstable environment, as it combines controllability, flexibility, and regular adaptive validation	The most balanced option for bank credit risk management under volatile conditions

Source: Compiled by the authors on the basis of a partial demonstration testing of scoring models

to a more heterogeneous risk environment. Random Forest and XGBoost, by contrast, demonstrate higher predictive power, but at the cost of greater complexity in explanation, control, and integration into a regulated banking process. For this reason, the hybrid model has the highest practical value according to the proposed matrix, as it combines an interpretable statistical core with the additional analytical capacity of the ML component and thus provides the most balanced combination of accuracy, transparency, regulatory acceptability, and adaptability to unstable banking conditions.

The data generation process took into account basic correlations between variables, as well as two operating regimes of the environment – baseline and shock – which made it possible to bring the demonstration sample closer to the realistic structure of a credit portfolio. The resulting model quality metrics should therefore be interpreted not as the outcome of a single run, but as generalised estimates obtained within a repeatable procedure of training, cross-validation, and testing on both the test and out-of-time samples. The demonstration example should therefore be understood not as factual empirical evidence, but as a reproducible methodological illustration of how scoring models can be built, compared, and evaluated using multiple criteria.

This conclusion is particularly important for Ukrainian banking practice. The domestic banking environment operates under a combination of macroeconomic volatility, war-related risks, structural breaks, and uneven quality of historical data. Under such conditions, the challenge lies not only in selecting the "most accurate" model, but in developing a scoring system capable of combining statistical transparency, automated analytics, expert adjustment, and regular revalidation. The requirements of the National Bank of Ukraine for banks' risk management systems confirm the importance of proper model governance, while the functioning of the Credit Register and the broader credit infrastructure highlights the value of a reliable and stable information base for credit risk management.

The demonstration testing does not seek to reproduce the actual results of a specific bank portfolio, but it does provide a methodological illustration of how scoring models can be built, compared, and assessed on a multicriteria basis. Its findings indicate the potential usefulness of a hybrid approach under conditions of heightened uncertainty, since it combines the interpretability of a baseline statistical model with the additional predictive sensitivity of an ML component. At the same time, these conclusions should be regarded as methodologically grounded and as requiring further empirical verification on real bank portfolio data where access to confidential information is available.

Its scientific and practical value lies in moving the discussion from a narrow comparison of algorithms to the design of a multilevel system of risk-oriented credit decision-making.

6. Limitations and Suggestions for Future Studies

The study has several limitations that should be considered when interpreting the findings and setting directions for further research. First, the testing of scoring models was carried out on a synthesised demonstration dataset constructed according to the logic of a real retail bank loan portfolio. This approach is methodologically justified given the limited access to primary banking information and the need to comply with confidentiality requirements; however, it cannot fully capture the multidimensional nature of the real credit environment, including the institutional specifics of individual banks, the actual behaviour of borrowers, and the full range of structural and behavioural relationships within the credit portfolio. Accordingly, the findings should be read primarily as a methodological test of the proposed comparative framework for scoring model evaluation rather than as directly generalisable empirical conclusions for the banking sector as a whole.

A further limitation is the study's primary focus on the retail lending segment. While this choice is fully reasonable given the widespread use of scoring technologies in retail banking, the specifics of corporate lending, SME finance, and certain specialised credit products require separate and more detailed analysis. In addition, the paper concentrates on comparing traditional statistical models, machine learning models, and a hybrid approach, whereas other modern directions – including deep neural networks, graph models, more complex ensemble architectures, and advanced explainable AI tools – remain outside the scope of direct empirical testing. This does not diminish the significance of the results obtained, but it does define the boundaries of their applicability within the chosen research design.

It should also be taken into account that the Ukrainian banking environment is characterised by elevated macroeconomic volatility, structural breaks, uneven historical data quality, and changing borrower behaviour patterns. Under such conditions, even a model that performs acceptably over a given time horizon may require further adaptation, recalibration, or revalidation as the credit risk environment itself changes. Therefore, the results of this study should not be interpreted as grounds for selecting a single optimal model. Instead, they support a flexible, adaptive, and context-dependent approach to scoring system design.

Future research should verify the proposed framework on real bank data, subject to confidentiality and regulatory safeguards. It also appears advisable to test the proposed matrix for evaluating scoring models across different lending segments, types of banking institutions, and macroeconomic regimes. Further development of hybrid scoring systems, combining the interpretability of classical models with the greater predictive sensitivity of machine learning algorithms, is of clear scientific interest, as is the study of how explainable AI tools can be integrated into banking validation, audit, and model governanceprocedures.

Another promising direction is deeper research in the context of IFRS 9, early detection of credit quality deterioration, stress testing, and scenario-based credit risk assessment in an unstable environment. For Ukrainian banking practice, especially relevant issues remain the improvement of model resilience to structural breaks, data drift, changes in borrower behaviour, and fragmented historical information. Further scientific and practical work is needed here to develop a more flexible, reliable, and regulatorily compatible architecture of credit scoring within bank credit risk management.

References:

Alves, B. C., & Dias, J. G. (2015). Survival mixture models in behavioral scoring. *Expert Systems with Applications.* Vol. 42, Issue 8. P. 3902–3910. DOI: https://doi.org/10.1016/j.eswa.2014.12.036

Dornadula, V. N., & Geetha, S. (2019). Credit card fraud detection using machine learning algorithms. *Procedia Computer Science.* Vol. 165. P. 631–641. https://doi.org/10.1016/j.procs.2020.01.057

Indu Singh, Narendra Kuma, Srinivasa K.G., Shivam Maini, Umang Ahuja, & Siddhant Jain (2021). A multi-level classification and modified PSO clustering based ensemble model for credit scoring. *Applied Soft Computing*. Vol. 111. November 2021, 107687. https://doi.org/10.1016/j.asoc.2021.107687

John Banasik, Jonathan Crook (16 December 2007). Reject inference, augmentation, and sample selection. *European Journal of Operational Research*. Volume 183, Issue 3, Pages 1582–1594. https://doi.org/10.1016/j.ejor.2006.06.072

Jonathan N. Crook, David B. Edelman, & Lyn C. Thomas (16 December 2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*. Volume 183, Issue 3, Pages 1447–1465. https://doi.org/10.1016/j.ejor.2006.09.100

Law of Ukraine "On Banks and Banking Activity" (Vidomosti Verkhovnoi Rady Ukrainy (VVR), 2001, No. 5–6, Art. 30). https://zakon.rada.gov.ua/laws/show/2121-14#Text

Law of Ukraine "On the Organization of Formation and Circulation of Credit Histories" (Vidomosti Verkhovnoi Rady Ukrainy (VVR), 2005, No. 32, Art. 421). https://zakon.rada.gov.ua/laws/show/2704-15#Text

Leonardo Gambacorta, Yiping Huang, Han Qiu, Jingyi Wang (August 2024). How do machine learning and non-traditional data affect credit scoring? New evidence from a Chinese fintech firm. *Journal of Financial Stability* Volume 73, 101284. https://doi.org/10.1016/j.jfs.2024.101284

Pang, S., Hou, X., & Xia, L. (2021). Borrowers' credit quality scoring model and applications, with default discriminant analysis based on the extreme learning machine. *Technological Forecasting and Social Change.* Vol. 165. Art. 120462. https://doi.org/10.1016/j.techfore.2020.120462

Ricardo Muñoz-Cancino, Cristián Bravo, Sebastián A. Ríos, Manuel Graña (15 May 2023). On the dynamics of credit history and social interaction features, and their impact on creditworthiness assessment performance. *Expert Systems with Applications.* Volume 218, 119599. https://doi.org/10.1016/j.eswa.2023.119599

Ripa Akter Evolution and Emerging Frontiers of Explainable Artificial Intelligence (XAI) in Financial Risk Management: A Bibliometric Analysis. *Strategic Business Research.* Available online 18 March 2026, 100118. https://doi.org/10.1016/j.sbr.2026.100118

Thomas, L. C., Crook, J. N., & Edelman, D. B. (2002). *Credit Scoring and Its Applications.* Philadelphia : SIAM, 248 p. https://www.econbiz.de/Record/credit-scoring-and-its-applications-thomas-lyn/10009458236

Weixin Sun, Xuantao Zhang, Minghao Li, Yong Wang (November 2023). Interpretable high-stakes decision support system for credit default forecasting. *Technological Forecasting and Social Change.* Volume 196, 122825. https://doi.org/10.1016/j.techfore.2023.122825

Yadira Salazar, Paloma Merello, Ana Zorio-Grima (September 2023). IFRS 9, banking risk and COVID-19: Evidence from Europe. *Finance Research Letters.* Volume 56, 104130. https://doi.org/10.1016/j.frl.2023.104130

Yujia Chen, Raffaella Calabrese, & Belen Martin-Barragan (1 January 2024). Interpretable machine learning for imbalanced credit scoring datasets. *European Journal of Operational Research.* Volume 312, Issue 1, Pages 357–372. https://doi.org/10.1016/j.ejor.2023.06.036

Received on: 05th of April, 2026
Accepted on: 10th of June, 2026
Published on: 31th of July, 2026