Available online www.ejaet.com

European Journal of Advances in Engineering and Technology, 2023, 10(4):86-96

Research Article ISSN: 2394 - 658X

An Explainable AI Model in Fintech Risk Management in Small and Medium Companies

Praneeth Reddy Amudala Puchakayala[1], Saurabh Kumar[2], Shafeeq Ur Rahaman[3]

[1]Data scientist, Regions Bank, AL, USA
[2]Kraft Heinz Foods, Senior Manager, Data Science
saurabh.hoa@gmail.com
[3]Monks, CA, USA

ABSTRACT

Financial technology is becoming more important in lending to small and medium businesses (SMEs). Due to advanced machine learning (ML) algorithms, this is now feasible; these algorithms can accurately predict a business's financial performance with the data that is currently available. Even though ML models are quite good at making predictions, they might not give users enough context for the outcomes. For instance, according to the recently proposed artificial intelligence (AI) laws, it could not be sufficient for making informed decisions. Applied to the context of model selection, Shapley values allowed us to close the gap. In light of this, we provide a state-of-the-art model selection method that is predictively accurate and can be used with any ML model, even with a probabilistic basis. We tested our proposal using a credit-scoring database with information on more than 100,000 SMEs. The empirical results show that a certain small and medium-sized firm (SME) can have its investment risk forecasted and understood using an accurate and explainable machine-learning model.

Keywords: AI Model, Risk Management, Small and Medium Companies, Explainable AI.

INTRODUCTION

Mystery container Regulated financial services do not fit artificial intelligence (AI) well. To overcome this issue, we need AI models that can be explained, which give reasoning and details to make AI work. Defining "Explainable" is the first step in developing such models. This year has seen the provision of certain crucial institutional benchmark definitions. Within the framework of the EU, we detail a few of them. "Explainability means that an interested stakeholder can comprehend the main drivers of a model-driven decision," according to the Bank of England [1]. "The lack of interpretability and auditability of AI and ML methods could become a macro-level risk," said the Financial Stability Board [2]. "The law may dictate a degree of explainability in some cases," the UK Financial Conduct Authority states, concluding their statement.

"The existence of automated decision-making should carry meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject," stated the European GDPR legislation. Therefore, under certain conditions, data subjects are entitled to obtain relevant information regarding the reasoning behind automated decision-making as per the GDPR. In addition, in April 2019, the Ethics Guidelines for Trustworthy Artificial Intelligence were released by the European Commission's High-Level Expert Group on AI. All artificial intelligence systems must fulfill the seven conditions stated here before being considered reliable. Only three of these are pertinent to XAI. Both human agency and monitoring require that decisions be based on solid information and that individuals be kept aware of developments. Openness: When communicating with stakeholders, it's crucial to personalize explanations of AI systems and their results.

Being aware that one is engaging with an AI system is essential for humans. Accountability: Accountability and responsibility, openness, and assessment of data, algorithms, and design processes should all be part of AI systems.

Many companies, big and small, have begun to use Explainable AI (XAI) models in response to the necessity to explain AI models voiced by lawmakers and regulators in many countries [3]. In mathematics, it is well-known that "simpler" statistical learning models, like logistic and linear regression models, offer good interpretability but may

have poor predictive accuracy. On the other hand, "more complicated" machine learning models, like neural networks and tree models, offer great predictive accuracy but poor interpretability.

We propose a novel approach to improving highly accurate machine learning models that can justify their anticipated output, thus resolving this trade-off. We propose a methodology that runs in post-processing rather than during analysis preparation. It maintains its agnosticism (technical neutrality) when applied to all three types of prediction models: neural networks, classification trees, or linear regression [4]. We base our proposed strategy on Shapley's principles. Our main area of interest is P2P lending, a form of financial technology driven by artificial intelligence. When many SMEs apply for loans on a P2P lending platform, the credit risk of these businesses is calculated using Shapley values.

Based on the data we collected, the results are as interpretable as, if not more so, a typical logistic regression model, and the predictions are even more accurate.

Recent years have seen tremendous advancements in artificial intelligence (AI), with major applications by internet giants like Amazon, Google, and Meta (previously Facebook). Examples of the services and products these corporations have merged with AI are the intelligent robotics applications sold by Amazon, the recommendation systems offered by Meta, and the AI-powered search algorithms provided by Google. Hence, artificial intelligence will exert an even greater influence in the future. The artificial intelligence market will likely expand further in the financial sector. Regardless, these advancements have led to AI models being opaque black boxes, which makes them harder to understand and work with. Consequently, creating and assessing remedies to this issue is critical, particularly in delicate domains like banking, where widespread use of AI systems necessitates more than pinpoint accuracy.

Machine Learning (ML) algorithms focus on algorithms that can learn from data and then make predictions and judgments on their own, which is why it is considered a subset of AI [6]. More than only machine learning, AI has many other potential uses. Machine learning (ML) models may be essential to certain AI systems, whereas rule-based systems may be the only means they function. Although ML is a part of AI, the fact that it is not the only method used in all AI applications demonstrates the area's diversity and complexity.

Explainable AI (XAI) aims to offer a framework for studying, describing, and comprehending complicated systems. Nevertheless, evaluating and determining whether an explanation is explicable is a difficult and complicated problem. Establishing trust and responsibility in decision-making requires providing accessible and interpretable financial explanations. Authors frequently draw parallels between the areas of finance and medicine to stress the necessity of explanations in both sectors, and XAI research supports this idea [7].

Alle ML community has settled upon no universally accepted yet of bacteria for explainability as yet sought to elucidate the meaning of terms like dependability, interpretability, trustworthiness, and explainability [8]. Interpretability and explainability are often used interchangeably, even though many research differentiate between the two. Imprecise concepts, common in explainable AI and interpretability, can cause people to draw the wrong conclusions. One measure of a model's interpretability is how well its reasoning can be understood and articulated. An AI system's underlying algorithms and logic are easier to know when users fully grasp the model's general notions [9]. Explainability refers to how clear the results of a model are.

On the flip side, the explainability of a machine learning system refers to how well humans can understand its inner workings and reasoning. When a model's training or decision-making processes demonstrate a high level of comprehension regarding these internal processes, its explainability increases. The argument goes like this: interpretability isn't enough; explainability is a prerequisite. Research by [10] suggests that interpretability is a more all-encompassing notion than explainability. Nevertheless, we used interpretable and explainable equally so that their relevance was not limited to certain situations. You can only put your faith in institutions with clear ways to make decisions. Regarding financial AI applications like credit scoring, bankruptcy forecasts, fraud detection, and portfolio optimization, XAI is head and shoulders above the competition thanks to its sophisticated models and ability to account for the consequences of biased or inaccurate results. Some potential future research directions include investigating the ethical and legal ramifications of employing opaque AI systems in the financial sector and how XAI could enhance the procedures for managing investment portfolios and making investment decisions.

In addition, XAI methods can make predictive models more understandable in many domains, including regulatory compliance, stock price prediction, insurance pricing, and loan underwriting.

When discussing AI techniques used in academic research, the words ML and AI are often thought to mean the same thing. Since this method incorporates several AI techniques, it provides a broader area picture. By using this terminology interchangeably, researchers can more effectively convey their findings to a wider audience and maintain clarity and consistency throughout the study.

Applying our proposed method, we generated credit scores that correctly predicted the probability of default (PD) for a collection of enterprises using random forest models. This allowed us to evaluate potential solutions.

Our enhanced credit scoring model and its increased accessibility for use in fintech lending platforms resulted from this improved utilization of the data and its nonlinear linkages [11]. When it comes to investing in small and

medium-sized businesses (SMEs), our concept clarifies for management what elements contribute to credit risk. We are unaware of any previous methodological study that used the XAI to select variables for credit scoring.

LITERATURE REVIEW

XAI in general

XAI's open-source AI integration initiative aims to help end users understand, trust, and operate AI systems more efficiently. The necessity for explanations was recognized in the early stages of rule-based expert systems [12]. As a result, XAI became a hot topic in academia with the introduction of Deep Learning (DL) systems. Numerous real-world situations call for ML models with ever-increasing predictive accuracy criteria, such as stock price prediction, online banking fraud detection, and bankruptcy prediction.

Although the model's predicted accuracy increases as its complexity increases, its ability to produce explainable predictions falls. For optimal performance, go with black-box models. These include ensemble models (e.g., XGBoost and Random forests) and deep learning models (e.g., generative adversarial networks, DNNs).

Nevertheless, as mentioned in [13], some models refuse to be explained. Conversely, white-box or intrinsic models provide a straightforward structure that clarifies the outcomes. This category includes decision trees, rule-based, and linear models.

Given the abundance of previous research on XAI, we will mostly focus on outlining its essential concepts [14]. Both researchers and industry professionals use a lot of jargon when describing the key aspects of XAI systems. Although some methods are model-specific, others are not model-specific at all. The strategies also differ in terms of stage (ante-hoc vs. post-hoc), scope (local vs. worldwide), and application field (e.g., finance, medical, education, transportation, ecology, agriculture, etc.).

The two steps of an explanation are Data interpretation and analysis employing explanatory data, data source specification, and model construction are all part of the ante-hoc stage. Now is the time to evaluate the data because doing so increases our model comprehension and yields substantial discoveries [15]. These strategies are called white-box approaches because of their intentional engineering to maintain a basic structure. They do not necessitate any explanation because they are inherent. In the post-doc phase, an additional explanation method is needed to produce an explanation after the ML model has been built.

One way to classify explainability is by whether it is based on a specific data point (local) or the complete model (global).

Model-agnostic approaches can operate independently without relying on any one ML model. In most cases, these methods are implemented after the fact and attempt to address the challenge of understanding complex models like Convolutional Neural Networks (CNN). Since they don't restrict themselves to one model structure, they can be easily modified and used with a wide range of models [16]. Conversely, the model-specific technique is structure-dependent and only useful for that particular model.

Practicality, ethics, and the law are reasons for wanting something to be explainable [17]. The EU governs artificial intelligence through the AI Act. The creation and usage of XAI must adhere to certain standards set out by the regulation. For many reasons, building complex XAI models is difficult to achieve explainability.

More complex models are usually required for better accuracy, raising an additional concern about the performance vs. explainability trade-off. Some of the challenges they address include identifying critical features, concerns about scaling, model acceptance, drawing attention to the requirement for substantial computational resources, and ensuring that explanations align with user intuitions.

Artificial intelligence in finance

Governments' concerted attempts to control AI use and practical applications highlight the technology's importance. One of these initiatives that stands out is the High-Level Expert Group on AI (HLEG) that the European Commission intends to establish in 2020. HLEG hopes to facilitate the ethical development and usage of robust AI systems by outlining guidelines and proposing a precise definition of AI. The major goal of HLEG is to discuss and improve policies on the social, ethical, and legal aspects of AI. In their discussion of AI, HLEG describes a system in which hardware and software work together to gather and analyze data from the physical world.

Using this analysis, the AI system learns new things and makes judgments to accomplish goals. To build its flexibility, the AI system looks at its previous activities and how they affected the operational environment. To carry out this assessment, one can use numerical models or symbolic rules [18]. Governments and expert bodies like HLEG are working hard to acknowledge the importance of AI and integrate it into many areas in a responsible and helpful way.

Several sectors, notably the financial sector, have found AI innovative and game-changing. Financial institutions that employ AI methods must have the capacity to forecast insolvency. Pay close attention to this work. Multiple research [19] have shown that AI has enormous potential to change financial decisions, reduce risks, and increase profitability. Financial organizations can improve customer service and acquire a competitive edge by utilizing AI. The categories employed in the economic arena were derived from an extensive, multi-dimensional, and problem-oriented economic-financial examination of previous research on AI in lending.

While AI has the potential to enhance and provide novel financial solutions, more obstacles remain to utilize cutting-edge algorithms fully.
According to recent studies [20], the banking industry is facing serious problems when implementing AI.
These obstacles include issues with localization, competency gaps, ethics, legality, trust, transparency, integrity, and the complexities of ML design and integration.
Scientists typically use a combination of ML methods to assess and enhance the performance of their models. Therefore, we defined a "multi-approach AI technique" as a set of procedures for solving a specific financial problem using a combination of different AI methods.

XAI in finance
According to [21], people should learn about AI because it's increasingly used in decision-making. This is necessary because there are worries about prejudice, developers' capacity to comprehend AI systems, and following rules and laws. They state that enhanced AI explainability can achieve more trust in the outcomes. A further rationale in support of XAI's objective of developing more comprehensible models—while simultaneously preserving efficiency and granting humans control over AI systems—is that interpretability aids in guaranteeing that AI decision-making is reliable, truthful, and equitable.
Discovery, control, and the enhancement of classification or regression tasks are among the several uses, viewpoints, and interpretations discovered in [22]. These viewpoints span a wide range of topics, from describing and guiding AI/ML methodologies to finding new insights and enhancing the accuracy of classification or regression tasks.
Guaranteeing the interpretability of AI/ML methods is essential to determining which input features substantially impact the results. When a model is completely understood, it can be enhanced by combining it with specific knowledge. Research on XAI in finance often focuses on credit rating and risk management.
Explainability categories
In published surveys like [23], researchers have examined and reported on the thorough examination of XAI methodologies, which includes a wide range of procedures, techniques, and performance measurements. I created a classification system to group explanatory approaches. It's now utilized for analyzing and comparing various XAI strategies and practices. It's great for learning about the pros and cons of each tool.
Feature relevance explanation: One major advancement in feature importance explanation, especially in XAI, is the Shapley Additive Explanations (SHAP) method, which aims to evaluate the impact of a feature on the model's output.
This approach builds a linear model around the example that needs explaining, and the features' relevance is deduced from the coefficients. Since this method does not reveal the interdependencies between features but concentrates on each feature's contribution independently, it is considered a roundabout way to generate explanations. When there are a lot of interrelated traits, the final scores could provide conflicting or inconclusive results. This must be taken into account when an analysis is being carried out. To illustrate the process of relevant feature selection, the study by [24] introduces a novel XAI model that can independently determine the causes of financial crises. The writers employed the pigeon optimizer to facilitate the feature selection procedure. After that, we use a gradient-boosting classifier to find the source using a subset of the most important features.
Explanation by simplification: Complicated Models are approximated to simpler ones through explanations by simplification. The simplification challenge primarily arises from comparing the simpler model's performance on classification problems to guarantee that it is flexible enough to mimic the complicated model and enhance its efficacy appropriately. Model-independent explanations can be made easier using rule-based learners and decision tree techniques. Furthermore, decision trees, rule-based learners, and distillation can all simplify explanations while providing model-specific details. Weighted Soft Decision Forest (WSDF) is one example of this type of explanation.
This method combines the results of many soft decision trees using weights. We want it to feel as natural as a credit score. Decision trees and the recursive rule extraction approach are also used in decision support systems for credit risk assessment.
These algorithms produce rules that humans can understand and use for credit assessment methods that rely on machine learning.
Local explanation: An example of local explainability would be a model's ability to shed light on a particular case's predictions down to the statistical unit level. Businesses and individuals benefit from this method when pinpointing the root causes of their financial problems. Consideration of alternative scenarios, rule-based learning, and linear approximations are all examples of local explanation approaches. One famous ML tool is the Local Interpretable Model-Agnostic Explanations (LIME) method, created by [25].
It finds the important features needed to produce predictions by removing input perturbations and helps to understand the behavior of black-box models. Despite being a locally linear model, LIME's accuracy could be compromised because it relies on another model. To better comprehend model predictions and suggestions for

future action, counterfactual explanations offer hypothetical causal situations in which the lack of Event A would lead to the absence of Event B. I also demonstrated how to use LIME in bankruptcy datasets to replicate the measurement of feature relevance in tree-based models, which helped with the bankruptcy prediction problem. They brought attention to the possibility of obtaining highly important features from other models that show better accuracy but do not have built-in feature measurement capabilities.

METHODOLOGY

Credit risk assessment

The primary metric to assess credit risk is the estimated probability of default (PD), or the likelihood that a corporation will not pay back its debts. The standard method for dealing with this issue is to forecast whether a company is in default by evaluating its credit score and establishing a threshold. Assume we have information on N firms' financial sheet metrics, which comprise T of the explanatory variables. A response variable Y indicates whether a company has defaulted or is still running, which is common in the following term for all companies. Companies are not in default if Y = 0 and in default if Y = 1. We have developed this credit scoring model to understand better the connection between the response variable and the T explanatory components.

There are primarily two credit rating models: black boxes and white boxes. In the former, we can see the final categorization, and we can't know how the explanatory variables relate to the response. Neural networks, random forests, and gradient boosting are examples of sophisticated ML models that fall into this category; these models have good predicted accuracy but low explainability. The opposite is true with white-box models, which are not opaque to the user and include examples such as logistic and linear regression. These elementary models detail their actions and the process of making predictions.

Logistic regression

Logistic regression is the 'white-box' statistical learning method most commonly used in credit scoring models.

The response variable is split into two groups, "default" and "active," according to the logistic regression model's classification of the variables. Here is a more formal way to specify the logistic regression:

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

The model intercept is represented by parameter α, the regression coefficient is denoted by γt, the likelihood of default for the nth firm is denoted by pn, and the T-dimensional vector of the borrower-specific explanatory factors is xn = (xn1, …, xnT).

Based on this, we can calculate the default probability as

$$p_n = exp(\alpha + \sum_{t=1}^{T} \beta_t x_{nt})(1 + exp(\alpha + \sum_{t=1}^{T} \beta_t x_{nt}))^{-1}$$

Logistic regression models are straightforward to grasp because of their linear functional form in the logarithm of odds, but this same linearity carries the risk of low prediction accuracy.

When dealing with complicated and huge datasets, logistic regression's predicted performance could fall short compared to a more sophisticated ML model.

Random forests

More sophisticated evaluations of credit risk rely on ML models. A random forest classifier, an ensemble of trees known as a random forest classifier, is one such tool; it identifies credit risk well.

A classification random forest model is similar to logistic regression, using a default response variable for each observation (T) and its associated vector of explanatory factors (xn-T) for each organization. Combining the rules derived from many classification trees, each trained on a different data set and augmented with explanatory variables makes up a random forest classifier. Although the building rules of each classification tree provide light on the process of generating various credit scores, the random-forest approach averages out all of the scores and makes them unintelligible. Because of its lack of transparency, a random forest model is unsuitable for use in the financial industry. One solution to this problem is to use explainable AI models, which explain how AI works by providing context and explanations.

Explainable Artificial Intelligence

Numerous rules are imposed on markets and financial institutions to keep customers and investors safe and the financial system stable. Credit risk models, in particular, are subject to oversight by financial regulators, who frequently seek confirmation of the key drivers. The idea that black-box AI isn't good for measuring credit risk prompted the creation of XAI models. As a model-agnostic post-processing tool for explaining and evaluating ML predictions, the Shapley values technique is extensively utilized as an explainable AI model.

The Shapley value method borrowed heavily from game theory and used a linear space to represent predictive judgments. We started with the idea that there should be a game where you can predict each row of observations.

Model predictors, or explanatory variables, were the players in each game. The expected value equals the total gain when all the projections are added together.

The Shapley value algorithm considers all potential coalitions (groups) of other variables to determine each variable's contribution to each prediction based on these assumptions.

Proposal

A stepwise variable selection technique based on the global Shapley values of each explanatory variable is proposed, which is applicable to both white box and black box models.

The procedure first creates a complete model with all variables. After that, we checked whether removing the variable with the lowest global explainability considerably reduced the predicted accuracy. If this happens, it will stop; otherwise, it will continue deleting variables until it reaches a similar point.

Our method relies heavily on a significance test that contrasts the AUC of two rival models that differ in the presence of a single variable. Calculated as the area beneath a model's Receiver Operating Curve (ROC), the Area Under the Curve (AUC) is the most often used statistic for predicting the correctness of binary variables. The receiver operating characteristic (ROC) curve was derived by combining the true and false positive rates at specific percentiles. While a perfect model would have a TPR of 1 and an FPR of 0, an AUC of 1 still indicates a solid model.

RESULTS AND STUDY

Data

We demonstrate how our solution can be applied to a massive dataset that includes the financial statements of more than 100,000 SMEs for the reporting year 2020. Modefinance (modefinance.com) provided the data. Regarding commercial financing for small and medium-sized firms (SMEs), Modefinance is the go-to credit rating agency. They're overseen by the European Securities and Markets Authority (ESMA). Given the pervasiveness of SMEs, These figures may represent a wider global phenomenon. Companies from the four biggest EU member states—Germany, France, Italy, and Spain—are included in the sample. In 2020, you may observe the distribution of these nations in Table 1.

Table 1. The Location of SMBs Selected for Sampling Around the World.

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 1 shows that Italy is home to most enterprises (49.37%), including several SMEs. Next on the list, with 27.93% of the companies' headquarters being in France, is Italy. At 1.30 percent, Germany has the lowest representation of any country in the table. The fact that public deposits are not required on firm balance sheets in Germany, despite the country's greater population, lends credence to this idea. Despite Germany's participation being minimal owing to the sample size, we prefer to keep all companies and not change the sample. Looking at the distribution of businesses in the sample by 'Industry Sector,' which shows the industry each SME is a part of, is an interesting exercise. The five most populous sectors are listed in Table 2.

Table 2. Distribution of small and medium-sized businesses in the sample by industry categories.

Industry Sector	No. of. Companies	Percentage
Retailing	30,201	24.91
Capital Goods	17,536	14.46
Material	11,969	9.87
Commercial and Professional Services	10,861	8.96
Food and Steples Retailing	8844	7.29

Take note that "Retailing" is the most populous industry, followed by "Capital Goods," "Materials," and "Commercial and Professional Services" (Table 2). A binary response variable that indicates whether a company is in trouble—possibly on the verge of default—and a collection of explanatory factors that can be viewed as probable or not as causes of this trouble are necessary to construct a credit score model using the provided data. The currently available information may be found in the 'MScore' variable, which represents the rating that each company was given by the rating agency Modefinance. An MScore can take on values that correspond to ratings like A, AA, AAA, B, BB, BBB, C, CC, CCC, or D, where A is the low10est credit risk (the possibility of default), and D is the highest credit risk (the probability of default). This is because MScore is based on multiple scoring systems.

Following the credit scoring context example given in Section 3.1, we allocated each company's rating to one of two potential classes to make the variable 'Mscore' a binary default variable.

We utilized the following criteria: C, CC, CCC, and D to determine Class 1 ratings, which indicated default, and Class 0 ratings, which indicated non-default. So, 14% of the sample of SMEs defaulted.
Figures 1 and 2 show the distributions of the sample default variable for each country and industry sector. In both representations, we show the default percentages that have been observed at the top of the bars, and the overall height of the bars is related to the number of businesses in each category (country or industry).

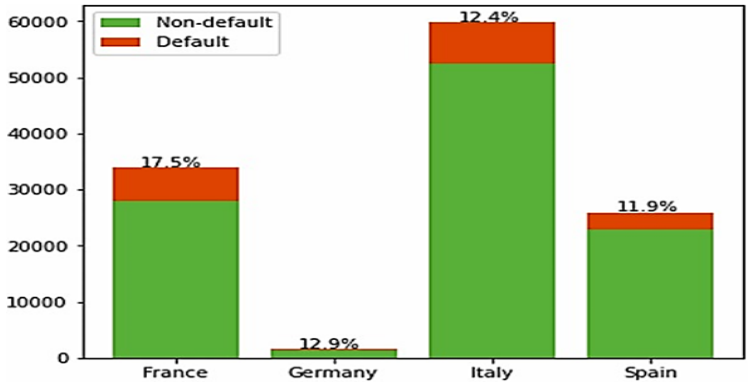

Fig. 1. *Comparison of defaulting and non-defaulting SMEs by nation.*

France is the riskiest country, with a default probability of around 17.5% (Fig. 1). Germany comes in second with a 12.9% default chance; however, compared to more populous countries like Italy and Spain, its frequency is modest. Therefore, its impact on the system is minimal. Fig. 2 also shows that the most dangerous businesses are "Consumer services," "Diversified financials," and "Media and entertainment." In contrast, there is a larger and more powerful commercial and professional services sector due to the sector's larger population.

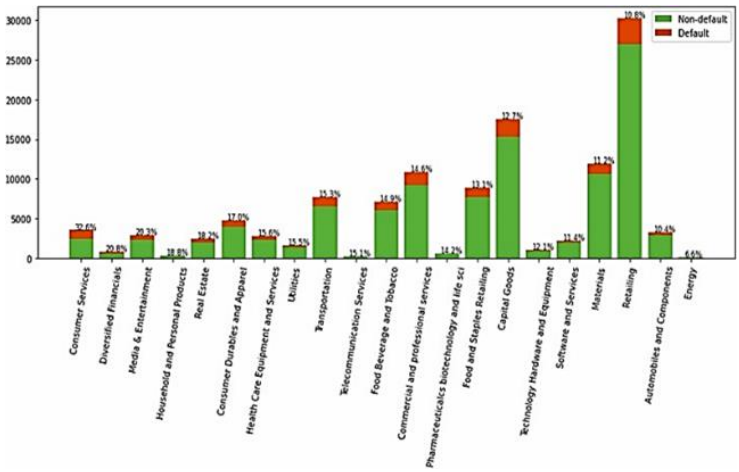

Fig. 2*. The business sector's distribution of defaulting and non-defaulting SMEs.*

In Table 3, you can see all of the financial ratios that mode-finance determined from the available 2020 company balance sheets. The variables in the sample data have been described to the fullest extent.

Table 3. Factors Used to Explain the Data.

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

Based on the 2020 financial statements, Table 3 shows that the six financial variables that can be used as explanatory variables are as follows: operational income (Turnover), operational scale (Total Assets), financial structure (Leverage), and profitability (EBIT, Profit, and Losses after Tax, Return on Equity).
Table 4 presents the summary statistics for each explanatory variable, broken down by company default status and non-default status.

Table 4. Financial variables' conditional means.

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

The most significant difference between defaulted and non-defaulted enterprises can be seen in the conditional means of EBIT, PLTax, and leverage (Table 4). These variables are expected to have the greatest impact on credit scores. On the other hand, there is a minor discrepancy between the conditional means of Turnover and Total Assets. The six financial ratios from Table 3 and the Country and Industry classifications were included as explanatory variables in the first full model test. To predict predicted coefficients, Z-scores, and p-values are presented in Table 5 to indicate a company's default using a logistic regression model.

Table 5. Credit scoring model estimated coefficients using complete logistic regression.

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

Given the vast amount of included training data (over 70,000 records), it is unsurprising that all variables are significant (Table 5). This results in a high goodness of fit. The greatest coefficients are for Country and Industry, but just because the scales of these variables are different doesn't imply they disproportionately influence the predictions. Utilizing the calculated model, we projected the scores of the organizations comprising the validation sample (30% of the total data). When we wanted to know how each variable impacted the predictions, we added the Shapley values from the test set (30% of the observations). We calculated the Global Shapley values for each variable. Figure 3 displays the results.

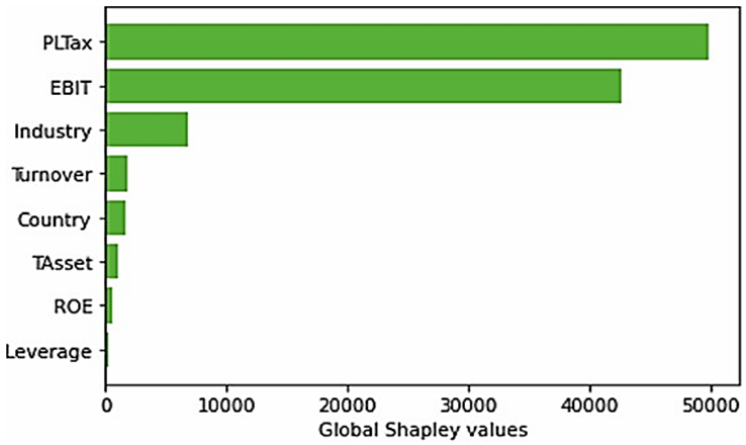

Fig. 3. *Worldwide Shapley values calculated from the logistic regression model's forecasts.*

Figure 3 reveals that "PLTax" has the highest Global Shapley value among the variables evaluated, making it the most influential in making predictions. "EBIT" comes in second. This finding partially agrees with Table 4's findings since PLTax and EBIT, two profitability factors, consistently show the largest difference in conditional means. At the same time, financial leverage is given modest importance.

We built an ML credit rating model with the random forest model presented as its core. Utilizing the identical data partitioning for logistic regression, we divided the data into training and validation samples, each comprising 70% of the observations. After training the training sample with Python's random forest GridSearch CV approach, we utilized the estimated model to determine the credit scores of the validation sample companies. After that, the model's predictions were compared to a predetermined threshold of 0.5, and each company in the validation sample was expected to either remain unchanged or experience a decline. Table 6 compares the logistic regression and the full random forest models, which use all six explanatory variables.

Table 6. Examining the entire Logistic Regression and Random Forest models with respect to prediction accuracy metrics.

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

Table 6 predicts that the random forest model will perform better than the others. Due to its improved compromise between sensitivity and specificity, the random forest model outperforms the others when considering the F1 score, sensitivity, and specificity. When employing AUC measurements with a defined threshold that is not equal to 0.5, the random forest model obtains an AUC of 0.93, whereas the logistic regression model produces an AUC of 0.63. These results are consistent with one another. The results demonstrate that the random forest credit scoring model outperforms the other model regarding prediction accuracy. This issue can be resolved by incorporating a "feature importance plot" into the random forest model's post-processing steps for scoring predictions.

With the random forest training data, the graph displays the average reduction in the Gini variability measure for all model variables as a function of each explanatory variable and tree splits. An explanatory variable's significance is proportional to its variability reduction for a given split; a smaller value indicates more significance. You can see the feature importance plot for the training data we used in Fig. 4.

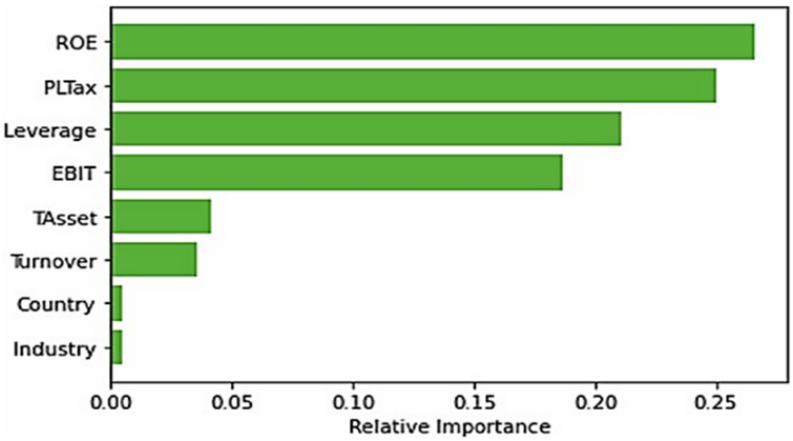

Fig. 4. *The Significance of Regular Forest Features.*

While the feature significance plot does help explain things, it isn't model-agnostic and can only be created for logistic regression models or random forests. This results in the inability to compare the explainability of various models. They employ Shapley values, computed using the expected credit ratings in their validation sample. Without the model, these numbers have no meaning. As can be seen from the global Shapley values, Figure 5 displays the total impact of every variable.

These values are the sum of all observations for each variable.

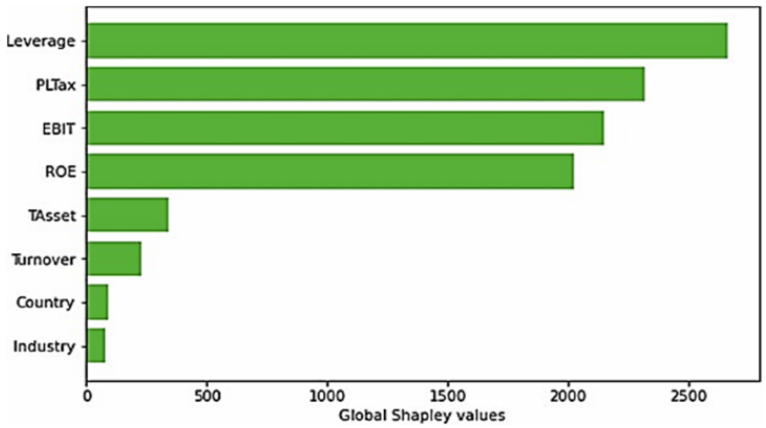

Fig. 5. *Shapley International prioritizes.*

The feature importance plot in Figure 4, the data gained from applying Shapley values to logistic regression, and the difference in conditional means all point to 'Leverage' being the most explainable variable. 'PLTax' and 'EBIT' follow closely behind (see Fig. 5). As an auxiliary variable, "Leverage" is used here. Figure 5 demonstrates that 'Country' and 'Industry' have relatively small global Shapley values. This aligns with the feature importance plot but contradicts the logistic regression results shown in Figure 3. While the size and operating revenues (as assessed by Tasset and Turnover) impact small and medium-sized firms (SMEs), the Shapley values of the random forest credit scores indicate that financial leverage and likelihood have a greater impact. This goes against the grain of what a logistic regression model would produce. Figure 5 shows the Global Shapley values, which we used to compare the AUC and implement our suggested selection process, which involves a step-by-step selection of components. Through this process, we determined which factors statistically explain the default likelihood. Classical stepwise approaches compare models based on their likelihood, but our procedure differed. Rather, we evaluated the models according to how well they could forecast the future. The benefit of this approach is its generalizability; for example, we may compare models using logistic regression and any ML techniques, including random forest models, with an underlying probabilistic model.

CONCLUSIONS

Credit scoring models can be improved with ensemble ML models like random forests, but these models must be explained. After processing credit ratings is complete, they can be made more transparent using explainable AI methods such as Shapley values. This study utilized Shapley values as a framework for variable selection and explainability to balance prediction accuracy and the degree to which the model could be explained. We get a model that is as simple as possible. To do this, we propose a model selection method that ranks possible explanatory variables by their predictive importance using global Shapley values. We then use a backward stepwise selection method to find the best factors statistically by comparing their predicted accuracy. We tested our hypothesis using a dataset that includes the credit ratings of European SMEs, their industry and place of origin, and the values of six financial ratios extracted from their 2020 financial statements. These results are corroborated by the fact that the nonlinear random forest model produced more precise credit scores than logistic regression. According to our methodology, the random forest model lacked a bias toward financial inclusion and was sparser than the logistic regression model. It was also country and industrial sector independent, relying solely on balance sheet ratios.

Our suggested technique satisfies a need in the literature by providing a standard procedure for comparing models according to their explainability and accuracy, and it also generates a credit-scoring model that effectively combines the two. From a purely administrative perspective, our method can prove useful to financial institutions, regulators, and fintech startups in their pursuit of legally compliant ML models for credit assessment, especially those pertaining to artificial intelligence.

REFERENCES

[1]. 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

[2]. 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

[3]. Dushimimana, B., Wambui, Y., Lubega, T., & McSharry, P. E. (2020). Use 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

[4]. 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
[5]. 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
[6]. 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
[7]. 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
[8]. Giudici, P., & Raffinetti, E. (2021). Shapley lorenz explainable artificial intelligence. Expert Systems with Applications, 114104, 167.
[9]. 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
[10]. 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
[11]. 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
[12]. 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
[13]. 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
[14]. Sundararajan, M. and Najmi, A. (2020). The many Shapley values for model explanation.In International conference on Machine Learning(9269–9278).: PMLR.
[15]. Ahelegbey D., Giudici P., Hadji-Misheva B. (2019). Latent factor models for credit scoring in P2P systems. Phys. A Stat. Mech. Appl. 522, 112–121. 10.1016/j.physa.2019.01.130 [CrossRef] [Google Scholar]
[16]. Chen T., Guestrin C. (2016). Xgboost: a scalable tree boosting system, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (New York, NY: ACM;), 785–794. [Google Scholar]
[17]. Croxson K., Bracke P., Jung C. (2019). Explaining Why the Computer Says 'no'. London, UK: FCA-Insight. [Google Scholar]
[18]. EU (2016). Regulation (EU) 2016/679–general data protection regulation (GDPR). Off. J. Eur. Union. [Google Scholar]
[19]. FSB (2017). Artificial Intelligence and Machine Learning in Financial Services–Market Developments and Financial Stability Implication. Technical report, Financial Stability Board. [Google Scholar]
[20]. Giudici P. (2018). Financial data science, in Statistics and Probability Letters, Vol. 136 (Elsevier:). 160–164. [Google Scholar]
[21]. Giudici P., Hadji Misheva B., Spelta A. (2019a). Correlation network models to improve P2P credit risk management. Artif. Intell. Finance. [PMC free article] [PubMed] [Google Scholar]
[22]. Giudici P., Hadji-Misheva B., Spelta A. (2019b). Network based credit risk models. Qual. Eng. 32, 199–211. 10.1080/08982112.2019.1655159 [PMC free article] [PubMed] [CrossRef] [Google Scholar]
[23]. Joseph A. (2019). Shapley Regressions: A Framework for Statistical Inference on Machine Learning Models. Resreport 784, Bank of England. [Google Scholar]
[24]. Lundberg S., Lee S.-I. (2017). A unified approach to interpreting model predictions, in Advances in Neural Information Processing Systems 30, eds Guyon I., Luxburg U. V., Bengio S., Wallach H., Fergus R., Vishwanathan S., et al. (New York, NY: Curran Associates, Inc;), 4765–4774. [Google Scholar]
[25]. Murdoch W. J., Singh C., Kumbier K., Abbasi-Asl R., Yu B. (2019). Definitions, methods, and applications in interpretable machine learning. Proc. Natl. Acad. Sci. U.S.A. 116, 22071–22080. 10.1073/pnas.1900654116 [PMC free article] [PubMed] [CrossRef] [Google Scholar
[26]. Saurabh Kumar, "Difference-in-Differences in Action: Measuring Brand Marketing Campaign Impact Through Survey Responses", International Journal of Science and Research (IJSR), Volume 7 Issue 9, September 2018, pp. 1669-1673, https://www.ijsr.net/getabstract.php?paperid=SR18920585948