Journal of Economics and Business 125-126 (2023) 106126

Contents lists available at ScienceDirect

Journal of Economics and Business

journal homepage: www.elsevier.com/locate/jeb

Explainable FinTech lending

Golnoosh Babaei, Paolo Giudici *, Emanuela Raffinetti

Phd in Computer Engineering and Department of Economics and Management, University of Pavia, Via San Felice 5, 27100 Pavia, Italy

ARTICLE INFO

Keywords:
Fintech
Credit scoring
Artificial intelligence
Machine learning
Shapley values

ABSTRACT

Lending activities, especially for small and medium enterprises (SMEs), are increasingly based on financial technologies, facilitated by the availability of advanced machine learning (ML) methods that can accurately predict the financial performance of a company from the available data sources. However, despite their high predictive accuracy, ML models may not give users sufficient interpretation of the results. Therefore, it may not be adequate for informed decision-making, as stated, for example, in the recently proposed artificial intelligence (AI) regulations. To fill the gap, we employed Shapley values in the context of model selection. Thus, we propose a model selection method based on predictive accuracy that can be employed for all types of ML models, those with a probabilistic background, as in the current state-of-the-art. We applied our proposal to a credit-scoring database with more than 100,000 SMEs. The empirical findings indicate that the risk of investing in a specific SME can be predicted and interpreted well using a machine-learning model which is both predictively accurate and explainable.

1. Introduction

The advent of financial technologies (fintech) has led to the emergence of several new companies in financial markets (Fasano and Cappa, 2022). Thanks to technologies such as Artificial Intelligence (AI), Blockchain, and Cloud computing, these companies offer additional services to compete with traditional banks (Kendall, 2017; Temelkov, 2018). In fact, in fintech platforms, such as crowdfunding, peer-to-peer lending, and robot advisory, based on digitalised business models, finance and technology meet (Ayadi et al., 2021), improving customer experience, lowering costs and increasing transparency (Romānova and Kudinska, 2016).

Fintech platforms have substantially changed several financial services. Among them, online lending platforms such as the Lending Club, one of the largest peer-to-peer lending organisations, have increased financial inclusion, allowing credit allocation to borrowers typically not funded by traditional banks while providing highly attractive returns for investors.

However, lending platforms bear high risks for investors, as borrowers are typically small and medium enterprises (SMEs) or low-income individuals (Correia and Martins, 2022), which, in addition, are largely interconnected. It follows that future perspectives of lending platforms largely depend on assessing credit risk and determining the causal drivers of such risks (Milne and Parboteeah, 2016).

This study contributes to fintech credit risk assessment for lending to SMEs. Providing credit to SMEs is a key research topic for policymakers (Berger and Udell, 2006; Ferri and Murro, 2015). Fintech lending facilitates credit services and financial inclusion, directly connecting individual lenders with company borrowers through a credit assessment platform that analyses all available data on borrowers to learn and continuously update their scores and classes of scores (ratings). Compared to traditional bank lending,

* Corresponding author.
E-mail address: giudici@unipv.it (P. Giudici).

https://doi.org/10.1016/j.jeconbus.2023.106126
Received 15 July 2022; Received in revised form 24 May 2023; Accepted 26 May 2023
Available online 2 June 2023

fintech lending improves customer experience and provides more credit to companies. However, they can suffer from information asymmetry between borrowers and lenders (Giudici and Hadji-Misheva, 2020; Bracke et al., 2019; Kumari and Kaur, 2021), possibly leading to inaccurate creditworthiness estimates.

To solve this problem, ML models, a combination of statistical models and computational algorithms able to learn from large databases regularities and relationships between a large number of variables, can be applied to lending data to obtain estimates of creditworthiness more accurate than those obtained with classical credit scoring models.

ML models have been widely used in many financial studies, such as credit scoring (Bastani and Asgari, 2019; Lee and Chen, 2005; Shen and Zhao, 2020), portfolio optimization (Guo et al., 2016), and profit scoring (Serrano-Cinca and Gutiérrez-Nieto, 2016; Babaei and Bamdad, 2020). These studies demonstrate that ML models perform well in terms of predictive accuracy. However, their predictions are not easily interpretable because the underlying model is a nontransparent black box.

To solve this problem, explainable Artificial Intelligence (XAI) methods, in which humans can understand the results of the solution, have recently been proposed (Lundberg and Lee, 2017; Ribeiro and Singh, 2016b; Bussmann et al., 2021; Sachan et al., 2020; Giudici and Raffinetti, 2021).

The XAI models achieve a good trade-off between explainability and predictive accuracy. However, massive computation may be involved when the number of explanatory variables is large. To reduce the computational burden, we propose to apply XAI models and, specifically, Shapley values to interpret ex-post the predictive power of each variable and as an ex-ante variable selection criterion. This leads to more parsimonious models, which, while maintaining a good predictive accuracy, can be better interpreted by users while maintaining good predictive accuracy.

Our proposed model can thus support banks and fintechs in developing an AI-based credit scoring model which is ‘trustworthy’, accurate, and explainable, with the potential of being validated by supervisory authorities and regulators.

From a statistical viewpoint, the proposed variable selection method combines predictive accuracy with explainability. Variable selection methods are well known in the literature and relate to Occam’s Razor principle: ‘Among competing hypotheses, the ones with the fewest assumptions should be selected.’ When many explanatory variables are available, applying this principle leads to the diffusion of stepwise variable selection algorithms that compare models comprising different sets of variables in terms of their statistical likelihood.

However, most ML models are non-probabilistic, and likelihoods are unavailable, preventing the use of stepwise selection algorithms. Alternative ML models, such as neural networks with different layers and hidden nodes or random forests with different input variables, can be compared in terms of predictive accuracy. This suggests that a different stepwise procedure is currently unavailable.

We propose to fill the gap by employing the Shapley value associated with each explanatory variable as the underlying metric to perform stepwise variable selection. Specifically, the variables that contributed the least to the predictions regarding their Shapley values were removed from the model. Variables were removed if the predictive accuracy of the model was not significantly reduced. Thus, both explainability and predictive accuracy were achieved.

We applied our methodological proposal to compare alternatives; we used random forest models that aimed to build credit scores that accurately predicted the probability of default (PD) for a set of companies. This helped to leverage the value of the data and the nonlinear relationships present in the data, leading to a more accurate and transparent credit scoring model that can be employed in fintech lending platforms.

From a managerial viewpoint, our proposal allows us to identify the variables that explain the credit risk of lending investments in SMEs.

To the best of our knowledge, this is the first methodological study to employ the XAI as a variable selection tool within the credit scoring context.

The rest of the paper is organised as follows. Section 2 contains a literature review on applying ML and AI to credit scoring. Section 3 introduces the proposed method. Section 4 presents the data and the main empirical findings. Finally, Section 5 concludes the study.

2. Literature review

Credit scoring is a research topic which has attracted many researchers, who have employed different statistical learning models to measure it (Bücker et al., 2022; Liu and Fan, 2022; Dushimimana et al., 2020). The modern ML methods have found one of their first fields of application to economics in credit scoring: among the first studies, we can mention (Srinivasan and Kim, 1987) in which decision trees are used; (Henley and Hand, 1996) in which k-nearest neighbours are employed; (West, 2000; Yobas and Crook, 2000), in which neural networks and support vector machines are applied; (Djeundje and Hamid, 2021), in which a range of ML models are applied to both traditional and alternative credit scoring data. See (Hand and Mannila, 2001) for a review of the data mining methods for credit scoring.

In the last few years, the emergence of ensemble methods, which aggregate results from different models, has substantially improved the performance of scoring models based on ML (Finlay, 2011; Lessmann et al., 2015). In this respect, Li and Chen (2020) provides a comparison of different ensemble methods: random forests, adaptive boosting, gradient boosting, and light gradient boosting, applied to five alternative credit scoring models: neural networks, classification trees, logistic regression, naïve Bayes, and support vector machines. Our study shows that the performance of ensemble credit scores was better than that of individual scores. We also show that the ensemble random forest model achieved the best accuracy metrics, such as the Area Under the ROC Curve, the Kolmogorov-Smirnov statistic, and the Brier score. In another study, Chopra and Bhilare (2018) compared random forests and gradient boosting using a credit-scoring model based on a classification tree. They showed that for their credit scoring application, ensemble methods (gradient boosting and random forest) outperformed individual classification tree models, thereby adding further evidence to the higher accuracy of ensemble methods. A third study, Tripathi et al. (2022), undertook a comparative analysis of nine ensemble methods applied to different scoring models, such as logistic regression, naïve Bayes, and classification trees. As in previous studies, they found that ensemble scoring methods improve the performance of single-credit scoring methods.

The previous discussion indicates a consensus on the superior predictive accuracy of ensemble credit scoring models concerning single models. However, the increased accuracy comes with a cost: while most single scoring models, such as logistic regression, tree models and naïve Bayes are 'explainable', as they can identify the contribution of each explanatory variable to the credit scores, ensemble methods are 'black boxes', and cannot explain the determinants of credit scores to their users (Bracke et al., 2019; Giudici and Hadji-Misheva, 2020). This is a problem from a regulatory viewpoint because the application of ML AI to credit scoring, a high-risk application, must be accurate and explainable, as stated in the recently proposed European Artificial Intelligence Act (https://artificialintelligenceact.eu).

To overcome this problem, ensemble methods should be complemented with explainable AI methods that are to be applied a posteriori on the obtained credit scores. Explainable AI methods can be classified into model-specific and model-agnostic (Adadi and Berrada, 2018). In contrast to model-specific methods, model-agnostic methods can be applied to any ML model. Local methods such as Local Interpretable Model agnostic explanations (LIME) (Ribeiro and Singh, 2016a) and Shapley values (Lundberg and Lee, 2017) are of particular interest, both explaining each specific credit score based on the additional contribution of each explanatory variable to their values.

Local methods have been recently applied to explain credit scores based on ML. For instance, Bussmann et al. (2021) propose a methodology based on Shapley values as a post-processing analysis to explain the credit scores obtained from ensemble models applied to data that concern a sample of Italian SMEs, which apply for peer-to-peer lending. Their empirical results demonstrated the capability of explainable AI methods to achieve predictive accuracy and explainability. A related study, Moscato and Picariello (2021), proposed a credit scoring model to predict whether a loan will be repaid on a P2P platform. It compared different ML models and explainability methods, including LIME and SHAP, showing their advantages. Similarly, Xia et al. (2021) showed how credit scores obtained with gradient boosting can be interpreted using Shapley values, and Tyagi (2022) compared various ML models for credit scoring in terms of Shapley values to develop new investment models and portfolio strategies. All these studies provide evidence of the advantage of using explainable AI methods in combination with ML models in credit scoring. Our study falls into this research stream. It proposes a credit-scoring model based on an ensemble machine-learning method that can be explained using the Shapley value approach. Our original contribution is that we propose achieving explainability, not a posteriori, by applying the Shapley value to the obtained credit scores but ex-ante as a variable selection criterion.

Our proposal is inspired by the acknowledged advantage of variable selection in improving the predictive accuracy of ML models. For example, Laborda and Ryoo (2021) discussed the performance of three variable selection models: a filter method (based on statistical tests) and two wrapper methods (based on stepwise model selection) to obtain a more parsimonious credit scoring model based on logistic regression, support vector machine, K-nearest neighbours, or random forest. They concluded that stepwise selection yielded a superior predictive performance for all models. A related study, Trivedi (2020) employed chi-square testing as a filter method for ML classifiers, such as naïve Bayes, random forest, classification trees, and support vector machines, to improve credit scoring predictions. They found that chi-square testing with credit scoring improved the predictive accuracy of all classifiers.

In this study, we combined variable selection with explainability, proposing a variable selection model that chooses the most explainable variables as model predictors. Thus, variable selection improves the predictive accuracy and interpretability of the credit scores obtained with an ML model. It does so before reaching a final model (ex-ante perspective) rather than after a model has been selected (as in the available applications of XAI models), thereby reducing the computational burden.

To achieve explainability, we considered traditional Shapley values (Shapely, 1953), implemented by following the Conditional Expectations approach (Lundberg and Lee, 2017). However, what is presented can be extended, without loss of generality, to more advanced approaches, such as the Integrated Gradients Shapley (Sundararajan and Taly, 2017) and the Baseline Shapley value (Sundararajan and Najmi, 2020). The integrated Gradients Shapley method extends Shapley values to the continuous setting. It can be applied to credit lending problems in which the response variable is continuous (such as when the loss-given default is considered the target variable). The Baseline Shapley value overcomes some counterintuitive results of the traditional Shapley approach, such as the assignment of nonzero values to features not used by the model, with a more general approach in which a missing feature for an observation is modelled randomly by drawing it from the sample feature distribution.

3. Methodology

3.1. Credit risk assessment

The evaluation of a company's credit risk depends mainly on its estimated probability of default (PD); that is, the probability that a company will fail to repay its financial obligations. This problem is usually addressed by estimating each company's credit score and setting a threshold to classify it predictively into two main classes: non-default and default. Imagine that information from T explanatory variables of N firms (usually balance sheet indicators) is available. For each firm, we also have a response variable Y, which indicates whether the company has defaulted or is still active (usually in the following period); that is, $Y = 1$ in the case of default and $Y = 0$ otherwise. In the credit scoring model, we aim to find a model that can describe the relationship between T explanatory variables and the response variable Y.

Credit scoring models can be classified into two main categories: black and white boxes. In the former, the relationship between the explanatory variables and the response is not transparent, and only the final classification is observed. Complex ML models such as neural networks, random forests or gradient boosting belong to this category, providing high predictive accuracy at the expense of explainability. In contrast, statistical learning models such as linear and logistic regression are transparent and considered white-box models. These simple models explain how they behave and how predictions are obtained.

3.1.1. Logistic regression

The most commonly used method for credit scoring is logistic regression, a ’white-box’ statistical learning method that finds application in many studies (Murdoch et al., 2019). Logistic regression models classify the response variable into two groups characterised by different statuses (default vs active). More formally, the logistic regression is specified as follows:

$$ln((p_n)/(1-p_n)) = \alpha + \sum\nolimits_{t=1}^{T} \beta_t x_{n_t}, \tag{1}$$

where p_n is the probability of default for the nth firm, $x_n = (x_{n1}, \ldots, x_{nT})$ is the T-dimensional vector of the borrower-specific explanatory variables, parameter α is the model intercept, and β_t is the tth regression coefficient. It follows that the probability of default can be found as

$$p_n = exp(\alpha + \sum\nolimits_{t=1}^{T} \beta_t x_{nt})(1 + exp(\alpha + \sum\nolimits_{t=1}^{T} \beta_t x_{nt}))^{-1} \tag{2}$$

Although the high interpretability of a logistic regression model follows from its explicit functional form, which is linear in the logarithm of odds, its predictive accuracy may be low because of its linear nature. When the available data are large and complex, the predictive accuracy of logistic regression may be inferior to that of a more complex ML model.

3.1.2. Random forests

ML models are increasingly used in complex credit risk assessments (Bussmann et al., 2021). Among them, the random forest classifier, an ensemble of classification trees (Breiman, 2001), performs well in many credit risk classification problems. Like logistic regression, in a classification random forest model, each observation-Ťfor example, a company with its corresponding vector of explanatory variables x_n-Ťis mapped to a default response variable. A random forest classifier merges the rules obtained from a set of classification trees, each based on a training data sample and explanatory variables.

Although each classification tree has explicit rules of construction and allows us to understand how different credit scores are generated, a random-forest model aggregates the scores from each tree on a single average, thereby losing interpretability. A random forest is a black box model which cannot meet the need for explainability in the finance sector (Murdoch et al., 2019). To overcome this limitation, explainable AI models that provide details or reasons to make the functioning of AI clear or easy to understand can be employed.

3.2. Explainable artificial intelligence

Financial institutions and markets are subject to many regulations to maintain the stability of the financial system and protect consumers and investors. An important aspect of financial regulation concerns the supervision of risk management models, particularly credit risk models, for which regulators may seek assurance on the key drivers (Giudici and Hadji-Misheva, 2020). This suggests that black-box AI is unsuitable for credit risk measurement, which motivated the development of XAI models.

The most commonly employed explainable AI model is the Shapley values approach, a model-agnostic post-processing tool used to explain and interpret ML predictions. The Shapley value approach was introduced by Shapely (1953), who leveraged concepts from game theory to map predictive inferences to a linear space.

Specifically, we assumed a game exists for predicting each observation (row). For each game, the players were model predictors (explanatory variables), and the total gain is equal to the predicted value, obtained as the sum of the contributions of each predictor.

Following these assumptions, the Shapley value algorithm calculates the contribution of each variable to each prediction by considering its additional effect on all possible coalitions (groups) of other variables. Specifically, the effect of each variable X_k, for each credit score $i = 1, \ldots, n$, is calculated as follows:

$$\phi(\widehat{f}(X_i)) = \sum_{X' \subseteq \mathscr{C}(X) \backslash X_k} \frac{|X'|!(K-|X'|-1)!}{K!} [\widehat{f}(X' \cup X_k)_i - \widehat{f}(X')_i], \tag{3}$$

where K represents the number of predictors, X' is a subset that contains $|X'|$ predictors, $\widehat{f}(X' \cup X_k)_i$ and $\widehat{f}(X')_i$ are the predictions of the i-th observation obtained with all possible subset configurations, respectively including variable X_k and not including variable X_k. Once Shapley values are calculated for each observation to be predicted, the overall contribution of each predictor, the ‘global’ Shapley value, is obtained as their sum.

3.3. Proposal

We propose employing the global Shapley values of each explanatory variable as the basis of a stepwise variable selection algorithm valid for all models, whether white box or black box.

The algorithm begins with a complete model containing all the available variables. It then removed the variable with the lowest global explainability from the model and evaluated whether the removal significantly decreased predictive accuracy. If so, it stops; otherwise, it proceeds with the deletion of variables until it stops.

A key element of our proposal is a significance test that compares the Area Under the Curve (AUC) of two alternative models differing in the presence of one variable. We recall that the AUC of a model is the most employed predictive accuracy measure for binary variables and is obtained as the area underlying the Receiver Operating Curve (ROC) of a model. The ROC curve was obtained by joining a set of coordinates which represented, for a given set of cutoff points (percentiles), the True Positive Rate against the False Positive Rate. While an ideal model should always have TPR= 1, FPR= 0, and an AUC equal to 1, the higher the AUC, the better the model.

The significance test for the AUC was based on DeLong's test (DeLong and DeLong, 1988). It calculated the Area Under the ROC Curve for each pair of models compared: to model $M_k(k = 1, \dots, K)$ against model M_{k-1}. The test statistic was based on the difference between two AUC values.

More formally, the null hypothesis of the statistical tests is the equivalence of models M_k and M_{k-1}. If the $p$value is more significant than a threshold significance level, such as 5%, we fail to reject the null hypothesis; therefore, model simplification is accepted: variable k is not statistically significant in predicting the response variable.

The stepwise variable selection process continues until the p-value exceeds the set threshold significance level (e.g. 5%). When this occurs, H_0 is rejected such that M_k cannot be simplified to M_{k-1}. Consequently, the procedure stops, and M_k is selected as the final model.

Note that the outlined procedure fully aligns with Occam's razor parsimony principle. If two models have similar predictive accuracy, we choose the simplest of the two (i.e. the one with the lowest number of predictors).

Finally, the choice of the AUC or other test statistics depended on the response variable. For a binary response, AUC is the most commonly employed measure. We employed the Mean Squared Error (MSE) and the corresponding Diebold-Mariano test (Diebold and Mariano, 2002) for a continuous response.

4. Application

4.1. Data

We illustrate the application of our proposal to a large data sample which contains the balance sheet data for over 100,000 SMEs, referred to as the 2020 reporting year. The data were supplied by Modefinance (modefinance.com), a rating agency in a European Credit Assessment Institution (ECAI) supervised by the European Securities and Markets Authority (ESMA), specialising in credit scores for P2P platforms focused on SME commercial lending. The presence of SMEs is a common trait in many countries; therefore, the data can be considered an instance of a more general situation. The companies in the available sample are headquartered in the largest European Union (EU) countries: Italy, France, Spain, and Germany. Their distribution across countries for 2020 is described in Table 1.

From Table 1, note that most companies (49.37%) are located in Italy, with several SMEs. Italy is followed by France, where 27.93% of the companies are headquartered. Germany has the least number of companies in the table, containing only 1.30% of the companies. This is consistent with the fact that although Germany has a larger population than other countries, it does not require public deposits on company balance sheets. Although Germany has a limited number of companies in the sample, limiting its contribution, we prefer not to alter the supplied sample and keep all companies.

Examining the distribution of companies in the sample by 'Industry Sector', which shows to which industrial sector each SME belongs, is interesting. Table 2 lists the five most populated industries.

From Table 2 note that 'Retailing' is the most populated industry, followed by 'Capital Goods', 'Materials', and 'Commercial and Professional Services'.

To estimate a credit scoring model from the data, we need a binary response variable that describes whether a company is in distress (indicating a likely default); and a set of explanatory variables, which may be considered likely causes (or not) of such distress. In the available data, such response variables can be obtained from the variable 'MScore', which is the rating assigned to each company by the rating agency modefinance. MScore can assume a set of ordered values that correspond to ratings of A, AA, AAA, B, BB, BBB, C, CC, CCC, and D, in which 'A′ is assigned to companies with the lowest level of credit risk (lowest probability of default), whereas 'D′ to those with the highest level (highest probability of default).

To convert the variable 'Mscore' into a binary default variable, as in the credit scoring context described in Section 3.1, we associated each company's rating to one of two alternative classes. On the one hand, we associated rating levels C, CC, CCC, and D with a perceived state of default (class 1); on the other, we associated rating levels A, AA, AAA, B, BB, and BBB with a perceived state of non-default (class 0). The resulting percentage of defaulting companies in the available SME sample was 14%.

Table 1
Distribution of Small and Medium Enterprises in the sample by Countries.

Country	No.of.Companies	Percentage
Italy	59,864	49.37
Spain	25,949	21.40
France	33,865	27.93
Germany	1575	1.30

Table 2
Small and Medium Enterprises distribution in the sample by Industry sectors.

Industry Sector	No. Of Companies	Percentage
Retailing	30,201	24.91
Capital Goods	17,536	14.46
Materials	11,969	9.87
Commercial and professional services	10,861	8.96
Food and Staples Retailing	8844	7.29

The distributions of the sample default variable for any given country and industry sector are presented in Figs. 1 and 2. In both figures, the total height of each bar is proportional to the number of companies in each group (country or industry sector), and at the top of each bar, we report the observed default percentages.

From Fig. 1, we can conclude that France is the riskiest country, with the highest default probability of approximately 17.5%. This is followed by Germany, with a 12.9% probability of default; however, its impact on the system is limited, as its frequency is low compared to those of more populated countries such as Italy and Spain. Similar conclusions can be obtained from Fig. 2: 'Consumer services', 'Diversified financials', and 'Media and entertainment' are the riskiest industries, but the impact of the 'Commercial and professional services' sector is higher, being much more populated.

To complete the description of the variables in the sample data, Table 3 shows the considered explanatory variables, which are all financial ratios calculated by modefinance from the 2020 balance sheets of the available companies.

Table 3 indicates that the available explanatory variables are six financial variables which measure, respectively: the operating revenues (Turnover); the financial structure (Leverage); the size (Total Assets), and the profitability (EBIT, Profit and Losses after Tax, Return on Equity) of each considered company, based on the 2020 balance sheets.

Table 4 provides, as summary statistics, the mean of each explanatory variable, separately for the defaulted and the non-defaulted companies.

Comparing the conditional means of each variable in Table 4 EBIT, PLTax, and Leverage present the largest difference between defaulted and non-defaulted companies: they are likely to be the most impactful on the credit scores. Conversely, Turnover and Total Assets show a small difference between the conditional means.

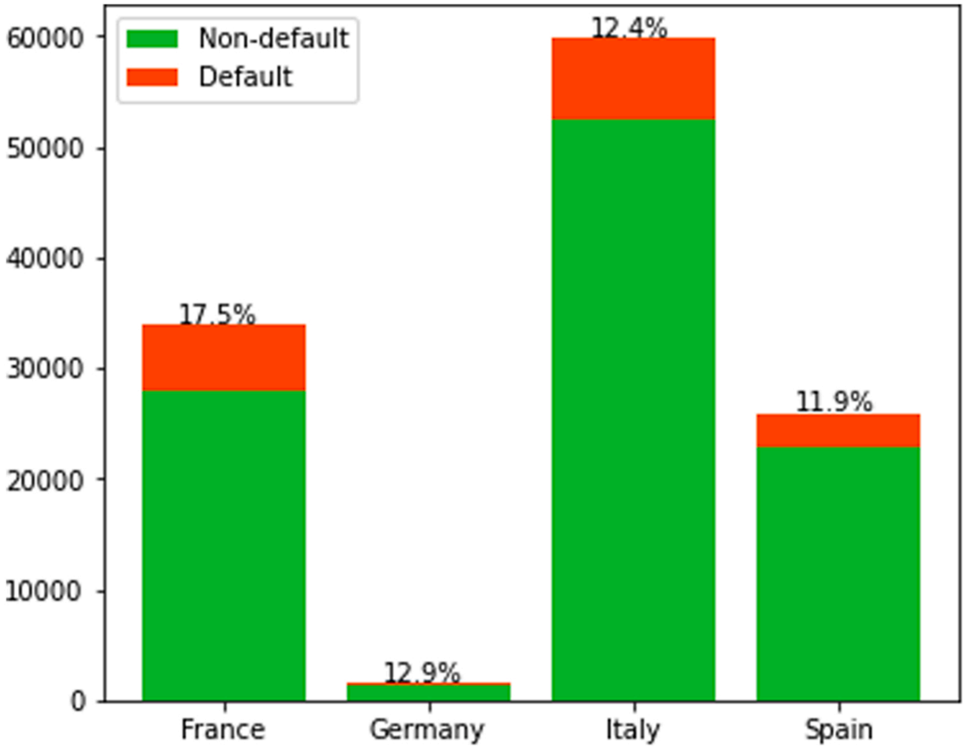

Fig. 1. Distribution of default and non-default SME by country.

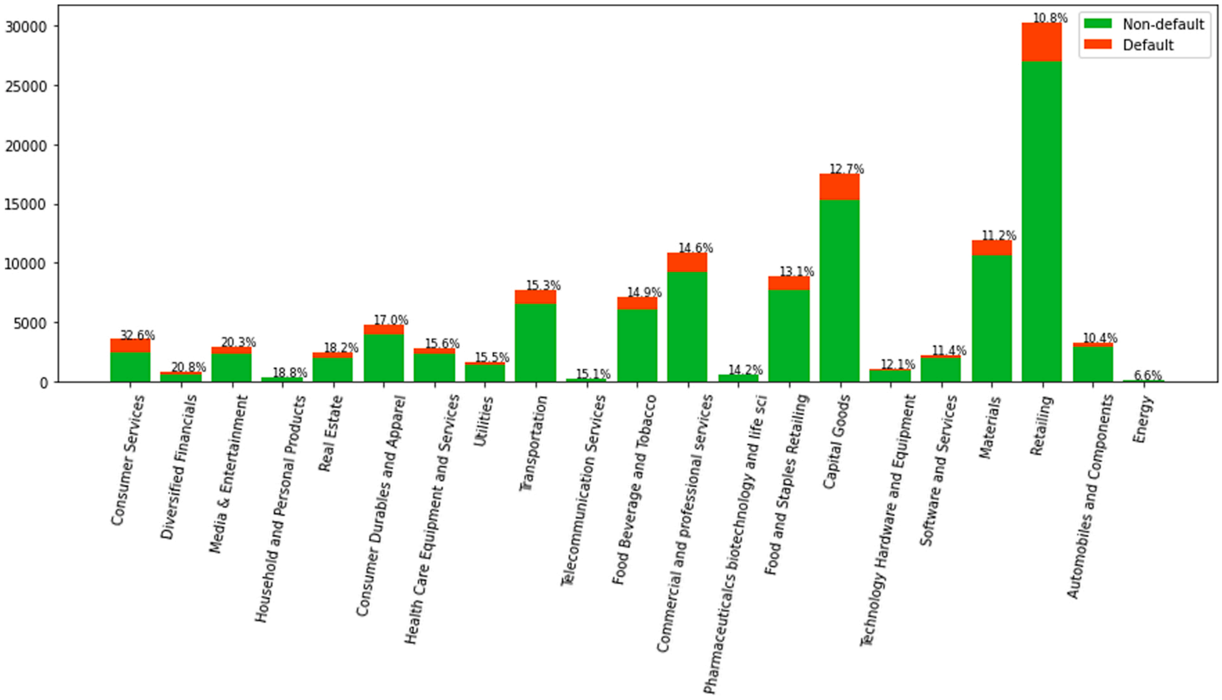

Fig. 2. Distribution of default and non-default SME by industry sector.

Table 3
Description of the Explanatory Variables.

Variable	Description
Turnover	Operating revenues in Thousands of Euro
Leverage	Leverage (ratio)
PLTax	Profit/Loss after tax in Thousands of Euro
TAsset	Total assets in Thousands of Euro
EBIT	Earnings Before Income Tax and Depreciation in Thousands of Euro
ROE	Return on Equity (percentage)

Table 4
Conditional means of the financial variables.

Class	Turnover.2020	EBIT.2020	PLTax.2020	Leverage.2020	ROE.2020	TAsset.2020
Non-Default (Class Zero)	10,950.948104	717.148144	521.539610	4.617414	13.895101	12,560.865164
Default (Class One)	10,261.416051	-828.175527	-1003.877095	994.987954	-4.896900	15,836.216495

Table 5
Estimated coefficients using a full logistic regression credit scoring model.

Variable	Coefficient	*Z*-value	*p*-value
Turnover	-0.001231	-64.568443	0.000001
Leverage	0.000163	3.945933	0.000074
EBIT	-0.001479	-33.018932	0.000001
PLTax	-0.001799	-35.065625	0.000001
ROE	0.000012	2.972022	0.002958
Country	0.130159	-5.213115	0.000001
Industry	-0.552089	-25.116592	0.000001
TAsset	-0.000001	-8.50125	0.003952

4.2. Results

We first build a ‘classic’ credit scoring model based on the logistic regression model in Section 3.1.1. Therefore, we randomly split the data into training (70% of the data) and validation samples (the remaining 30% of the data). For comparison, the same data partitioning was used when applying the random forest model.

Initially, we considered, as explanatory variables, a full model, with all the six financial ratios described in Table 3, along with the Country and Industry sector classes. Applying a logistic regression model to predict a company’s default and a full logistic regression model to the training data led to the estimated coefficients shown in Table 5, along with their corresponding *Z* and *p*values.

From Table 5, we see that all variables are significant, as may be expected, given a large amount of considered training data (more than seventy thousand), which leads to high goodness of fit. Note that Country and Industry have the highest coefficients, but this does not mean they mostly impact the predictions because the variable scales differ. To understand the effect of each variable on the predictions, we employed the estimated model to predict the scores of the companies in the validation sample (30% of all data) and then calculated the Global Shapley values for each variable, summing the Shapley values for the observations in the test set (30% of the observations). The results are presented in Fig. 3.

Fig. 3 shows that ‘PLTax’ has the largest Global Shapley value: it is the variable that mostly impacts the predictions, followed by ‘EBIT’. This result partially aligns with that observed in Table 4, as it consistently indicates the two profitability variables that present the highest difference in conditional means (PLTax and EBIT) but give low importance to financial leverage.

We built an ML credit scoring model based on the random forest model in Section 3.1.2. After splitting data into training (70% of the observations) and validation samples (30% of the observations), with the same partitioning for the logistic regression, we applied the random forest GridSearch CV algorithm of Python to the training sample and used the estimated model to calculate the credit scores of the companies in the validation sample. Each company in the validation sample was then predicted: to default or not to default, comparing the model scores to a set threshold of 0.5. The performance of the full random forest model, which employs all six explanatory variables, is shown in Table 6 in comparison with the logistic regression model previously described.

Table 6 shows that, as expected, the accuracy of the random forest model is higher. The joint consideration of Sensitivity, Specificity, and F1 Score further indicates that the random forest model performs better because it balances sensitivity and specificity better. Consistent with this result, when the set threshold varies from 0.5, as in the *AUC* metrics, the random forest model strongly outperforms the logistic regression, with $AUC = 0.93$ versus $AUC = 0.63$. In conclusion, the available data indicate clear superiority in the predictive accuracy of the random forest credit scoring model over the logistic regression model.

However, although the random forest model is highly accurate, it does not produce a set of estimated coefficients, as shown in Table 5, indicating the relative impact of each explanatory variable. From a managerial perspective, we can predict whether to invest in an SME; however, we do not know why. From an SME perspective, it is unclear which variables improve credit scores.

To overcome this problem, we can post-process the predicted scores obtained with the random forest model using a ‘feature importance plot’. For each variable in the model, the plot represents the decrease in the Gini variability measure determined by each split of the tree induced by a given explanatory variable averaged over all tree models in the random forest built on the training data. Recall that, for a given split induced by an explanatory variable, the higher the reduction in variability, the more important the

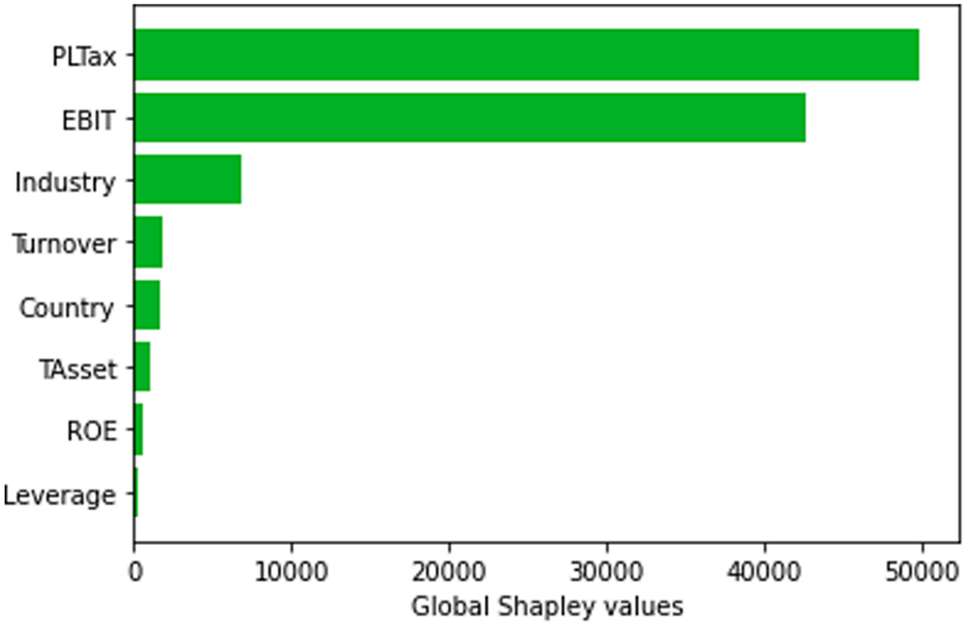

Fig. 3. Global Shapley values based on the predictions generated by the full logistic regression model.

Table 6
Comparison of Logistic Regression and Random Forest full models, regarding predictive accuracy measures.

Measure	Accuracy	Sensitivity	Specificity	F1 Score	AUC
Random Forest	0.97066	0.98687	0.86884	0.89051	0.92785
Logistic Regression	0.89646	0.98748	0.32459	0.46261	0.65603

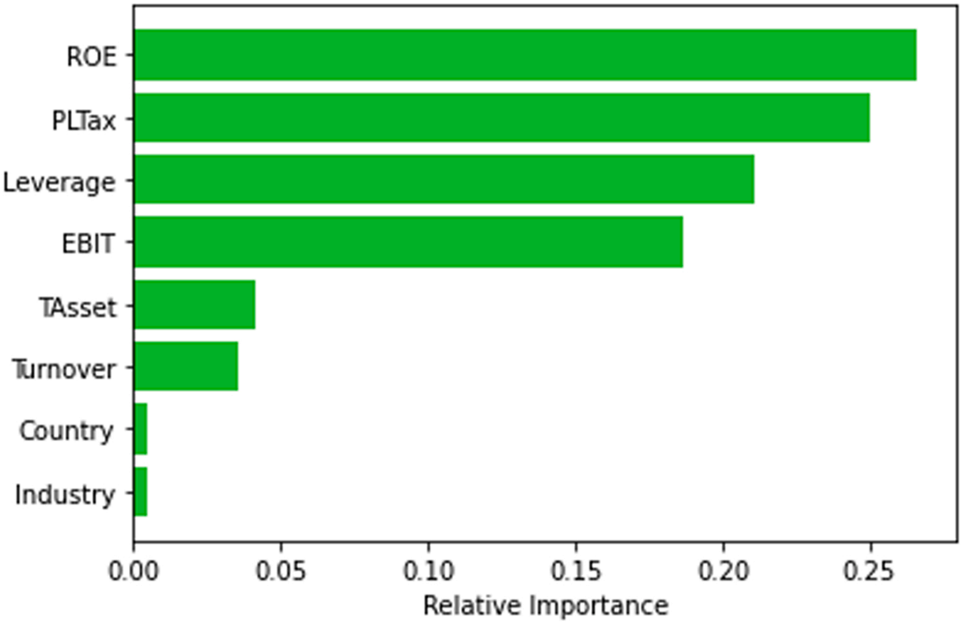

Fig. 4. Random Forest Feature Importance.

variable. The feature importance plot for our considered training data is shown in Fig. 4.

Fig. 4 shows that ROE, PLTax, Leverage, and EBIT lead to the highest reduction of the Gini measure and are thus individuated as the most important variables. However, the lowest importance is related to Industry and Country, the only two variables in the sample that are not derived from the balance sheet.

Although the feature importance plot addresses, to some extent, explainability, it is not model-agnostic; it cannot be obtained for models different from random forests, such as logistic regression. Consequently, it does not allow a comparison of model explainability. Thus, we resort to a model-agnostic tool, Shapley values, calculated from the predicted credit scores in their validation sample.

Fig. 5 shows the overall contribution of each variable, as described by the global Shapley values: the Shapley values for each variable, summed across all observations.

From Fig. 5 note that the most explainable variable is ‘Leverage’, followed by ‘PLTax’ and ‘EBIT’, consistently with the difference in conditional means, and with the feature importance plot in Fig. 4, but differently from what obtained applying Shapley values to logistic regression. Here, ‘Leverage’ is the least important variable. The same figure shows that the global Shapley values for ‘Country’ and ‘Industry’ are small, consistent with the feature importance plot but different to what occurs for the logistic regression in Fig. 3.

From a financial viewpoint, the Shapley values of the random forest credit scores indicate that the probability of default of an SME is mainly determined by its probability (as measured by ROE, PLTax, and ROE) and by its financial leverage; less affected by its size and operating revenues (as measured by Tasset and Turnover); and little affected by its corresponding Country or Industry, differently from what occurs using a logistic regression model.

To understand which variables are statistically significant to explain the probability of default, we applied our proposed selection procedure: a stepwise variable selection based on the comparison of the *AUC* and the ordering established by the Global Shapley values in Fig. 5. Our procedure differed from classical stepwise procedures that compare models in terms of their likelihood. Instead, we compared the models in terms of their predictive accuracy. The advantage of doing so is generality: we can compare models with an underlying probabilistic model, such as logistic regression and all ML models, such as random forest models.

More precisely, we employed a backward selection procedure, which progressively eliminated variables from the full model, following the order determined by the global Shapley values in Fig. 5: from the least explainable (‘Industry’) to the most explainable (‘ROE’). Each variable was removed from the least explainable to the most explainable variable. Specifically, a variable is removed from the model when its additional contribution to predictive accuracy, as measured by the Area Under the Receiver Operating Characteristics (*AUC*), is not statistically significant; that is, it leads to a DeLong test with *p*value larger than a threshold (e.g. 5%). The procedure was stopped when *p* was lower than the set threshold. As a result of our proposed procedure, the selected model is highly

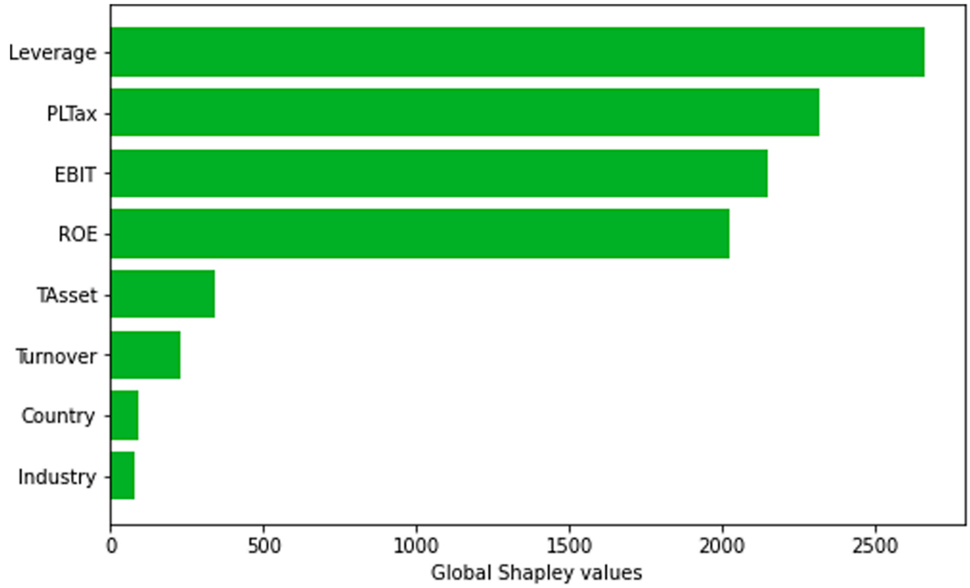

Fig. 5. Global Shapley values importance.

Table 7
DeLong Tests of the considered pairs of Random Forest models.

Removed Variable	Number of Variables in model	AUC	P-value	H Measure
-	8	0.927856	-	0.818273
Industry	7	0.927604	0.167723	0.817766
Country	6	0.928045	0.258918	0.818519
Turnover	5	0.924413	0.000009	0.808577

predictive and explainable using a parsimonious set of predictors.

Table 7 lists the results of this stepwise procedure. It contains the *AUC* values and the *p*-values of the DeLong test, corresponding to comparing subsequent pairs of *AUC*s.

Table 7 shows that 'Industry' is the least explainable variable: the first candidate for removal. The comparison of the *AUC* of the entire model, when all predictors are used in the random forest model, against the model without 'Industry', leads to a *p*-value of the DeLong test equal to 0.16772, leading to select the simpler model, without 'Industry'.

The next variable candidate for removal is 'Country'. Comparing the *AUC* of the model without 'Industry' and 'Country' against the model which excludes only 'Industry', we found that the *p*-value of the DeLong test equals 0.25892, so the model can be further simplified.

The procedure continues until a variable whose exclusion leads to a significant decrease in predictive accuracy is identified. In our case, this is the third most explainable variable, 'Turnover', for which the *p*-value is smaller than the threshold, leading to a rejection of model simplification and stopping the variable removal procedure. In conclusion, from Table 7, we obtain that the best trade-off between explainability and predictive accuracy is provided by a model that includes all available financial indicators but not the variables which describe the 'Industry' and the 'Country' of the companies.

From a financial viewpoint, this result indicates that the binarised ratings assigned by the rating agency are 'fair' across countries and sectors, with no bias in terms of financial inclusion.

To verify the results obtained in Table 7, we provide the results from applying Hand's *H* statistics Hand (2009) to our models. The results are consistent with those of the *AUC*: the values for the *H* measure are similar, approximately 0.818, for the first three models, before removing 'Turnover' and, when 'Turnover' is removed, *H* drops down to 0.808577, showing that the model should not be simplified any further, consistent with the results obtained applying DeLong's test to the *AUC*s.

Thus, our proposed selection procedure leads to a simpler random forest credit scoring model than the entire model, with six variables instead of eight. However, this does not result in a significant loss of predictive accuracy.

For completeness and comparison, we should apply our proposed stepwise procedure also to the logistic regression scoring model, following the variable ordering determined by Fig. 3. In this case, removing 'Leverage', the least explainable variable in Fig. 3, leads to a *p*-value smaller than 0.05. Hence, the null hypothesis is rejected, and the full model cannot be simplified without a significant loss of accuracy. Thus, the selected random forest model with six variables is more parsimonious than the selected logistic regression model, a full model with eight variables.

Therefore, we conclude that the random forest model selected by our proposed procedure is more accurate and parsimonious than the selected logistic regression model.

5. Conclusions

Ensemble ML models, such as random forests, can improve the accuracy of credit scoring models but are not explainable. Explainable AI methods such as Shapley values can be employed to post-process credit scores to achieve explainability.

This study employed Shapley values to achieve explainability and guide variable selection, leading to a parsimonious model that is a good trade-off between predictive accuracy and explainability.

To achieve this goal, we proposed a model selection strategy in which global Shapley values ordered the candidate explanatory variables in terms of their predictive importance, and a backward stepwise selection procedure, based on the comparison of predictive accuracy, was implemented to select a 'statistically optimal' subset of variables.

Our proposal is applied to a database containing credit ratings for a large set of European SMEs, the values of six financial ratios from their 2020 balance sheets, and their country and sector of belonging. These results indicated that the nonlinear random forest credit scoring model was more accurate than the logistic regression. The application of our procedure also showed that the selected random forest model was more parsimonious than the selected logistic regression model because it depended only on balance sheet ratios and not on the country or industry sector of a company, with no bias in terms of financial inclusion.

From a methodological viewpoint, our proposed method: i) fills a gap, as it provides a model comparison procedure based on both accuracy and explainability, which can be equally applied to all types of ML models; and ii) leads to a credit scoring model which is a good trade-off between predictive accuracy and explainability.

From a managerial viewpoint, our model can support banks, fintech companies, and regulators in developing and supervising ML models for credit scoring compliant with regulatory requirements, particularly those concerning AI.

Our proposal makes three main contributions to literature. For research scholars, it proposes a novel model comparison approach, which combines explainability with predictive accuracy; for financial and fintech managers, it proposes a way to make AI applications explainable and, therefore, acceptable; for policymakers and regulators, it provides a methodology able to check whether a specific AI application for credit scoring is compliant with the existing regulations.

Our study is built on the standard axiomatisation of the Shapley value, which is only suitable for binary responses such as credit default. However, when continuous response variables such as loss given default or exposure at default are considered, the proposed method can be extended by considering the Baseline Shapley value (Bshap) or the Random Baseline Shapley value (Sundararajan and Najmi, 2020) when implementing random forest or other ML approaches.

Further research is needed to experiment with the proposed model selection procedure using alternative Shapley axiomatisations. An interesting avenue of research would be to understand the impact of balance sheet variables on companies' financial exposure by extending the linear regression analysis of Fasano and Cappa (2022) to an ML context.

Further research is also needed to consider the interpretation of the predictions in terms of their 'fairness', that is, to establish how independent the credit scores from country and industry sector or, for consumer credit applications, from gender, race or other types of stratifications.

Acknowledgments

The authors acknowledge support from the European Commission Horizon2020 research project PERISCOPE: Pan-European Response to the ImpactS of COvid-19 and future Pandemics and Epidemics, contract number n. 101016233. The authors also specify that the paper is the result of a close collaboration among the three authors.

References

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: a Survey on Explainable Artificial Intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Ayadi, R., Bongini, P., Casu, B., and Cucinelli, D., 2021, Bank Business Model Migrations in Europe: Determinants and Effects.British Journal of Management, 32(4), 1007–1026, 10.1111/1467–8551.12437.

Babaei, G., & Bamdad, S. (2020). A multi-objective instance-based decision support system for investment recommendation in peer-to-peer lending. *Expert Systems with Applications, 150*, Article 113278. https://doi.org/10.1016/j.eswa.2020.113278

Bastani, K., Asgari, E., & Namavari, H. (2019). Wide and deep learning for peer-to-peer lending. *Expert Systems with Applications, 134*, 209–224. https://doi.org/10.1016/j.eswa.2019.05.042

Berger, A. N., & Udell, G. F. (2006). A more complete conceptual framework for SME finance. *Journal of Banking and Finance, 30*(11), 2945–2966. https://doi.org/10.1016/j.jbankfin.2006.05.008

Bracke, P., Datta, A., Jung, C., and Sen, S., 2019, Machine learning explainability in finance: an application to default risk analysis.10.2139/ssrn.3435104.

Breiman, L. (2001). Random Forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

Bücker, M., Szepannek, G., Gosiewska, A., & Biecek, P. (2022). Transparency, Auditability and eXplainability of Machine Learning Models in Credit Scoring. *Journal of the Operational Research Society, 73*(1), 70–90. https://doi.org/10.1080/01605682.2021.1922098

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics, 57*, 203–216. https://doi.org/10.1007/s10614-020-10042-0

Chopra, A., & Bhilare, P. (2018). Application of ensemble models in credit scoring models. *Business Perspectives and Research, 6*(2), 129–141. https://doi.org/10.1177/2278533718765531

Correia, F., Martins, A., & Waikel, A. (2022). Online financing without FinTech: Evidence from online informal loans. *Journal of Economics and Business, 121*, Article 106080. https://doi.org/10.1016/j.jeconbus.2022.106080

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics, 44*(3), 837–845. https://doi.org/10.2307/2531595

Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business and Economic Statistics, 20*(1), 134–144. https://doi.org/10.1198/073500102753410444

Djeundje, VB, C. J, C. R., & Hamid, M. (2021). Enhancing credit scoring with alternative data. *Expert Systems with Applications, 163*. https://doi.org/10.1016/j.eswa.2020.113766

Dushimimana, B., Wambui, Y., Lubega, T., & McSharry, P. E. (2020). Use of Machine Learning Techniques to Create a Credit Score Model for Airtime Loans. *Journal of Risk and Financial Management, 13*(8), 180. https://doi.org/10.3390/jrfm13080180

Fasano, F., & Cappa, F. (2022). How do banking fintech services affect SME debt? *Journal of Economics and Business, 121*, Article 106070. https://doi.org/10.1016/j.jeconbus.2022.106070

Ferri, G., & Murro, P. (2015). Do firm-bank 'odd couples' exacerbate credit rationing? *Journal of Financial Intermediation, 24*(2), 231–251. https://doi.org/10.1016/j.jfi.2014.09.002

Finlay, S. (2011). Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research, 210*(2), 368–378. https://doi.org/10.1016/j.ejor.2010.09.029

Giudici, P., Hadji-Misheva, B., & Spelta, A (2020). Network based credit risk models. *Quality Engineering, 32*(2), 199–211. https://doi.org/10.1080/08982112.2019.1655159

Giudici, P., & Raffinetti, E. (2021). Shapley lorenz explainable artificial intelligence. *Expert Systems with Applications, 114104*, 167.

Guo, Y., Zhou, W., Luo, C., Liu, C., & Xiong, H. (2016). Instance-based credit risk assessment for investment decisions in P2P lending. *European Journal of Operational Research, 249*(2), 417–426. https://doi.org/10.1016/j.ejor.2015.05.050

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of Data Mining*. MIT Press,.

Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning, 77*(1), 103–123. https://doi.org/10.1007/s10994-009-5119-5

Henley, W., & Hand, D. J. (1996). A k-nearest-neighbour classifier for assessing consumer credit risk. *Journal of the Royal Statistical Society: Series D (The Statistician), 45*(1), 77–95. https://doi.org/10.2307/2348414

Kendall, J. (2017). Fintech companies could give billions of people more banking options. *Harvard Business Review*, 1.

Kumari, B., Kaur, J., and Swami, S. (2021). System Dynamics Approach for Adoption of Artificial Intelligence in Finance. In Advances in Systems Engineering: Select Proceedings of NSC 2019 (555–575).Springer.

Laborda, J., & Ryoo, S. (2021). Feature selection in a credit scoring model. *Mathematics, 9*(7), 746. https://doi.org/10.3390/math9070746

Lee, T.-S., & Chen, I.-F. (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications, 28*(4), 743–752. https://doi.org/10.1016/j.eswa.2004.12.031

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research, 247*(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Li, Y., & Chen, W. (2020). A comparative performance assessment of ensemble learning for credit scoring. *Mathematics, 8*(10), 1756. https://doi.org/10.3390/math8101756

Liu, W., Fan, H., & Xia, M. (2022). Credit scoring based on tree-enhanced gradient boosting decision trees. *Expert Systems with Applications, 189*, Article 116034. https://doi.org/10.1016/j.eswa.2021.116034

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 30.

Milne, A. and Parboteeah, P., 2016, The Business Models and Economics of Peer-to-Peer Lending. ECRI Research Report dl, 10.2139/ssrn.2763682.

Moscato, V., Picariello, A., & Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications, 165*, Article 113986. https://doi.org/10.1016/j.eswa.2020.113986

Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B., 2019, Definitions, methods, and applications in interpretable machine learning.Proceedings of the National Academy of Sciences, 116(44), 22071–22080, 10.1073/pnas.1900654116.

Ribeiro, M.T., Singh, S., and Guestrin, C., 2016a, Model-Agnostic Interpretability of Machine Learning. arXiv:http://arXiv.org/abs/arXiv:1606.05386, 10.48550/arXiv.1606.05386.

Ribeiro, M.T., Singh, S., and Guestrin, C.(2016b)Why Should I Trust You? Explaining the Predictions of Any Classifier.in Proceedings of the 22nd ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (1135–1144).

Romānova, I. and Kudinska, M. (2016). Banking and Fintech: A Challenge or Opportunity? In Contemporary issues in finance: Current challenges from across Europe, 98 (21–35). Emerald Group Publishing Limited.

Sachan, S., Yang, J.-B., Xu, D.-L., Benavides, D. E., & Li, Y. (2020). An explainable ai decision-support-system to automate loan underwriting. *Expert Systems with Applications, 144*, Article 113100. https://doi.org/10.1016/j.eswa.2019.113100

Serrano-Cinca, C., & Gutiérrez-Nieto, B. (2016). The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. *Decision Support Systems, 89*, 113–122. https://doi.org/10.1016/j.dss.2016.06.014

Shapely, L. (1953). A value for n-person games. *Contributions to the Theory of Games, volume II* (pp. 307–317). Princeton University Press,.

Shen, F., Zhao, X., & Kou, G. (2020). Three-stage reject inference learning framework for credit scoring using unsupervised transfer learning and three-way decision theory. *Decision Support Systems, 137*, Article 113366. https://doi.org/10.1016/j.dss.2020.113366

Srinivasan, V., & Kim, Y. H. (1987). Credit Granting: A Comparative Analysis of Classification Procedures. *The Journal of Finance, 42*(3), 665–681. https://doi.org/10.1111/j.1540-6261.1987.tb04576.x

Sundararajan, M. and Najmi, A. (2020). The many Shapley values for model explanation.In International conference on Machine Learning(9269–9278).: PMLR.

Sundararajan, M., Taly, A., and Yan, Q.(2017). Axiomatic Attribution for Deep Networks.In International Conference on Machine Learning (3319–3328).: PMLR.

Temelkov, Z. (2018). Fintech firms opportunity or threat for banks? *International Journal of Information, Business and Management, 10*(1), 137–143.

Tripathi, D., Shukla, A. K., Reddy, B. R., Bopche, G. S., & Chandramohan, D. (2022). Credit scoring models using ensemble learning and classification approaches: a comprehensive survey. *Wireless Personal Communications*, 1–28. https://doi.org/10.1007/s11277-021-09158-9

Trivedi, S. K. (2020). A study on credit scoring modeling with different feature selection and machine learning approaches. *Technology in Society, 63*, Article 101413. https://doi.org/10.1016/j.techsoc.2020.101413

Tyagi, S. (2022). Analyzing machine learning models for credit scoring with explainable ai and optimizing investment decisions. *American International Journal of Business Management, 5*(1), 1–161.

West, D. (2000). Neural network credit scoring models. *Computers and Operations Research, 27*(11–12), 1131–1152. https://doi.org/10.1016/S0305-0548(99)00149-5

Xia, Y., Yinguo, L., Lingyun, H., Yixin, H., & Yigun, M. (2021). Incorporating multilevel macroeconomic variables into credit scoring for online consumer lending. *Electronic Commerce Research and Applications*, 10195. https://doi.org/10.1016/j.elerap.2021.101095

Yobas, M. B., Crook, J. N., & Ross, P. (2000). Credit scoring using neural and evolutionary techniques. *IMA Journal of Management Mathematics, 11*(2), 111–-125. https://doi.org/10.1093/imaman/11.2.111