Annals of Operations Research
https://doi.org/10.1007/s10479-024-06134-x

ORIGINAL RESEARCH

Prediction of bank credit worthiness through credit risk analysis: an explainable machine learning study

Victor Chang[1] · Qianwen Ariel Xu[1] · Shola Habib Akinloye[2] · Vladlena Benson[1] · Karl Hall[2]

Received: 24 March 2023 / Accepted: 20 June 2024

Abstract

The control of credit risk is an important topic in the development of supply chain finance. Financial service providers should distinguish between low- and high-quality customers to predict credit risk accurately. Proper management of credit risk exposure contributes to the long-term viability and profitability of banks, systemic stability, and efficient capital allocation in the economy. Moreover, it benefits the development of supply chain finance. Supply chain finance offers convenient loan transactions that benefit all participants, including the buyer, supplier, and bank. However, poor credit risk management in supply chain finance may cause losses for finance providers and hamper the development of supply chain finance. Machine learning algorithms have significantly improved the accuracy of credit risk prediction systems in supply chain finance. However, their lack of interpretability or transparency makes decision-makers skeptical. Therefore, this study aims to improve AI transparency by ranking the importance of features influencing the decisions made by the system. This study identifies two effective algorithms, Random Forest and Gradient Boosting models, for credit risk detection. The factors that influenced the decision of the models to make them transparent are explicitly illustrated. This study also contributes to the literature on explainable credit risk detection for supply chain finance and provides practical implications for financial institutions to inform decision making.

Keywords Credit risk analysis · Machine learning · Explainable artificial intelligence · Supply chain finance

✉ Victor Chang
v.chang1@aston.ac.uk; victorchang.research@gmail.com

[1] Department of Operations and Information Management, Aston Business School, Aston University, Birmingham, UK

[2] School of Computing, Engineering and Digital Technologies, Teesside University, Middlesbrough, UK

Published online: 08 July 2024

1 Introduction

1.1 Background

Credit risk is one of the most serious dangers that banks seek to mitigate in operations due to the nature of their business. In recent years, numerous studies have been done on these themes. Hosna et al. (2009) discovered a link between credit risk management and commercial bank profitability in Sweden. Credit risk management is positively associated with bank profitability in Nigeria, according to Kolapo, Ayeni, and Oke (2012). Credit risk management has little impact on commercial bank profitability in Kenya, according to Kithinji (2010). Ruziqa (2013) evaluated how traditional Indonesian banks performed in terms of credit and liquidity risk. Liquidity risk was shown to be positively related to profitability, while credit risk was found to be adversely associated with profitability. These studies show that no clear conclusion has been reached, suggesting this is a subject worth further investigating.

By correctly managing credit risk exposure, banks contribute to their own long-term viability, profitability, systemic stability and efficient capital allocation in the economy (Psillaki, Tsolas, and Margaritis, 2010, p.873). “A few late customers might cost the bank a lot of money” (Gestel & Baesems, 2009, p. 24). During the early stages of the Basel Accord, the Basel Committee recognized this as a substantial source of risk. Tabari et al. (2013) also claim that “bank profitability and capital sufficiency are necessary for the financial system’s stability”. In addition, the contribution of credit risk management to the health of banks can further benefit the development of supply chain finance. Supply chain finance (SCF) is a collection of technologically advanced business and financing processes that improve efficiency, reduce costs, or increase revenues for all participants. To be specific, it provides convenient loan transactions in which the buyer gets more time to pay, the supplier gets paid back faster, and the bank or other funding provider gets interested as the benefit (Choi, T. M., 2020; Liang et al., 2021). However, the nature of SCFs also poses unique credit risk challenges. Whereas in a traditional credit system, banks extend loans directly to individuals or businesses based on their creditworthiness, in SCF, this dynamic relationship is intricately intertwined with that between buyers and suppliers. This tripartite structure highlights the importance of identifying the credit risk associated with the ultimate borrower as well as the supply chain transaction. Poor credit risk management in supply chain finance may bring losses to finance providers and, in turn, make the development of supply chain finance suffer. Therefore, finance providers, including banks, must distinguish between low and high-quality customers to accurately predict credit risk.

Credit risk analysis is a collection of classifiers and the procedures that support them that help banks determine creditworthiness to reduce risk in operations research. Credit risk analysis is undoubtedly one of the most “conventional” applications of predictive modeling, since it predicts whether loans granted to an applicant will result in profit or loss for the bank. People, corporations, and other organizations are given credit for a variety of reasons (equipment purchases, Purchases of real estate and products for consumption, and so on) and through a variety of credit facilities (credit card, loan, delayed payment plan). On the other hand, a bank provides finance to a person or a business with the understanding that it will be paid back on schedule, with interest calculated according to the risk of default.

Several recent studies have been conducted on machine learning-based credit risk analysis. For example, Belhadi et al. (2021) proposed an ensemble method based on the Rotation Forest algorithm and Logit Boosting algorithm to predict the credit risk of Small and Medium-sized Enterprises in agriculture 4.0. Targeting SMEs in China, Wang et al. (2022) employed resampling methods, under-sampling, over-sampling, synthetic minority oversampling technique, and cost-sensitive learning to several classical machining learning algorithms for predicting the credit risk of SMEs. Mahbobi et al. (2021) applied DNN, SVM, KNN, and ANN to the prediction of default payments. More innovative models or applications will be discussed in the literature review. While these algorithms have made significant contributions to the improvement of the accuracy of the credit risk prediction systems that can be used in supply chain finance, they lack interpretability or transparency, thus making decision-makers skeptical of or even rejecting AI systems (Shin, 2021).

Therefore, this study aims to provide a credit risk detection system with great performance using machine learning algorithms. In particular, this study aims to improve AI transparency in data-driven decision-making for supply chain finance by ranking the importance of features that influence the decisions made by the system.

1.2 Research aims and objectives

By using a recent and large source of the dataset of bank loan defaults, this research aims to perform a credit risk analysis by comparing classification performances among different machine learning algorithms and enhance the AI transparency of these classifiers by ranking the importance of features that determines the detection results. In pursuit of this aim, our objectives are as follows:

- To examine various credit risk analysis approaches and their applications in predicting bank credit worthiness.
- Determine the most accurate method for forecasting whether a loan applicant will repay the loan in full or default.
- To calculate the likelihood of loan defaults based on the characteristics of loan applicants.
- To develop an AI transparency-enhanced model based on open-source data that predicts loan repayment or default.

1.3 Research question and research contributions

We focus on one major research question for the financial service firm that we have worked with. "How can machine learning models be leveraged to accurately predict credit risk while providing transparent explanations that align with the principles of responsible AI?" In this paper, we have demonstrated in-depth theories, steps, algorithmic development, simulations and interpretations to show our work can fully validate this research question.

This work significantly adds to the field of responsible and transparent AI in financial services in several ways. First, Gradient Boosting is shown to be the best-performing algorithm after a thorough assessment of the most recent machine learning algorithms for credit risk prediction. Second, it provides feature important explanations that go beyond model accuracy and transparently reveal the main factors influencing credit risk. Third, it offers a guide on how banks might create credit scoring models that align with ethical AI best practices, are highly predictive, and are understandable. When combined, these efforts improve our knowledge of how to create reliable and responsible AI systems for critical financial judgments.

1.4 Paper structure

The rest of this paper is structured as follows: Sect. 2 explores important credit worthiness related literature. More particularly, it focuses on machine learning (ML) methods for bank credit worthiness prediction and explainable AI for credit risk predictions. Section 3 covers the methodology conducted in the study, including details of the algorithms employed. Section 4 covers the model training process and evaluation. Section 5 provides conclusions and recommendations for future research directions in this area.

2 Literature review

This study reviews two main strands of related work. First, the implementation of ML models is reviewed and their applications in predicting credit worthiness. Second, studies focusing on the importance of transparency and explainability of such models and their associated techniques are reviewed. Understanding ML methods well and selecting suitable methods for research can be helpful to perform better in analysis and reduce risk in operations.

2.1 Machine learning models for Bank Credit Worthiness Prediction

A genuine, good assessment of bank credit worthiness has been proven to be an effective management tool. There have been many approaches for predicting credit worthiness. Several types of literature have attempted to classify them using statistical methodologies, artificial intelligence, data mining, and machine learning techniques. It is a good idea to take a glance at how things have evolved in the past before diving into the most cutting-edge concepts in this field. There has been a rise in interest in quantitative assessment of financial credit risk since Beaver and Altman's work in the 1960s. In the 1970s, methods such as ordinary least squares regression (Meyer & Pifer, 1970), discriminant analysis (Deakin, 1972), and logistic regression (Martin, 1977) were utilized to solve credit worthiness classification difficulties. Due to their ability to account for variable correlation, discriminant analysis variants surpassed univariate analysis in terms of performance. At this point, Altman had achieved a classification accuracy of more than 95% one time before credit worthiness and more than 70% three times prior to bankruptcy (Haldeman et al., 1977). By the 1980s, logit analysis (Ohlson, 1980), factor analysis (West, 1985), and other credit worthiness prediction approaches had been applied to the problem. Altman's initial Z-score technique was broadened in the 1990s to include private businesses, non-manufacturers, and emerging market companies (Altman, 1998). On one hand, this demonstrated that the same machine learning approach could be applied across a wide range of industries. However, it also emphasized how probability distributions differed amongst markets, challenging extrapolating training results outside the market. Using data from the US steel sector to develop an accurate prediction model for the European clothing market, for example, would virtually surely result in a significant category error.

By the late 2000s, Bayesian approaches were used to combine financial ratios and maturity schedule components (Philosophov et al., 2007). Other than financial ratios, predictors in the 2010s included accounting-based metrics, stock prices, company features, industry predictions, macroeconomic data, and agents' perspectives (Altman et al., 2011; Yu et al., 2011). The restricted expressiveness of financial ratios on the complex system of financial distress drove the model's growing number of parts. Observing "real market aspects" differs from estimating financial ratios; these hidden components must be identified indirectly. Additional elements typically result in a more accurate representation of the company's true health, but they come at a cost. Collecting the data can be time-consuming and costly (if not impossible due to confidentiality), making bank credit worthiness prediction a multi-dimensional classification task. This makes it more difficult to apply machine learning algorithms that must cope with the dimensionality curse efficiently. Most of the publications in this series examine datasets with up to 57 criteria (Du Jardin, 2009). In recent years, new solutions to these issues have arisen. There are a variety of artificial neural network modifications provided, including probabilistic neural networks (Sang, 2021) and self-organizing maps (Suleiman et al., 2021). Decision trees (Liu et al., 2022), support vector machines (Teles et al., 2021), soft computing (Lappas & Yannacopoulos, 2021), genetic algorithms (Lappas & Yannacopoulos, 2021; Yu & Cui, 2022), AdaBoost (Machado & Karray, 2022), and the Gaussian process are all examples of case-based reasoning successfully employed to analyze credit worthiness.

2.2 Explainable artificial intelligence (AI) in credit risk analysis

Machine learning (ML) methods frequently perform predictions better than standard statistical models (Cascarino et al., 2022). Researchers have used this property in many different disciplines, particularly when there are a lot of predictors available and the connection between the features and the outcomes is non-linear.

However, when these methods are applied in a commercial context, they bring several benefits while also revealing some important drawbacks. For example, in regulated financial services, every decision must be meaningful, and users need to comprehend the underlying reasons for the model's designation of the customer as a defaulter (Biecek et al., 2021). The excessive focus on the performance of predictive models or systems usually comes at the cost of a lack of interpretability or explainability, which leads decision-makers to be skeptical of or even reject AI systems (Shin, 2021).

Bussmann et al. (2021) also considered AI models as "black boxes" and believed that they are not well-suited for use in regulated financial systems and can present new difficulties for organizations working in the regulated financial services industry and the people in charge of them. For example, when evaluating the default risk of a potential borrower, individual results often lack interpretability, i.e., it is difficult to relate the predicted probability of default to the characteristics of the borrower, which poses a challenge for auditing and accountability of the estimated probability of default based on expert knowledge or the actual characteristics of the borrower. In addition, from a regulatory standpoint, the explainability of credit decisions is a consumer right that regulators should ensure its existence

(Cascarino et al., 2022). Explainable AI (XAI) models, which offer specifics or explanations to make AI operations understandable or clear, are required to solve this issue.

XAI allows interested stakeholders to understand the key drivers of model-driven decisions and to understand how AI algorithms decide, predict, and execute the process of their operations (Croxson et al., 2019; Rai, 2020). According to the European General Data Protection Regulation (GDPR), "the existence of automated decision-making should carry out meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject." Accordingly, under some conditions, the GDPR legislation gives data subjects the right to get relevant information about the reasoning behind automated decision-making (Regulation, 2016).

In the existing literature, several XAI methods have been applied to credit risk models for operations research, including measuring variable importance (Breiman, 2001; Fisher et al., 2018), Accumulated Local Effects (ALE) plot (Apley & Zhu, 2020), Shapley values (Shapley, 1953; Štrumbelj & Kononenko, 2014) and LIME (Visani et al., 2020).

To enhance the explainability of credit risk ratings provided by peer-to-peer lending platforms, Ariza-Garzón et al. (2020) and Bussmann et al. (2020) proposed an XAI model which builds on a similarity network model applied to the Shapley values of individual predictions.

Using a relatively novel XAI approach, Visani et al. (2020) interpreted a credit scoring model constructed using a gradient boosting tree and developed an approach to evaluate the stability of the resulting interpretation (Roa et al., 2021). In contrast to traditional bureau data, measured variable importance and SHAP values were to examine the impact of proxy data from an app-based marketplace on credit scoring models.

Kuiper et al. (2021) sought out the perspectives of banks and financial supervisory authorities to gain insight into how XAI approaches can be improved in the financial sector in the future. They investigated three major areas: credit risk, anti-money laundering and consumer credit by employing semi-structured interviews with businesses in the Netherlands. Major discrepancies were found in the desired scope of XAI systems between the banks and supervisory authorities. The differentiation between the technical aspects of AI modeling and explainability requirements should be clarified to overcome this issue.

Demajo et al. (2020) presented a credit scoring model for use in XAI that is highly interpretable while also being accurate. They achieved state-of-the-art performance on the Lending Club and HELOC datasets by implementing an XGBoost algorithm, augmenting it with a 360-degree explanation framework in order to provide local feature-based, local instance-based and global explanations. This was supplemented with manual analysis, proving that the explainability was consistent and understandable, satisfying hypotheses for trustworthiness, detail sufficiency, ease of understanding, effectiveness and correctness. One key direction they suggested for advancements in XAI was to combine global and local explanations to create decision trees, further increasing explainability.

Misheva et al. (2021) implemented two major model agnostic explainability algorithms – LIME and SHAP – to augment ML credit score algorithms employed for analyzing US-based peer-to-peer lending platforms. More specifically, LIME was used to enhance the explainability of local instances, whereas SHAP was used for both global and local instances. Furthermore, the issue of AI adoption in credit risk management was addressed, positing that the key to increase the use of AI in this domain is to increase trust by adopting XAI methodologies. Therefore, their work showed both LIME and SHAP to demonstrate consistent explanations correlating to financial theory to develop trust in AI technology in the finance sector.

Torrent et al. (2020) utilized a range of ML models for credit risk analysis, namely Logistic Regression, Random Forest, Gradient Boosting, CatBoost, and stochastic gradient neural networks. The major contribution of this work was by using SHAP to evaluate the performance of the ML predictions from both local and global perspectives. Notably, this proved that universal approaches should be avoided and that solutions should be tailored to the specific environment combination of credit scoring methodologies and financial establishments to understand the relationship between banks and their customers fully. Based on the findings of Torrent et al. (2020), it is clear that while SHAP values and other XAI methods provide insights into model predictions, they are not without limitations. It is critical to understand that the interpretations provided by these methods are strongly influenced by the quality and representativeness of the data to which they are applied. Moreover, while XAI can shed light on the mechanisms by which predictors influence outcomes, it is important to cross-check these interpretations with economics and common sense to ensure that conclusions are consistent with broader financial realities. Therefore, as researchers continue to utilise XAI in finance, it is paramount that these interpretations are treated with caution to ensure that they are both grounded in data-driven insights and economically sound.

Although there is a growing corpus of research on explainable AI techniques and machine learning for credit risk prediction, studies described above still thoroughly assesses various modeling approaches and produce feature importance explanations in this particular context. Furthermore, there still needs to be more consensus in the literature regarding the best ways to create accountable and transparent AI credit rating systems. By comparing cutting-edge machine learning models, figuring out what influences model predictions, and providing a framework that financial institutions can use to create AI systems that stakeholders can comprehend and rely on, this study seeks to close these gaps.

In summary, credit risk management is critical to the operation of banking, and related domains, such as supply chain finance. To better conduct credit risk analysis, a variety of advanced machine learning algorithms have been applied to this area by existing literature and achieved great performance in terms of accuracy. However, due to the black-box nature of AI, the output of ML-based detection systems cannot be easily understandable to decision-makers in the financial services industry. Moreover, the research on this domain is nascent. Therefore, our study aims to contribute to the literature by improving the explainability of the credit risk models by determining the threshold using the precision-recall AUC curve and F-measure and then measuring the feature importance. Providing explainable models can indicate which variables contribute more to the default prediction and provide meaningful information for the decision-makers in supply chain finance.

3 Methodology

This research aims to perform a bank credit worthiness analysis by comparing classification performances among different machine learning algorithms (logistic regression, support vector machine, decision tree, multi-layer perception, probabilistic, neural network and so on) by using a recent and large source of the dataset for bank loans default.

3.1 Machine learning algorithms

3.1.1 Logistic regression

Logistic regression models are efficient because they are fast and easy to train, and their results are simple to understand yet relatively accurate. As a result, they are often used in classification analyses. The most notable difference between a linear and a logistic function is that the latter is dichotomous in nature. After this is considered, the essential assumptions for both models are the same, with no limits on the homoscedasticity and normality of the variables used in the analysis being applied in either scenario.

3.1.2 Decision tree

Decision trees classify data by recursively dividing the data set into mutually exclusive subsets to better explain the variation in the dependent variable under observation (Tickle et al., 1998). In order to classify instances (data points), decision trees arrange along the tree from the root node to the leaf node at the end of the branches. This leaf node is responsible for categorizing the instance. The decision tree's branches depict several situations for decision-making, as well as the results of those scenarios.

3.1.3 K-Nearest neighbor

The *kth* and nearest neighbor distance can also be utilized as a density functional metric. Therefore, it is a popular outlier detection technique. The smaller the density, the more likely the outlier is and the closer the reference point is. The nearest neighbor classifier is the closest intuitive neighbor of the form classifier, and it assigns point x to the class of its closest neighbor in the function space.

The k-nearest neighbor classifier can be considered as a technique that assigns the k nearest neighbors a weight $\frac{1}{k}$ and assigns a weight of 0 to all other parameters. This can be generalized to weighted nearest-neighbor classifiers.

3.1.4 Random forest

The Random Forest algorithm chooses the most relevant splitting point and variable for each tree and every node to reduce the errors. The model uses every variable to give predictions, unlike CART, where the variables not selected do not interfere with the response. Therefore, the RF model provides a prediction based on all explanatory variables. This model can train with class imbalance due to how the voting process works. Consequently, it has the potential to perform well when dealing with credit data, which is often extremely unbalanced when compared to the input and response variables (Breiman, 2001).

3.1.5 Naive bayes

The Naive Bayes algorithm is a classification method based on Bayes' Theorem and assumes that predictors are unrelated. In other words, the Naive Bayes classifier assumes that the existence of one feature in a class has no bearing on the presence of any other feature in the same class.

3.1.6 Light gradient boosted machine

To address the limitations of the histogram-based algorithm, which is commonly employed in all gradient boosting decision tree frameworks, LGBM employs two methods, exclusive feature bundling and gradient-based one-side sampling. These techniques increase the model's efficiency and use less memory.

3.1.7 Adaptive boosting

Freund and Schapire (1997) developed an adaptive boosting (Adaboost) machine learning technique based on reinforcement learning. Adaboost gathers all weak classifiers in one location to build a strong classifier. Any machine learning system that performs slightly better than random guesses in terms of accuracy is referred to as a weak learner. Weak learners' outputs are linked with a weighted accumulation, which indicates the weighted classifier's results if they are erroneous. As a result of combining the weak classifiers, a superior classification will be created.

3.1.8 Gradient boosting

Gradient Boosting is an ensemble technique that aims to improve the accuracy of a predictive function by progressively lowering the error term in a predictive function (Friedman, 2001). Following the development of a base learner, the most common of which is a decision tree, each of the trees in a series is fitted to the residuals to decrease the error from prior trees. Gradient Boosting employs the bagging technique to improve its prediction ability. It also has a strong capacity to deal with class imbalances, making it ideal for bank credit worthiness modeling.

3.1.9 Linear discriminant analysis

Linear Discriminant Analysis makes predictions by estimating the probability that a new set of inputs corresponds to each of the four classes. The class with the best chance of being chosen as the output class is identified, with the prediction made.

The Bayes Theorem is used to calculate the model's probability. In its most basic version, Bayes' Theorem may be used to estimate the likelihood of an output class (k) given an input class (x) by considering the probability of each class and the probability of the data belonging to each class.

3.1.10 Multilayer perceptron

Multilayer perceptron (MLP) typically refers to any feedforward artificial neural network. The MLP in our bank credit worthiness modeling example has three layers: a single neuron in the output layer, a hidden layer with any number of hidden neurons, and an input layer with all the neurons for the explanatory variables. During training, each connection between neurons is given a weight, W_i, which determines how well each neuron understands its input/explanatory variables and then sends that value to the neurons in the next layer. An activation function $f^{(1)}$ is applied to the weighted inputs, together with the bias term $b_i^{(1)}$ to compute the output value of the hidden neuron in such a manner that:

$$h_i = f^{(1)}(b_i^{(1)} + \sum_{j=1}^{m} w_{ij}x_j).$$

3.2 Performance evaluation metrics

To run these models properly, the data must be categorized into training and testing sets. This allows the model to be run on the training data to figure out what parameters to use and then compare the results to how the model performs on the test data. It is vital to examine how accurate the model's predictions are after it has been built. A range of indicators is available for measuring the effectiveness of a model and quantifying the quality of the predictions. Some of the measures offered are based on true positive, true negative, false positive and false negative classifications, namely accuracy, precision, recall, and F1 score. Additionally, ROC-AUC curves are presented as an alternative way of evaluating the models' performance.

3.2.1 Discrimination threshold

Figure 1 shows an example of a discrimination threshold plot, which has five main components, including precision, recall, F1 score, queue rate, and discrimination threshold rate. The discrimination threshold is a critical value that is used to determine whether the predicted probability is high enough to label the positive class. Generally, this threshold is set at 50% probability. However, it may not be optimal for all scenarios, and adjusting it can help fine-tune the classifier to perform optimally. This is achieved by considering key metrics such as precision, recall, and F1 score, which can be optimized to achieve a balance between minimizing false positives and avoiding missed positive cases. Additionally, the queue rate is another crucial metric that describes the percentage of instances that require review. It should be optimized in accordance with business requirements, especially when the review cost is high. There are several methods to determine the threshold value. In this study, the model is executed several times over multiple train/test splits. The metrics are then represented as a median curve with a fill area around it. Then, the discrimination threshold can be set to maximize the performance of the model on the F1 score. In Fig. 1, the threshold is 0.28, corresponding to the queue rate of 0.21, which marks the labels with a score in the upper 21 percentiles as positive/risky. Figure 2 shows loan default distribution for our analysis.

One reason this paper selects this method is that it allows for the visualization of how different discrimination thresholds affect the trade-offs between sensitivity (true positive rate) and specificity (true negative rate). By adjusting the discrimination threshold, we can optimize the performance of the classifier for the specific requirements. In the context of bank loans default classification, for example, the decision-makers may want to optimize

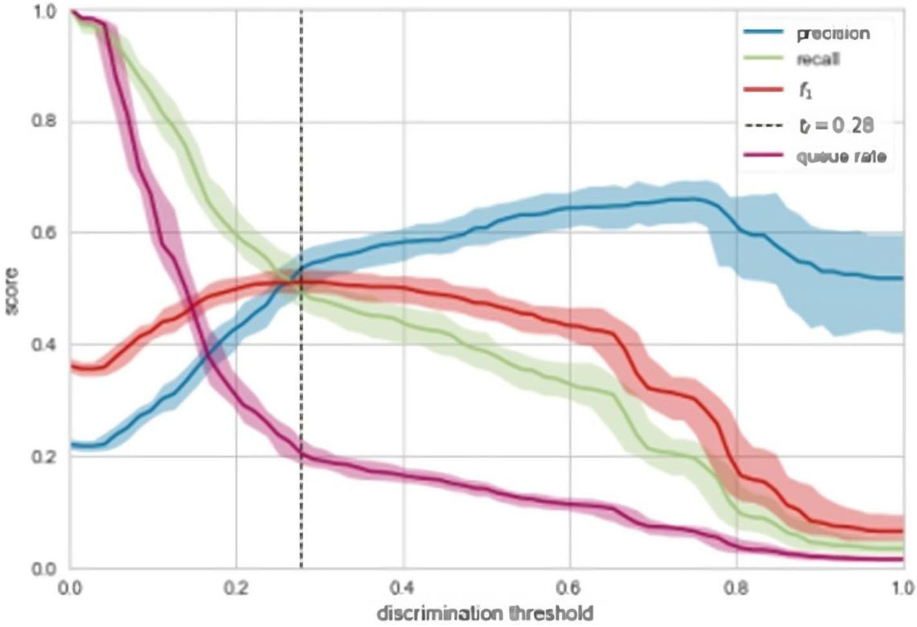

Fig. 1 Discrimination Threshold Plot

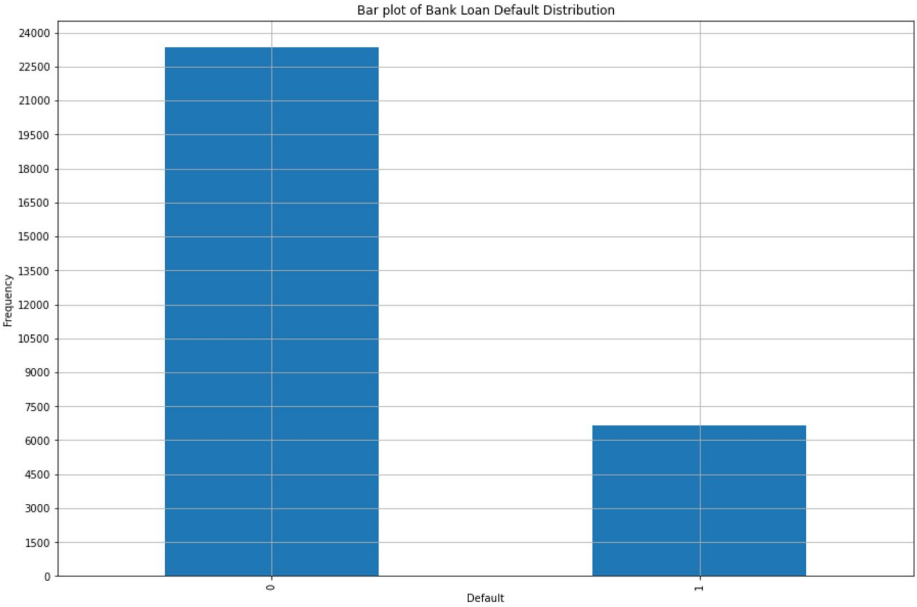

Fig. 2 Bar plot of loan default distribution

the model to minimize the number of false positives (good borrowers misclassified as risky) or false negatives (risky borrowers misclassified as good), or find the best balance between the two measures, as measured by the F1 score. What is also important is that this method accounts for the cost of investigation, as measured by the queue rate. This can help decision-makers of banks to balance the trade-offs between performance and cost, as they can see how changes to the discrimination threshold affect both the quality of the predictions and the number of potentially risky cases that require further review.

4 Model training and evaluation

4.1 Overview of the data

In order to develop a model for bank credit worthiness analysis through bank loan default classification fusing different machine learning algorithms, an investigation of available datasets was carried out and a dataset on customer payment default cases in the UK was finally selected[1]. It is the most suitable dataset for our study as it is the most recent dataset available and the data sample is relatively large to build a stable model. This dataset contains over 30,000 attributes of bank users who had taken loans across different banks in the United Kingdom. These customers were tracked over some periods to observe whether they had loan defaults or not. The various features considered in the data are the amount borrowed, which is a combination of both the individual customer credit as well as their supplementary credit, gender, education level, marital status and age in years. The dataset also includes prior payment history, which was compiled from the previous six-monthly payment data, amount of bill payments over the same period and the amount of previous payments.

Data had to be cleaned and organized in an analytically valuable style before it could be analyzed. All categorical variables were translated to numeric equivalents, with a unique number given to each category. There were no missing values found. Since not all the data sets were scaled evenly, the whole dataset was rescaled using a Z-score standardization; this was only required for the logistic regression.

SMOTE, KNN, and Tomek Links are all sampling approaches that may be used to deal with unbalanced datasets like the one utilized in this study. Owing to time and technology restrictions, a random over-sampling strategy was used for the target variable. The data were then randomly divided into a training and a test set using a 75/25% split. The test set was used to see how well the models predicted possible bank loan default, while the training set was used to make classification predictions.

4.2 Exploratory data analysis

In order to examine the distribution of the target variable, the bar chart presented below was computed and it can be observed from the chart that over 22,500 (75%) of the bank transactions did not have loan defaults, while well over 7000 out of the 30,000 transactions considered resulted in loan defaults.

[1] https://www.kaggle.com/datasets/praveengovi/credit-risk-classification-dataset.

To develop a better understanding of the features provided in the bank loan data used in this study, the summary statistics of all the features are computed and presented in Table 1:

Descriptions of the features can be found in Appendix Features description. From the total of 30,000 bank account attributes for this research study, the mean amount of the given credit (LIMIT_BAL) is £167,484.30, with a standard deviation (SD) of £129,747.70. The credit limit ranges from £10,000 to £1,000,000. The clients' ages (AGE) range from 21 to 79, with a mean age of around 35 years old, and an SD of 9.21. The table also shows the payment history of the account owners (PAY_0 to PAY_6), where the average monthly payment records from April to September 2005 are given as -0.291, -0.266, -0.221, -0.166, -0.133, and −0.016 for April, May, June, July, August, and September, respectively. The negative mean values imply that, on average, the clients paid their previous monthly payments on time. However, there were some clients who had payment delays of up to eight months or more (maximum values for PAY_0 to PAY_6 are 8).

Furthermore, the table displays the total number of bill statements from April (i.e., BILL_AMT6) to September (i.e., BILL_AMT1) of 2005. The average amount of bill statements for April is 38871.76, SD = 59554.11, May is 40311.4, SD = 60797.16, June is 43262.95, SD = 64332.86, July is 47013.15, SD = 69349.39, August is 49179.08, SD = 71173.77, and September is 51223.33, SD = 73635.86. The average amount of the bill statements tends to increase from month to month, from around 38871.76 in April to 51223.33 in September, which suggests that clients may be accumulating debt over time.

Finally, the average amount of the previous payment (pounds) in April is 5215.53, SD = 17777.47, May is 4799.39, SD = 15278.31, June is 4826.08, SD = 15666.16, July is 5225.68, SD = 17606.96, August is 5921.16, SD = 23040.87, and September 2005 is 5663.58, SD = 16563.87. The average payment amount tends to decrease from month to month, from

Table 1 Descriptive statistics of the features

Features	Count	Mean	SD	Min	Max
LIMIT_BAL	30,000	167484.3	129747.7	10,000	1,000,000
AGE	30,000	35.4855	9.217904	21	79
PAY_0	30,000	-0.0167	1.123802	-2	8
PAY_2	30,000	-0.1338	1.197186	-2	8
PAY_3	30,000	-0.1662	1.196868	-2	8
PAY_4	30,000	-0.2207	1.169139	-2	8
PAY_5	30,000	-0.2662	1.133187	-2	8
PAY_6	30,000	-0.2911	1.149988	-2	8
BILL_AMT1	30,000	51223.33	73635.86	-165,580	964,511
BILL_AMT2	30,000	49179.08	71173.77	-69,777	983,931
BILL_AMT3	30,000	47013.15	69349.39	-157,264	1,664,089
BILL_AMT4	30,000	43262.95	64332.86	-170,000	891,586
BILL_AMT5	30,000	40311.4	60797.16	-81,334	927,171
BILL_AMT6	30,000	38871.76	59554.11	-339,603	961,664
PAY_AMT1	30,000	5663.581	16563.28	0	873,552
PAY_AMT2	30,000	5921.164	23040.87	0	1,684,259
PAY_AMT3	30,000	5225.682	17606.96	0	896,040
PAY_AMT4	30,000	4826.077	15666.16	0	621,000
PAY_AMT5	30,000	4799.388	15278.31	0	426,529
PAY_AMT6	30,000	5215.503	17777.47	0	528,666

around 5,216 in April to 5,664 in September, which suggests that clients may be struggling to keep up with their debt payments.

Based on this analysis, it seems that many customers are carrying a significant amount of debt and may be struggling to make their debt payments on time. This could be a cause for concern, as it may lead to increased default rates and financial distress for both the clients and the banks. If banks do not have enough cash and a high number of customers want to withdraw cash, it may lead to possible collapse like what Silicon Valley Bank has experienced. Therefore, understanding the debt and cash flow situations are crucial to financial operations research and risk management. By using XAI approach, we can understand in-depth data analysis better.

4.2.1 Correlation matrix

In order to examine the inter-correlation among the features, a correlation matrix was computed and displayed in the form of a heatmap to indicate how the pairwise relationship among the features of the bank account users that obtained loans in the UK. The lighter the color of the color block where two features intersect, the stronger the positive correlation between these two features. See Fig. 3 for details.

The features can be divided into five types, including the amount of overdraft available on the account (LIMIT_BAL), AGE, the user's repayment status (PAY-related features), the monthly bill amount (BILL_AMT-related features), and the monthly repayment amount

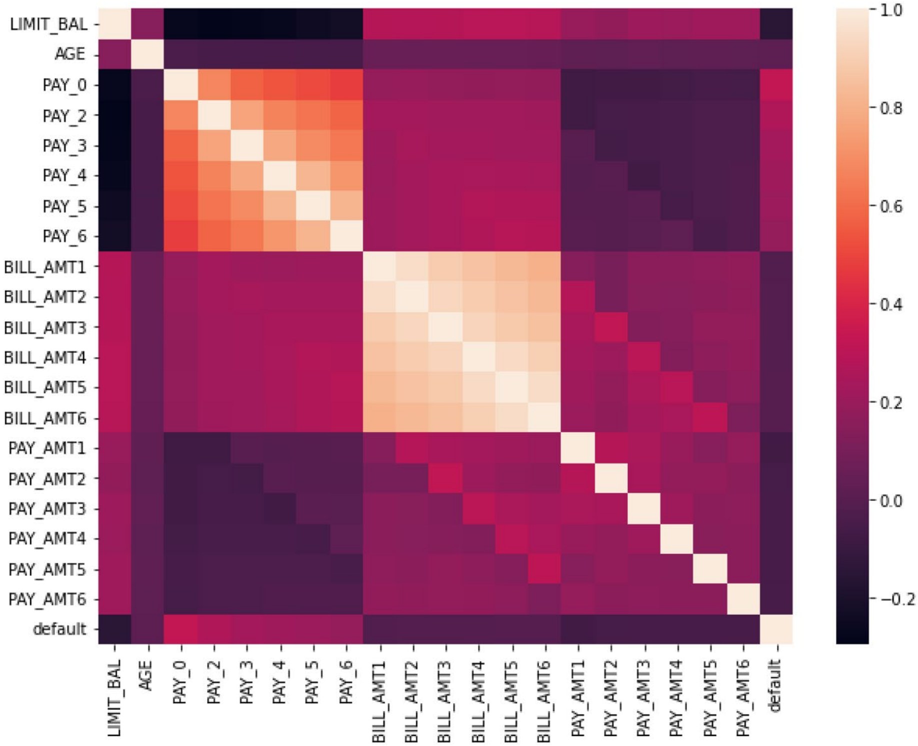

Fig. 3 Correlation matrix of data features

(PAY_AMT-related features). According to the correlation matrix, the monthly status (PAY-related features) are moderately correlated with each other, indicating that a customer who was late in paying their bills in one month is likely to be late in the following months as well. The monthly bill amounts (BILL_AMT-related features) are highly correlated with each other, indicating that a customer's bill amount tends to be consistent across months. Interestingly, the monthly repayments (PAY_AMT-related features) are relatively less correlated with each other, which suggests that customers do not always pay their bills in a consistent manner across months. Moreover, the correlations with the target variable of this study, default, are in the order of PAY-related features, BILL_AMT-related features, AGE, and PAY_AMT-related features and LIMIT_BAL. Clients with highly correlated features may be more likely to default. The contribution of features to default detection will be further explored in measuring feature importance.

4.3 Discussion of model results

The data were subject to the machine learning algorithms described in Sect. 3, and the model's performance was assessed using the assessment criteria provided. Finding a consistent metric for judging performance may be a complex and subjective procedure, depending on the work at hand. In reality, the best assessment model is a profit function that is a function of accuracy and recall, both of which must be improved in the future. A tradeoff between the TP (profit) and the FP (cost) must be made to calculate the profit, and both variables are recorded by the F1 score and precision-recall AUC. On the other hand, F1 score has been selected as the metric for evaluating the models in this research. The model with the best F1 score may be updated and tested further to improve prediction and, as a result, evaluation.

The highest-performing model for each class was chosen using a stratified 5-fold cross-validation technique. Table 2 compares model performance when the following metrics were used to assess their performance.

The Gradient Boost classifier has the best Accuracy score (82.49%), followed by Random Forest (82.39%) and Adaboost (82.11%). This indicates that Gradient Boost was the most effective algorithm at accurately classifying instances as default or non-default overall. Random Forest, AdaBoost, and Gradient Boost, had the highest Precision score of 0.81. In the context of credit risk detection, precision measures the accuracy of the model in predicting actual defaults among all cases that were predicted to default. The high precision scores of these three algorithms mean that they correctly identify the 81% of cases that are actu-

Table 2 Model evaluation

Algorithm	Precision	Recall	F1 Score	Accuracy	AUC
Logistic Regression	0.61	0.78	0.68	0.78	0.50
Decision Tree	0.80	0.81	**0.80**	0.81	0.66
KNN	0.70	0.75	0.71	0.75	0.54
Random Forest	**0.81**	**0.82**	**0.80**	0.82	0.66
Naïve Bayes	0.74	0.39	0.39	0.39	0.56
LGBM	0.77	0.79	0.78	0.79	0.63
ADABoost	**0.81**	**0.82**	0.79	0.82	0.64
Gradient Boost	**0.81**	**0.82**	**0.80**	**0.82**	**0.66**
LDA	0.80	0.81	0.78	0.81	0.61
MLP	0.72	0.74	0.73	0.74	0.59

ally defaults. Random Forest, Adaboost and Gradient Boost (82%) also returned the joint highest recall rate, followed by Decision Tree and LDA (81%). Recall measures the ability of the model to correctly identify all cases of default in the dataset. A Recall score of 82% means that these three algorithms correctly identified 82% of the actual defaults. Decision Tree, Random Forest and Gradient Boosting jointly returned the highest F1 score (80%), followed by Adaboost (79%), indicating that they were effective at balancing Precision and Recall. All of the models' Precision-Recall (PR) AUC is higher than 0.5, indicating that they performed well in capturing false negatives and positives. This also indicates that there are enough features for the models to train correctly. Gradient Boost (0.6605) achieved the highest AUC score, followed by Random Forest (0.6587), indicating that they were able to correctly classify most of the default and non-default cases in the dataset, regardless of the balance between the two categories. By measuring all indicators, Random Forest and Gradient Boosting were the two most effective algorithms for detecting credit card default risk.

When examining bank credit worthiness, there is the need to determine the proportion of accuracy acceptable to the financial institution being considered. False positives will lead to the exclusion of clients who would have been profitable otherwise but for whom the models mistakenly rated them as providing a high risk to the bank. Consequently, false negatives would raise the bank's risk by classifying customers as non-risky when they were more likely to fail, resulting in a financial loss for the bank. Since both situations are undesirable for the company, we should establish the best weighting for these two factors to ensure that the bank's expenses are reduced while earnings are increased. The most common method for selecting weights is to use hyperparameter optimization to find the most appropriate threshold between the two. The AUC curve depicts all possible threshold combinations that may be determined between recall and precision.

Given the computational and temporal constraints, hyperparameter optimization for all of the models was not practical; as a result, the study concentrated on improving the two best-performing models based on the F1 score and the PR AUC score, Random Forest and Gradient Boosting. In order to find the optimal parameters, a grid search was performed, and model tuning was performed on the two models as a consequence. The results are shown in Figs. 4 and 5. For the Gradient Boosting model, the optimal threshold is 0.27, meaning that if the predicted probability of the default class is greater than or equal to 0.27, the predicted class is the default, and otherwise, it is non-default. For the Random Forest model, the optimal threshold is 0.25. These two optimal thresholds are chosen to balance the false positive rate (good borrowers misclassified as risky) or false negative rate (risky borrowers misclassified as good), which enables Random Forest and Gradient Boosting models to achieve high AUC scores and overall good performance, respectively.

While various evaluation metrics measure the performance of machine learning algorithms, feature importance ranking plots are great visualization tools to show decision-makers in the financial services industry which factors highly indicate a customer's likelihood of high default risk. It also enhances the transparency of AI and makes the output of the black-box nature of algorithms more understandable. According to Figs. 4 and 5, one crucial observation is that the two models have a similar arrangement and degree of feature importance across the attributes considered in the models. More specifically, in both instances, the attributes contributing most to the likelihood of risk are PAY_0 and PAY_2, which outline the repayment status in September and August 2005, respectively. Regarding repayment status, the data shows the length of payment delay on these dates. Given that PAY_0 and

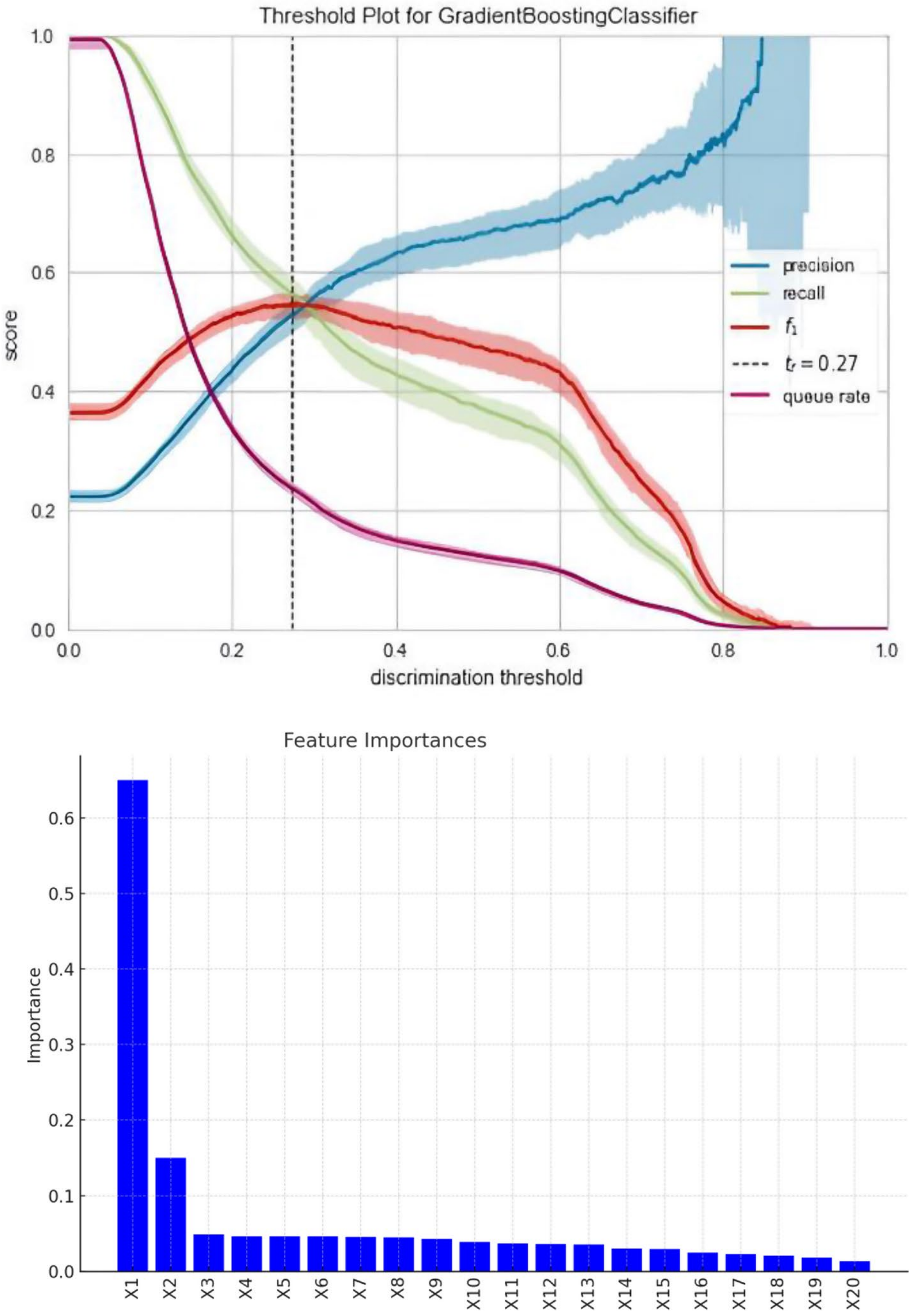

Fig. 4 Threshold plot and feature importance ranking for gradient boost classifier

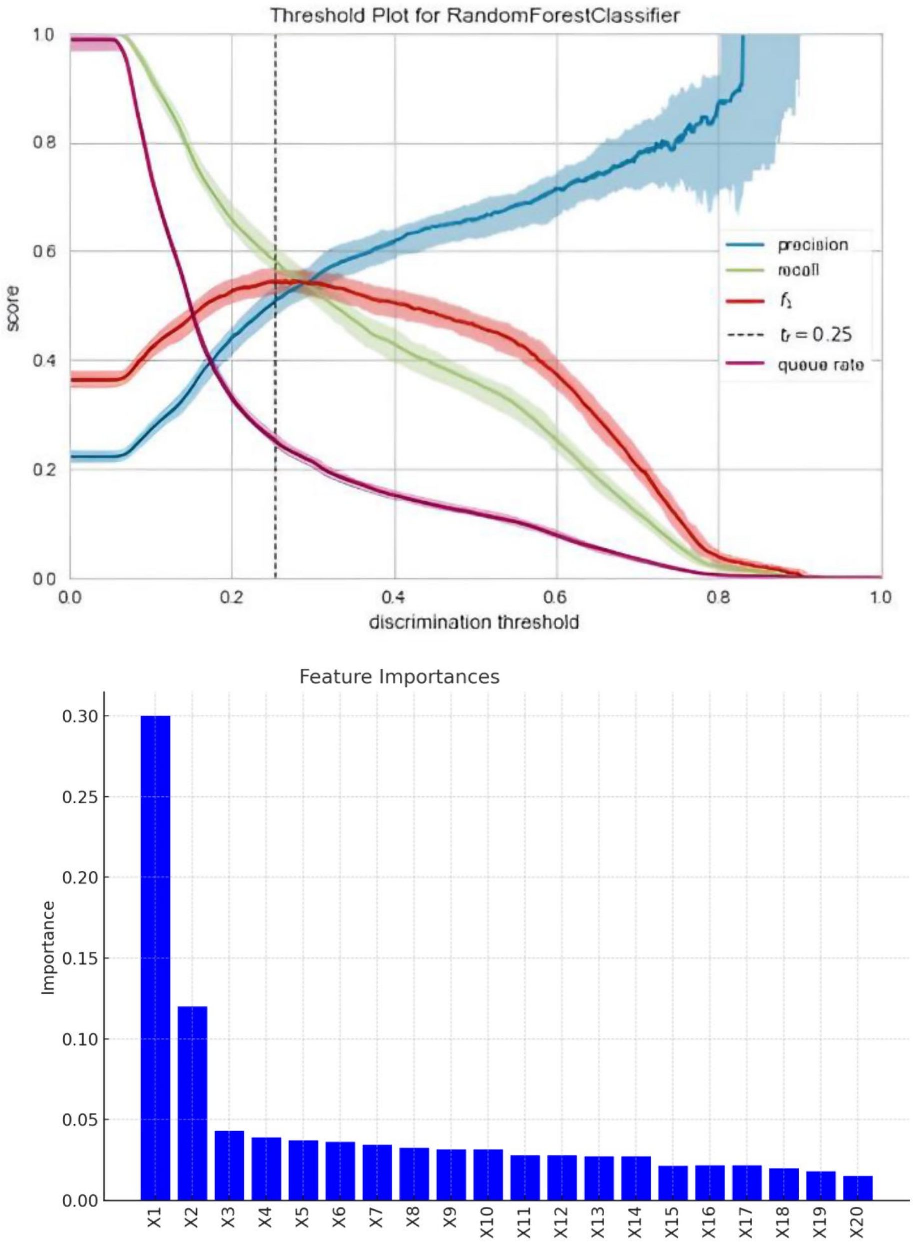

Fig. 5 Threshold plot and feature importance ranking for random forest classifier

PAY_2 refer to the most recent dates in the study, the most recent repayment status updates contribute the most towards the likelihood of risk. As this is a real case study for an anonymous financial service, therefore,

we have named them X1 to X20 for feature importance to meet anonymization and privacy requirements.

From these results, we can confirm that the Gradient Boosting classifier produced a better F1 score after model optimization as compared to the random forest model. This implies that the Gradient boost classifier best fits all models when predicting bank credit worthiness.

After determining the best classifier for this study, the importance of each figure determined by the Gradient boost classifier is then discussed to improve the explainability of this algorithm. Overall, the user's repayment status (features related to PAY) contributed the most to predicting default risk, followed by the amount of overdraft available on the account (LIMIT_BAL), the monthly repayment amount (features related to PAY_AMT), and then the monthly bill amount (features related to BILL_AMT), with gender (SEX) scoring near zero and representing almost irrelevant. Regarding the repayment status of each month, the importance of the features related to repayment is ranked as PAY_0, PAY_2, PAY_3, and PAY_4, PAY_5 and PAY_6, meaning that the closer the monthly repayment status is to the prediction time, the more meaningful it is for predicting the user's default risk, especially the repayment data of the last three months is considered the most important of all features by both algorithms, where the importance of the last month far exceeds the other features. In addition, the Gradient boost classifier considers the user's monthly repayment amount to be more meaningful than the bill amount for predicting the user's default risk in the next month, since most repayment amount features rank higher than the bill amount.

4.4 Discussion: how our findings support credit risk modeling

Our research findings add to and support the body of knowledge already available on transparent credit risk modeling. The Gradient Boosting model's excellent performance is consistent with earlier studies showing the ensemble learning approaches' predictive capacity in credit scoring scenarios (Abedin et al., 2023; Mushava & Murray, 2024).

Gradient Boosting's dominance over more conventional techniques like logistic regression also aligns with the larger body of machine learning literature, which has constantly shown the benefits of more adaptable, nonparametric algorithms (Belsti et al., 2023).

However, in addition to analyzing a larger variety of machine learning models, this study goes one step further by quantifying the relative significance of certain characteristics in influencing model predictions. The feature importance results are externally validated by identifying recent repayment history as the most significant predictor of default risk, consistent with domain expertise and economic intuition.

4.5 Limitations of the study

Unforeseen factors altered how the research was conducted, leading to the limitations of this study. First, the data used in the study lacked sufficient features to allow for enhanced model training, resulting in less accurate results. The number of hidden feature interactions would have increased due to the inclusion of more features, allowing the models to learn and make more accurate predictions and better measure metrics. Second, data preprocessing and preparation is frequently the most time-consuming portion of any machine learning project, and more might have been done to improve performance, especially in the case of feature engineering, but a lack of appropriate features slowed the process. Last, an estimate of future profits would have been a beneficial addition to this study, but it was outside the scope of this study. Since this is such an important aspect of any business, the study is constrained due to not having investigated it.

5 Conclusions and recommendations

5.1 Conclusions

Credit risk management is critical to the operation of banking, and failing to control the risk may bring significant negative effects to other related domains, such as supply chain finance and instability to financial operations. To better conduct credit risk analysis, in addition, to providing an effective credit risk detection model using machine learning algorithms, this study aims to improve AI transparency in data-driven decision-making for supply chain finance by ranking the importance of features that determines the detection results and reporting price and risk analysis.

The statistics show that determining a limit for assessing binary classification bank credit worthiness is challenging. According to the study, predicting bank credit worthiness is difficult, since deciding who is a high-risk customer is seldom a binary decision but rather a continuous system. Furthermore, the unavailability of a balanced dataset complicates calculations, which is a typical issue with imbalanced datasets, as shown by concentrates (Khandani et al., 2010).

The findings of this research demonstrate that machine learning models present a better technique for bank credit worthiness prediction when compared with the traditional statistical modeling approach. As observed, the Gradient Boosting classifier is the best fit among all models when predicting bank credit worthiness. The Logistic Regression model did badly in terms of performance, whereas the Random Forest classifier model came in second. The capacity of the model to boost organizational productivity and profitability by decreasing the open-door cost of false up-sides and raising revenue by lowering the cost of misleading negatives was classified as competent in this research study. The F1-score was used to represent the harmonic mean of precision and recall. Regardless, Logistic Regression scored well on the accuracy measure, which compares the proportion of adequately detected defaults to the absolute perception, with over 70% accuracy. Despite this, the accuracy measure is inadequate as a proxy for performance since it ignores the cost of misclassification, represented in the number of false positives and negatives, which is why the score was left out of the study.

Individual models often outperformed ensemble models, as seen by a range of measures, including the concentration of models by Butaru et al. (2016). Overfitting was a concern due to the imbalanced nature of the data, and the Gradient Boosting classifier's resistance to overfitting difficulties may have played a factor in why it outperformed the other models. The Random Forest model was successfully beaten because of the review's completion of hyper boundary optimization and model change to get a higher limit for precision and recall, with the F1-score increasing by 2.7%.

According to the study's findings, male customers were more likely than female customers to default, with a 14% and 9% likelihood of defaulting, respectively. Aside from that, this study revealed that those aged 18 to 33 were the most likely to default, while those aged 54 and older were the least likely. The researchers also observed that inconsistent payment history patterns play an essential role in determining whether a customer will default. When dealing with imbalanced datasets, the study revealed that the ensemble models perform better at demonstrating bank credit worthiness, similar to credit data sets; nevertheless, the models may have done better if the dataset included a larger number of highlights. More advanced testing procedures, such as SMOTE, may have contributed even more to the uneven data set and performance than traditional statistical modeling techniques.

In addition, this study argues that the joint use of complex models using ML and Explainable AI techniques can help understand the determinants of financial risk. Their combination helps to identify several characteristics (e.g., thresholds) of the relationship between credit risk and its determinants and gives greater weight to determinants that show non-monotonic or non-linear relationships with outcomes than logit models.

Aversion to AI systems appears to be especially pervasive in the finance industry, despite strong indications that such systems would improve the operations of the supply chain finance business. The key to increasing the adoption of such systems lies in gaining the trust of those who would benefit the most from implementing them. Ironically, these tend to be the people with the most skepticism – and understandably so. It then follows that it is not enough for modern AI systems to merely offer an improvement over traditional methods – they must be easily explainable. Those looking to employ AI systems, in supply chain finance or otherwise, want to be able to understand exactly how they work for them to be fully trusted. Indeed, given the complexity of these systems, this is not an easy task, and the onus is on the developers of AI systems to demonstrate the level of transparency and explainability that is becoming increasingly required.

5.2 Theoretical implications

This study has important theoretical implications for the development of transparent and responsible AI systems in the financial sector and beyond. The results demonstrate that it is indeed possible to build machine learning models for high-stakes decisions like credit underwriting that achieve state-of-the-art accuracy while also providing clear explanations of the factors driving model predictions.

The top performance of Gradient Boosting model, coupled with the generation of global and local feature importance values, provides a concrete example of how the often-competing objectives of predictive power and interpretability can be balanced. This research also outlines a generalizable framework - consisting of techniques like feature importance, SHAP values, and visualization of partial dependence plots - that can be applied to build explainable AI systems across a range of problem domains.

More broadly, this research makes a theoretical contribution by empirically validating the feasibility and value of combining sophisticated machine learning with explanation techniques to promote transparency and accountability. It extends the growing body of XAI research into the specific context of credit risk modeling, demonstrating how domain-specific considerations (e.g., the need to provide adverse action notices under fair lending laws) make explainability especially critical.

At the same time, the study highlights open theoretical questions around how to define and measure key constructs like “transparency”, “explainability”, and “responsibility” in the context of AI systems. While feature importance provides one lens of explanatory insight, further research is needed to explore other perspectives, such as those focused on the understandability of explanations to different stakeholder audiences. Ongoing work is also needed to develop theoretical frameworks that specify the necessary conditions for AI systems across different industries to be considered sufficiently transparent and responsible.

5.3 Practical implications for banks and financial services

The application of SHAP values to pinpoint the primary causes of default risk at the global and borrower levels demonstrates how banks can produce justifications that enhance the clarity and usefulness of model projections. Essentially, banks ought to combine advanced machine learning models with elucidation methods to identify the primary determinants of a borrower's risk score. This is particularly important since it enables banks to give the specific, model-driven justifications needed to comply with fair lending laws when borrowers are denied credit.

More strategically, banks can directly influence how they prioritize different data items in their credit risk models by taking into account the precise factors found in this study that were most predictive, such as a borrower's recent repayment history. Although the precise feature importance might fluctuate among distinct borrower populations and product categories, this study offers a broadly applicable proof-of-concept for employing model introspection methodologies to enhance data gathering and feature development endeavors.

Finally, this research can provide a guide for banks looking to combine the interpretability of explanatory techniques with the power of machine learning to create credit risk models that are transparent, highly accurate, and consistent with regulatory standards. Banks should see explainable AI as a chance to improve the ethical, equitable, and robustness of their models rather than as an additional regulatory burden. Banks can cultivate more trust from investors, customers, and regulatory bodies through the proactive adoption of the strategies defined in this research.

5.4 Recommendations

As previously stated, despite its limitations, this study provides several recommendations for financial operations research:

1. To enable feature engineering, which may improve performance by using a more optimal subset, the analysis should be performed on data with many more characteristics, such as monthly transactional information from client accounts.
2. AI provides transparency in deep data analysis and XAI will be utilized more frequently in complex data analysis.
3. It is recommended for institutions to analyze data with many more features, such as monthly transactional information on customers' accounts and monthly data. Additionally, considering the rising popularity of technology in the financial industry, alternative ways to perform predictive modeling will be useful for any institutions.
4. Future research may examine deep learning models and their performance. However, one disadvantage of these models is that they are opaque, making them difficult to explain using theoretical frameworks. The blended approach of using XAI and deep learning models can be effective.

Appendix A: Features description

Features	Description
LIMIT_BAL	Credit limit provided to the client
AGE	Age of the client (in years)
PAY_0 to PAY_6	Repayment status of the client's previous monthly payments. The measurement scale for the repayment status is as follows: -2 = no consumption; -1 = paid in full; 0 = use of revolving credit; 1 = payment delay for one month; 2 = payment delay for two months; …; 8 = payment delay for eight months or more.
BILL_AMT1 to BILL_AMT6	Amount of the client's bill statement (in NT dollar) for the respective months of April to September 2005.
PAY_AMT1 to PAY_AMT6	Amount of the client's previous payment (in NT dollar) for the respective months of April to September 2005.

Acknowledgements This work is partly supported by VC Research (VCR 0000182).

Declarations

Conflict of interest Authors confirm that there is no conflict of interest with anyone involved.

References

Abedin, M. Z., Guotai, C., Hajek, P., & Zhang, T. (2023). Combining weighted SMOTE with ensemble learning for the class-imbalanced prediction of small business credit risk. *Complex & Intelligent Systems*, *9*(4), 3559–3579.

Altman, E. I. (2011). Default Recovery Rates and Lgd in Credit Risk Modelling and Practice.

Altman, E. I., Hartzell, J., & Peck, M. (1998). Emerging market corporate bonds—a scoring system. In Emerging Market Capital Flows: Proceedings of a Conference held at the Stern School of Business, New York University on May 23–24, 1996 (pp. 391–400). Springer US.

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *82*(4), 1059–1086.

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

Belhadi, A., Kamble, S. S., Mani, V., Benkhati, I., & Touriki, F. E. (2021). *An ensemble machine learning approach for forecasting credit risk of agricultural SMEs' investments in agriculture 4.0 through supply chain finance* (pp. 1–29). Annals of Operations Research.

Belsti, Y., Moran, L., Du, L., Mousa, A., De Silva, K., Enticott, J., & Teede, H. (2023). Comparison of machine learning and conventional logistic regression-based prediction models for gestational diabetes in an ethnically diverse population; the Monash GDM Machine learning model. *International Journal of Medical Informatics*, *179*, 105228.

Biecek, P., Chlebus, M., Gajda, J., Gosiewska, A., Kozak, A., Ogonowski, D., & Wojewnik, P. (2021). Enabling machine learning algorithms for credit scoring–explainable Artificial Intelligence (XAI) methods for clear understanding complex predictive models. *arXiv Preprint arXiv*:210406735.

Bis.org (2014). *Basel Committee on Banking Supervision*. [online] Available at.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020). Explainable AI in fintech risk management. *Frontiers in Artificial Intelligence*, *3*, 26.

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics*, *57*(1), 203–216.

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

Cascarino, G., Moscatelli, M., & Parlapiano, F. (2022). Explainable Artificial Intelligence: interpreting default forecasting models based on Machine Learning. Bank of Italy Occasional Paper, (674).

Choi, T. M. (2020). Supply chain financing using blockchain: Impacts on supply chains selling fashionable products. *Annals of Operations Research*, 1–23.

Croxson, K., Bracke, P., & Jung, C. (2019). Explaining why the computer says 'no'. *FCA-Insight*, *5*, 31.

Deakin, E. (1972). A discriminant analysis of predictors of business failure. *Journal of Accounting Research*, *10*(1), 167–179.

Demajo, L. M., Vella, V., & Dingli, A. (2020). Explainable ai for interpretable credit scoring. arXiv preprint arXiv:2012.03749.

Du Jardin, P. (2009). Bankruptcy prediction models: How to choose the most relevant variables? *Bankers, Markets & Investors, 98*, 39–46. *Edition*. John Wiley & Sons.

Fisher, A., Rudin, C., & Dominici, F. (2018). Model class reliance: Variable importance measures for any machine learning model class, from the rashomon perspective. arXiv preprint arXiv:1801.01489, 68.

Freund, Y., & Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232.

Gestel, T. V., & Baesens, B. (2009). *Credit risk management*. [E-book] Available through: Oxford Scholarship Online http://www.oxfordscholarship.com.proxy.ub.umu.se/view/https://doi.org/10.1093/acprof:oso/9780199545117.001.0001/acprof-9780199545117-chapter-1.

Haldeman, R., et al. (1977). Zeta analysis: A new model to identify bankruptcy risk of corporations. *Journal of Banking and Finance*, *1*, 29–35.

Hosna, A., Manzura, B., & Juanjuan, S. (2009). Credit risk management and profitability in commercial banks in Sweden. rapport nr.: Master Degree Project 2009: 36.

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, *34*(11), 2767–2787.

Kithinji, A. M. (2010). *Credit Risk Management and profitability of commercial banks in Kenya*. School of Business, University of Nairobi.

Kolapo, T. F., et al. (2012). Credit risk and commercial banks' performance in Nigeria: A panel model approach. *Australian Journal of Business and Management Research*, *2*(2), 31–38.

Kuiper, O., Berg, M. V. D., Burgt, J. V. D., & Leijnen, S. (2021, November). Exploring explainable AI in the financial sector: perspectives of banks and supervisory authorities. In Benelux Conference on Artificial Intelligence (pp. 105–119). Springer, Cham.

Lappas, P. Z., & Yannacopoulos, A. N. (2021). A machine learning approach combining expert knowledge with genetic algorithms in feature selection for credit risk assessment. *Applied Soft Computing*, *107*, 107391.

Liang, D., Cao, W., & Wang, M. (2021). Credit rating of sustainable agricultural supply chain finance by integrating heterogeneous evaluation information and misclassification risk. *Annals of Operations Research*, 1–31.

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

Machado, M. R., & Karray, S. (2022). Assessing credit risk of commercial customers using hybrid machine learning algorithms. *Expert Systems with Applications*, *200*, 116889.

Mahbobi, M., Kimiagari, S., & Vasudevan, M. (2021). Credit risk classification: An integrated predictive accuracy algorithm using artificial and deep neural networks. *Annals of Operations Research*, 1–29.

Martin, D. (1977). Early warning of banking failure. *Journal of Banking and Finance*, *7*, 249–276.

Meyer, P., & Pifer, H. (1970). Prediction of bank failures. *The Journal of Finance*, *25*(4), 853–868.

Misheva, B. H., Osterrieder, J., Hirsa, A., Kulkarni, O., & Lin, S. F. (2021). Explainable AI in credit risk management. *arXiv Preprint arXiv*:210300949.

Mushava, J., & Murray, M. (2024). Flexible loss functions for binary classification in gradient-boosted decision trees: An application to credit scoring. *Expert Systems with Applications*, *238*, 121876.

Ohlson, J. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, *18*(1), 109–131.

Philosophov, L. (2007). Predicting the event and time horizon of bankruptcy using financial ratios and the maturity schedule of long-term debt. *EFA 2005 Moscow Meetings Paper*.

Psillaki, M., et al. (2010). Evaluation of credit risk based on firm performance. *European Journal of Operational Research*, *201*(3), 873–888.

Rai, A., & Explainable, A. I. (2020). From black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137–141.

Regulation, E. U. (2016). 679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union L*, 119, 1–88.

Roa, L., Correa-Bahnsen, A., Suarez, G., Cortés-Tejada, F., Luque, M. A., & Bravo, C. (2021). Super-app behavioral patterns in credit risk models: Financial, statistical and regulatory implications. *Expert Systems with Applications*, *169*, 114486.

Ruziqa, A. (2013). The impact of credit and liquidity risk on bank financial performance: The case of Indonesian Conventional Bank with total asset above 10 trillion Rupiah. *International Journal of Economic Policy in Emerging Economies*, *6*(2), 93–106.

Sang, B. (2021). Application of genetic algorithm and BP neural network in supply chain finance under information sharing. *Journal of Computational and Applied Mathematics*, *384*, 113170.

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, *2*(28), 307–317.

Shin, D. (2021). Why does explainability matter in news analytic systems? Proposing explainable analytic journalism. *Journalism Studies*, *22*(8), 1047–1065.

Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, *41*(3), 647–665.

Suleiman, S., Ibrahim, A., Usman, D., Yabo, B. I., & Muhammad, H. U. (2021). Improving credit scoring classification performance using self Organizing Map-based machine learning techniques. *European Journal of Advances in Engineering and Technology*, *8*(10), 28–35.

Tabari, N. A. Y., Ahmadi, M., & Emami, M. (2013). The effect of liquidity risk on the performance of commercial banks. *International Research Journal of Applied and Basic Sciences*, *4*(6), 1624–1631.

Teles, G., Rodrigues, J. J., Rabêlo, R. A., & Kozlov, S. A. (2021). Comparative study of support vector machines and random forests machine learning algorithms on credit operation. *Software: Practice and Experience*, *51*(12), 2492–2500.

Tickle, A. B., Andrews, R., Golea, M., & Diederich, J. (1998). The truth will come to light: Directions and challenges in extracting the knowledge embedded within trained artificial neural networks. *IEEE Transactions on Neural Networks*, *9*(6), 1057–1068.

Torrent, N. L., Visani, G., & Bagli, E. (2020). PSD2 explainable AI model for credit scoring. arXiv preprint arXiv:2011.10367.

Visani, G., Bagli, E., Chesani, F., Poluzzi, A., & Capuzzo, D. (2020). Statistical stability indices for LIME: Obtaining reliable explanations for machine learning models. *Journal of the Operational Research Society*, 1–11.

Wang, L., Jia, F., Chen, L., & Xu, Q. (2022). Forecasting SMEs' credit risk in supply chain finance with a sampling strategy based on machine learning techniques. *Annals of Operations Research*, 1–33.

West, R. (1985). A factor analytic approach to bank condition. *Journal of Banking and Finance*, *9*, 253–266.

Yu, J., & Cui, H. (2022). Rural Financial Decision Support System Based on Database and Genetic Algorithm. Wireless Communications and Mobile Computing, 2022.

Yu, L., Yao, X., Wang, S., & Lai, K. K. (2011). Credit risk evaluation using a weighted least squares SVM classifier with design of experiment for parameter selection. *Expert Systems with Applications*, *38*(12), 15392–15399.