Received: 14 April 2022 | Revised: 6 July 2023 | Accepted: 10 July 2023

DOI: 10.1002/asmb.2803

RESEARCH ARTICLE

WILEY

Lost in a black-box? Interpretable machine learning for assessing Italian SMEs default

Lisa Crosato[1] | **Caterina Liberati**[2] | **Marco Repetto**[2,3]

[1]Department of Economics and Bliss–Digital Impact Lab, Ca' Foscari University of Venice, Venice, Italy

[2]Department of Economics, Management and Statistics (DEMS) and Center for European Studies (CefES-DEMS), University of Milano-Bicocca, Milan, Lombardy, Italy

[3]Digital Industries, Siemens Italy, Milan, Italy

Correspondence
Caterina Liberati, Department of Economics, Management and Statistics (DEMS) and Center for European Studies (CefES-DEMS), University of Milano-Bicocca, Milano, Piazza dell'Ateneo Nuovo, 1, 20126 Milan, Italy.
Email: caterina.liberati@unimib.it

Abstract

Academic research and the financial industry have recently shown great interest in Machine Learning algorithms capable of solving complex learning tasks, although in the field of firms' default prediction the lack of interpretability has prevented an extensive adoption of the black-box type of models. In order to overcome this drawback and maintain the high performances of black-boxes, this paper has chosen a model-agnostic approach. Accumulated Local Effects and Shapley values are used to shape the predictors' impact on the likelihood of default and rank them according to their contribution to the model outcome. Prediction is achieved by two Machine Learning algorithms (eXtreme Gradient Boosting and FeedForward Neural Networks) compared with three standard discriminant models. Results show that our analysis of the Italian Small and Medium Enterprises manufacturing industry benefits from the overall highest classification power by the eXtreme Gradient Boosting algorithm still maintaining a rich interpretation framework to support decisions.

KEYWORDS

accumulated local effects, default prediction, interpretability, machine learning, small and medium sized enterprises

1 | INTRODUCTION

The European Union (EU) economy is deeply grounded in Small and Medium Enterprises (SMEs) which represent about 99.8% of the active enterprises in the EU-28 non-financial business sector (NFBS), accounting for almost 60% of value-added within the NFBS and fostering the EU workforce with two out of three jobs.[1]

Consequently, a wide literature has grown covering various economic aspects of SMEs, mainly focused on default prediction (for an up-to-date review see Reference 2), interesting for scholars as well as for practitioners such as financial intermediaries and for policy makers in their effort to support SMEs and to ease credit constraints to which they are naturally exposed.[3]

Whether for private credit-risk assessment or for public funding, independently of the type of data imputed to measure a firm health status, prediction of default should succeed in two aspects: maximise correct classification and clarify the role of the variables involved in the process. Most of the times, the contributions based on Machine Learning (ML) techniques neglect the latter aspect, often with better results with respect to standard parametric techniques that provide, on the

contrary, a clear framework for interpretation. In other words ML techniques rarely deal with *interpretability* which, according to a recent document released by the European Commission, should be kept "in mind from the start".[4]

Interpretability is central when applying a model in practice, both in terms of managerial decisions and compliance: it is a fundamental requisite to bring a model into production.[5] Interpretable models allow risk managers and decision makers to understand their outcome and to knowingly take courses of actions. The European Commission itself encourages organizations to build trustworthy Artificial Intelligence (AI) systems (including ML techniques) around several pillars: one of them is transparency, which encompasses traceability, explainability and open communication about the limitations of the AI system.[6]

Accordingly, ML models-no matter how good in classifying default-should be made readable to avoid that their inherent uninterpretable nature may prevent their spreading in the literature on firms' default prediction as well as their use in other contexts regulated by transparency norms.

This work tries to fill this gap by applying two different kind of ML models, FeedForward Neural Networks[7] and eXtreme Gradient Boosting,[8] to Italian Manufacturing SMEs' default prediction, with a special attention to interpretability. Italy represents an ideal testing ground for SMEs default prediction since its economic framework is more extensively configured by firms up to this size than the average of EU countries.[1] Default was assessed on the basis of the firms' accounting information retrieved from Orbis, a Bureau van Dijk (BvD) dataset.

The main original contribution of the paper is to address ML models' interpretability in the context of default prediction. Our approach is based on model agnostic-techniques and adds Accumulated Local Effects (ALEs),[9] to the Shapley values already applied in Reference 10. Using these techniques we can rank the variables in terms of their contribution to the classification and determine their impact on default prediction. Robustness of the ML models hyperparameters was taken care of by Montecarlo Cross-Validation and substantial class imbalance between defaulted and survived firms was reduced through undersampling of the latter into the cross-validation training sets. Another contribution of the paper is the benchmarking of the ML models' outcome with Logistic, Probit and with Binary Generalized Extreme Value Additive (BGEVA) classifications, both according to standard performance metrics and to the role played by the input features. Moving a step forward with respect to the current use of ALEs, we fully exploit the tool and supply them also for the parametric models, in order to unfold what is compressed within the single variables coefficients and significance and guarantee a common ground for comparison.

We obtain a few interesting results. First, eXtreme Gradient Boosting (XGBoost) outperformed the other models mainly for total classification accuracy and default prediction rate. Second, the impact of the variables assessed by XGBoost is fully consistent with the economic literature, whereas the same cannot be said for its competitors. Thanks to the ALEs framework for interpretability, risky thresholds, non-linear patterns and other additional insights emerge for predictors even in standard models.

The remainder of the paper is organized as follows. Section 2 gives an overview of the (necessarily) recent literature concerning ML intepretability. Section 3 provides a description of the dataset and of the features we use throughout the modelling. Section 4 discusses our methodology, briefly reviewing the models fundamentals, the techniques employed for interpretability and the research design. Section 5 presents the results and discusses the most relevant findings. Section 6 concludes.

2 | LITERATURE REVIEW

The ability to predict corporate failure has been largely investigated in credit risk literature. On the one hand, the academic interest in the topic has increased after the global financial crisis (2007–2009) and is being renewed today due to the current pandemic impact on the companies of all sizes.[2,11] On the other hand, a good part of the financial industry has shown great attention to statistical algorithms that prioritize the pursuit of predictive power. Such a trend has been registered by recent surveys, showing that credit institutions are gradually embracing ML techniques in different areas of credit risk management, credit scoring and monitoring.[12-14] Among all, the biggest annual growth in the adoption of highly performing algorithms has been observed in the SMEs sector.[15]

For these reasons, new modeling techniques have been successfully employed in predicting SMEs default, including Deep Learning,[16] Support Vector Machines,[17,18] Neural Networks,[19] and Hazards models,[20,21] to name only a few. However, they have been applied mainly in order to improve classification accuracy with respect to the standard linear models, supporting decisions through reduced uncertainty but leaving somewhat unsolved the issue of interpretability. But the latter is no longer a negligible aspect, both for academic research and for management of regulated financial services:

it has become overriding, since the European Commission and other European Institutions have released a number of regulatory standards on Machine Learning modeling.

The Ethics Guidelines for Trustworthy AI[4] and the Report on Big Data and Advanced Analytics[22] illustrate the principle of explicability of ML algorithms which must be transparent and fully interpretable to the ones directly and indirectly affected. Indeed, as the Commission points out, predictions, even accurate, without explainability measures are unable to foster responsible and sustainable AI innovation. The pillar of transparency (fourth among seven), somewhat combines explainability and interpretability of a model, referring to interpretability as the "concept of comprehensibility, explainability, or understandability".[6]

The difference in meaning between interpretability and explainability, synonymous in the dictionary, has been addressed by the recent ML literature which recognizes the two words a conceptual distinction related to different properties of the model and knowledge aspects.[23,24] A clear indication about the distinction is given by Reference 25 that defines interpretation as a mapping of an abstract concept into a domain that the human expert can perceive and comprehend and explanation as a collection of features of the interpretable domain that have contributed to produce a decision. Roughly speaking, interpretability is defined as the ability to spell out or to provide the meaning in understandable terms to a human,[26,27] whereas explainability is identified as the capacity of revealing the causes underlying the decision driven by a ML method.[28]

There are several approaches to ML interpretability in literature, classified in two main categories: ante-hoc and post-hoc methods. Ante-hoc methods employ intrinsically interpretable models (e.g., simple decision trees or regression models, also called white-box) characterized by a basic structure. They rely on model-specific interpretations depending on examination of internal model parameters. Post-hoc methods instead provide a reconstructed interpretation of decision rules produced by a black-box model in a reverse engineering logic,[29,30] reckoning on model-agnostic interpretation where internal model parameters are not inspected.

So far, ante-hoc approaches were widely used in the SMEs default prediction literature that counts contributions employing mainly white-box models as Logistic regression (see e.g., References 31-33), Survival analysis,[34-36] or Generalised Extreme Value regression.[37] The empirical evidences and the variables' effect on the outcome are interpreted in an inferential testing setting, so that the impact of the predictors and the results' implications are always clear to the reader.

On the contrary, post-hoc methods have been rarely used in this field and comprehend Partial Dependence (PD) plots,[38] Local Interpretable Model-agnostic Explanations (LIME)[39] and the SHAP,[40] all of them providing detailed model-agnostic interpretation of the complex ML algorithms employed either focusing on a global or a local scale[41-43] used the PD to identify the relevant variables' subset and to measure the change of the average probability of default with respect to the single features. A PD-based framework for making transparent, auditable, and explainable black-box models both at the global level and for single instances was developed in the ambit of credit scoring by Reference 44. LIME and SHAP were applied in References 45,46 to rank the variables and to provide their impact on the output prediction respectively.

Alternative strategies to enhance interpretability combine the above approaches to get the most out of both. Surrogate models emulate the black-boxes with one or more white-boxes to clarify the output of the former.[47,48] Another strand of literature links together complex ML models for feature selection/transformation and white-box models for fitting/interpretation in two-layer frameworks.[49,50] The rationale under these combinations is to exploit each class of models in what they do better: black-boxes for coping with high-dimensionality and non-linearities and white-boxes for plain explanations, treating all issues within and between data ex-ante and leaving thus space for simpler models ex-post. This approach seems promising, although evidences on its advantages have been so far limited.

This paper contributes to the literature investigating global level interpretability and to the literature on SMEs default: we compare black-box with white-box models on both performance and interpretability domains, thus bridging both sides of the empirical work in the field. We do this by fully exploiting post-hoc methods on all models. Building on a set of features recommended by experts from the well-established literature on firm default, we employ the Accumulated Local Effects (ALEs),[9] a model-agnostic technique that represents a suitable alternative to PDs when the features are highly correlated, without providing incoherent values.[49] Since ALEs are a newest approach, their usage is still limited and not yet spread in the bankruptcy prediction area.

3 | DATA DESCRIPTION

The data of this study are retrieved from BvD-Orbis database, which provides financial and accounting ratios from balance sheets of the European private companies. We have restricted our focus on Italian manufacturing SMEs for

several reasons. Italy is the second-largest manufacturing country in the EU[51] and this sector generates more than 30% of the Italian GDP.[52] Differently from SMEs in other EU countries, Italian SMEs trade substantially more than large firms, the manufacturing sector, in particular, driving both imports and exports. Moreover, according to Reference 19, predictive models have better performances when trained for a specific sector in that pooling heterogeneous firms is avoided.

To define our sample, we filtered the database both by country and NACE codes (from 10 to 33) and we employed the European Commission definition[53] of Small and Medium Enterprises. We retrieved only firms with an annual turnover of fewer than 50 million euros, a number of employees lower than 250 and a total balance sheet of fewer than 43 million euros. Among those, we classified as defaults all the enterprises that entered bankruptcy or a liquidation procedure, as well as active companies that had not repaid debt (default of payment), active companies in administration or receivership or under a scheme of the arrangement, (insolvency proceedings), which in Orbis are also considered in default. Consistently with the literature, we excluded dissolved firms that no longer exist as a legal entity when the reason for dissolution is not specified.[54,55] This category in fact encompasses firms that may not necessarily experience financial difficulties. The resulting dataset contains 105,058 SMEs with a proportion of 1.72% (1807) failed companies.

The accounting indicators, which refer to 2016 to predict the firm status in 2017, have been selected among the most frequently used in the SMEs default literature and are the following:[19,31,37,54-58]

1. Cash flow: computed as net income plus depreciations
2. Gearing ratio: computed as the ratio between total debt and total assets
3. Number of employees, as a size measure from an input perspective
4. Profit margin: measured as profit/loss before tax over the operating revenue
5. ROCE: computed as profit/loss before tax over capital employed, which is given as total assets minus current liabilities
6. ROE: computed as profit/loss before tax over shareholders' funds
7. Sales: in thousands Euro, measuring the output side of firm size
8. Solvency ratio: computed as shareholders' funds over total assets
9. Total assets: in thousands Euro, as a measure of total firm resources

As a quick preview of the expected relationship between the single predictors and the likelihood of default, we have computed the averages and standard deviations of the variables for survived and defaulted firms (see Table 1). In line with Reference 55, we can see on average weakest liquidity, smallest size and deficient leverage for defaulted firms.

The Profit margin is higher for surviving firms, whereas the remaining profitability indexes, ROE and ROCE, show a larger median and mean among defaulted firms respectively. They should both be negatively related to default, although some studies found ROCE's impact non-significant coherently with the low-equity dependency of small businesses,[59] while others attest its positive effect on default with a caveat for large values.[37] We will get more valuable insights into these profitability indicators when discussing the models' outcome.

4 | METHODOLOGY

4.1 | White-box versus black-box models

The models we apply can be broadly classified as white-box, or interpretable, and black-box but post-hoc interpretable in the model-agnostic framework.

In the first category, Logistic Regression (LR) and Probit were selected among the most recurrent models in the economics literature, where the accent on the factors impacting default is certainly of primary importance. These models frequently serve as a benchmark for classification when a new method is proposed. The third model, BGEVA,[37] comes from the Operational Research literature and is based on the quantile function of a Generalized Extreme Value random variable. The main strengths of BGEVA are robustness, accounting for non-linearities and preserving interpretability.

The black-box models we use are XGBoost and FeedForward Neural Networks (FANN). These models are by nature uninterpretable since the explanatory variables pass multiple trees (XGBoost) or layers (FANN), thus generating an output for which an understandable explanation cannot be provided.

TABLE 1 Summary statistics by survived and failed firms.

Variable	Min	Mean	St. dev.	Median	Max
			Survived		
Cash flow	−43,142.000	236.802	934.877	55.000	89,591.000
Gearing ratio	0.000	24.807	23.093	22.198	99.882
No. of employees	1.000	16.506	24.385	9.000	249.000
Profit margin	−87,700.000	−2.736	610.488	2.673	141,300.000
ROCE	−86,250.000	12.335	516.765	7.955	114,233.333
ROE	−35,961.110	23.020	314.135	17.647	39,500.000
Sales	1.000	3,427.163	6,301.229	1,165.000	49,995.000
Solvency ratio	−99.970	27.101	24.315	22.400	100.000
Total assets	1.000	3,904.129	12,098.087	1,194.000	1,758,577.000
			Failed		
Cash flow	−19,497.00	−278.521	1,636.028	−15.000	41,186.000
Gearing ratio	0.000	22.166	26.010	12.594	98.134
No. of employees	1.000	11.080	19.531	5.000	228.000
Profit margin	−87,762.50	−106.845	2,190.012	−9.677	21,700.000
ROCE	−23,600.00	66.367	2,284.001	5.818	90,800.000
ROE	−28,800.00	7.146	971.112	32.692	5,366.667
Sales	1.000	1,259.695	2,940.010	380.000	32,522.000
Solvency ratio	−99.430	−1.044	37.342	3.080	100.000
Total assets	1.000	1,921.689	5,149.559	526.000	110,501.000

The XGBoost algorithm was found to provide the best performance in default prediction with respect to LR, Linear Discriminant Analysis, and Artificial Neural Networks.[10,60] The algorithm builds a sequence of shallow decision trees, which are trees with few leaves. Considering a single tree one would get an interpretable model taking the following functional form:

$$f(x) = \sum_{m=1}^{M} \theta_m I(x \in R_m) \tag{1}$$

where M covers the whole input space with $R_1, \ldots R_M$ non-overlapping partitions, $I(\cdot)$ is the indicator function, and θ_m is the coefficient associated with partition R_m. In this layout, each subsequent tree learns from the previous one, thus improving the prediction.[38]

As a competing black-box model we chose the FANN, which is widely used and well performing in SMEs' default prediction[2] and in several works on retail credit risk modeling.[61-63] FANN consists of a direct acyclic network of nodes organized in densely connected layers, where inputs, weighted and shifted by a bias term, are fed into the node's activation function and influence each subsequent layer until the final output layer. In a binary classification task, the output of a single layer FANN can be described as in Reference 64 by:

$$f(x) = \phi\left(\beta_0 + \sum_{j=1}^{d} \beta_j G\left(\gamma_{j0} + \sum_{i=1}^{p} \gamma_{ji} x_i\right)\right) \tag{2}$$

where G is the activation function, in our case $G(x) = \frac{1}{1+e^{-\alpha x}}$, β and γ represent weights and biases at each layer, whereas $\phi(\cdot)$ is the network output function that in our case is also a sigmoid function as for $G(\cdot)$.

4.2 | Model-agnostic interpretability

To achieve the goal of interpretability, we make use of two different and complementary model-agnostic techniques. First, we use the global Shapley Values[65] to provide comparable information on the single feature contributions to the model output. Global Shapley Values have been already proposed in the SMEs default prediction literature by Reference 10. They differ from standard feature importance metrics based on feature permutation because of feature attribution evaluation based on possible coalitions capturing feature interactions.[66] Although model-agnostic, they share some of the axioms that characterize gradient-based interpretability methods such as Integrated Gradients.[67]

However, global Shapley Values do not provide any information about the shape of the variable effects, therefore we resort to ALEs.[9] ALEs, contrary to Shapley Values, offer a visualization of the path according to which the single variables impact on the estimated probability of default.

To further clarify the improvement that ALEs bring to interpretability in our setting, we briefly contextualize the method and outline its fundamentals.

The first model-agnostic approach for ML models' interpretation to appear in the literature was Partial Dependence (PD), proposed by Reference 68 in the early 1990s. PD plots evaluate the change in the average predicted value as specified features vary over their marginal distribution.[69] In other words, they measure the dependence of the outcome on a single feature when all of the others are marginalized out. Since their first formulation, PD plots have been used extensively in many fields but seldom in the credit risk literature, with a recent application by Reference 70.

One of the main criticisms moved to PD concerns its managing the relationships within features. The PD evaluation on all the possible feature configurations carries the risk of computing points outside the data envelope: such points, intrinsically artificial, can result in a misleading effect of some features when working on real datasets.

Due to this fallacy, and because of the renewed interest in complex deep learning models as Artificial Neural Networks, many new methodologies have been proposed. With Average Marginal Effects (AMEs),[71] suggested to condition the PD to specified values of the data envelope.[39] went the opposite direction presenting a local approximation of the model through simpler linear models, the so-called Local Interpretable Model-agnostic Explanations (LIME). In subsequent research, they also worked on rule-based local explanations of complex black-box models.[72] Shapley Additive exPlanations (SHAP) was introduced by Reference 40 to provide a human understandable and local Shapley evaluation.

In this framework, ALEs constitute a further refinement of both PD and AMEs. They avoid the PD plots-drawback of assessing variables' effects outside the data envelope, generally occurring when features are highly correlated,[9] as in the case of many accounting indicators.[33,54] Furthermore, ALEs do not simply condition on specified values of the data envelope as AMEs do, but take first-order differences conditional on the feature space partitioning, eventually eliminating possible bias derived from features' relationships.

Specifically, computing the ALE implies the evaluation of the following type of function:

$$ALE_{\hat{f},S}(x) = \int_{z_{0,S}}^{x} \left[\int \frac{\partial \hat{f}(z_S, X_{\setminus S})}{\partial z_S} d\mathcal{P}(X_{\setminus S} | z_S) \right] dz_S - constant \tag{3}$$

where $\hat{f}$ is the black-box model, S is the subset of variables' index, X is the matrix containing all the variables, x is the vector containing the feature values and z identifies the boundaries of the K partitions, such that $z_{0,S} = min(x_S)$.

The expression in Equation (3) is in principle not model-agnostic as it requires accessing the gradient of the model: $\nabla_{z_S}\hat{f} = \frac{\partial \hat{f}(z_S, X_{\setminus S})}{\partial z_S}$ but this is not known or even non existent in certain black-boxes. As a replacement, finite differences are taken to the boundaries of the partitions, z_{k-1} and z_k.

Hence, the resulting formula to evaluate ALEs is:

$$ALE_{\hat{f},S}(x_S) = \sum_{k=1}^{k_S(x)} \frac{1}{n_S(k)} \sum_{i: x_S^{(i)} \in N_S(k)} \left[\hat{f}(z_{k,j}, x_{\setminus S}^{(i)}) - \hat{f}(z_{k-1,j}, x_{\setminus S}^{(i)}) \right] - \frac{1}{n} \sum_{i=1}^{n} ALE_{\hat{f},S}(x_S^{(i)}) \tag{4}$$

The replacement of the constant term in Equation (3) by $-\frac{1}{n}\sum_{i=1}^{n} ALE_{\hat{f},S}(x_S^{(i)})$ in Equation (4) centers the plot, which is something missing in PD. This makes it clear that, by evaluating predictions' finite differences conditional on S and integrating the derivative over features S, ALEs disentangle the interaction between covariates. This way the main disadvantage of PD is solved.

TABLE 2 Models' performances on the test set.

Model	Sensitivity	Specificity	E-I	E-II	H	AUC	BS	KS
FANN	0.694	**0.829**	**0.171**	0.306	**0.391**	0.827	0.187	0.501
XGBoost	**0.821**	0.719	0.281	**0.179**	0.383	**0.843**	**0.146**	**0.552**
BGEVA	0.752	0.727	0.273	0.248	0.331	0.819	0.178	0.481
LR	0.745	0.736	0.264	0.246	0.303	0.809	0.151	0.483
Probit	0.738	0.737	0.263	0.262	0.299	0.809	0.190	0.448

4.3 | Research design

Our research design has been carried out according to Reference 73. We split the initial dataset into training (70%) and test (30%) sets*. Then, through the Monte Carlo Cross-Validation procedure,[74] we estimate the models parameters and validate the estimated rules. More in detail, at each iteration we create a sub-training set and a validation set via random sampling without replacement so that the models learn from the training set whereas the assessment, based on performance metrics, is done on the validation set. This way, we also tune the hyperparameters of the algorithms when necessary.

The training set serves as well to compute the Shapley values, based on the optimal rule, and to calculate the ALEs with corresponding bootstrap non-parametric confidence intervals.[9,75] Finally, we evaluated the models' performance on the test set.

We took into account also the severe unbalance in favour of survived firms to avoid over-classification of the majority class.[76] After testing several techniques for addressing imbalance[77] in the learning phase, we have chosen random Undersampling, which consists of sampling randomly among the majority class observations to achieve balancing†.

Obviously the undersampling scheme was applied only to the training data, to avoid over-optimistic performance metrics on either the validation or the test set.[78,79]

5 | RESULTS

The results are organized according to the performance and interpretation of the five models. The performance is measured by the proportion of failed and survived firms correctly identified (sensitivity and specificity) together with the Errors of the first and second type (E-I and E-II, respectively) as well as by four global performance metrics: the Area Under the Receiver Operating Curve (AUC), the H-measure, the Brier Score (BS) and the Kolmogorov-Smirnov statistic (KS) (see Table 2). We have chosen these indicators‡ for two reasons: they are popular in credit scoring and account for three different aspects of the discriminating rule. The AUC and H-measure assess discriminatory ability, with the H-measure normalizing classifiers' cost distribution, the BS evaluates the accuracy of probability predictions and KS appraises the correctness of categorical predictions.[73]

Second, we cross-compare the role and weight of the variables among models and contextualize the results within the literature. The post-hoc interpretation of the black-box models is based on the Shapley values and ALEs. We report the ALEs also for interpretable models to exploit a common basis for predictors comparison without incurring in the "p-value arbitrage" when evaluating white-box models via p-values and ML models via other criteria.[80]

5.1 | Performance

All competing models offer fair correct classification rates, but the ones that score globally best are black-box models, in terms of all metrics. The FANN reaches the highest H-measure and specificity while it's last as far as correct classification

*Results based on alternative data splits are reported in Appendix A.

†Complete results about the resampling schemes are reported in Appendix B.

‡The F1-score is not included in our analysis since in our case it would be confounding. The score, defined as the harmonic mean of precision and sensitivity, is largely driven by the imbalance between the two groups in the test set. Results on the F1 score are available upon request.

TABLE 3 Estimates and summary statistics for the probit, logistic regression, and BGEVA models on the test set (significant variables in bold).

	Probit model			Logistic regression			BGEVA model		
	Odds ratio	**Std. error**	**p-value**	**Odds ratio**	**Std error**	**p-value**	**Estimate**	**Std. error**	**p-value**
(Intercept)	6.195	0.134	0.000	21.256	0.233	0.000	2.087	0.137	0.000
Cash flow	**1.000**	0.000	**0.000**	**0.999**	0.000	**0.000**	**−0.001**	0.000	**0.000**
Gearing ratio	1.000	0.001	0.713	1.001	0.002	0.594	0.000	0.001	0.756
Number of employees	1.081	0.078	0.319	1.135	0.131	0.332	0.033	0.081	0.683
Profit margin	1.000	0.000	0.535	1.000	0.000	0.947	0.000	0.000	0.800
ROCE	1.000	0.000	0.256	1.000	0.000	0.302	0.000	0.000	0.027
ROE	1.000	0.000	0.240	1.000	0.000	0.275	0.000	0.000	0.285
Sales	**0.526**	0.066	**0.000**	**0.316**	0.120	**0.000**	**−0.637**	0.064	**0.000**
Solvency ratio	**0.985**	0.001	**0.000**	**0.973**	0.002	**0.000**	**−0.015**	0.001	**0.000**
Total assets	1.044	0.064	0.503	1.166	0.112	0.172	0.090	0.066	0.174

of default is concerned (with a sensitivity not reaching 70%, see Table 2). On the contrary, the XGBoost algorithm provides the best default prediction (showing, by far, the largest sensitivity) with a reasonable classification of survivors, as well as the highest global metrics but for the H-measure, which ranks it second.

The interpretable models are ranked consistently by AUC and H-measure in the following order: BGEVA, LR and Probit, whereas the Brier score and the KS provide alternative rankings. Anyway, these results confirm the trade-off between performance and interpretability highlighted in previous works on Italian SMEs.[19]

All in all, undersampling the training set has a balancing effect on the rate of correct prediction for either class.[77] This improves global classification not only through FANN, but also when applying Logistic Regression, as compared for instance with the results on the same kind of variables of References 19 or 32, the latter for both techniques.

5.2 | Interpretation

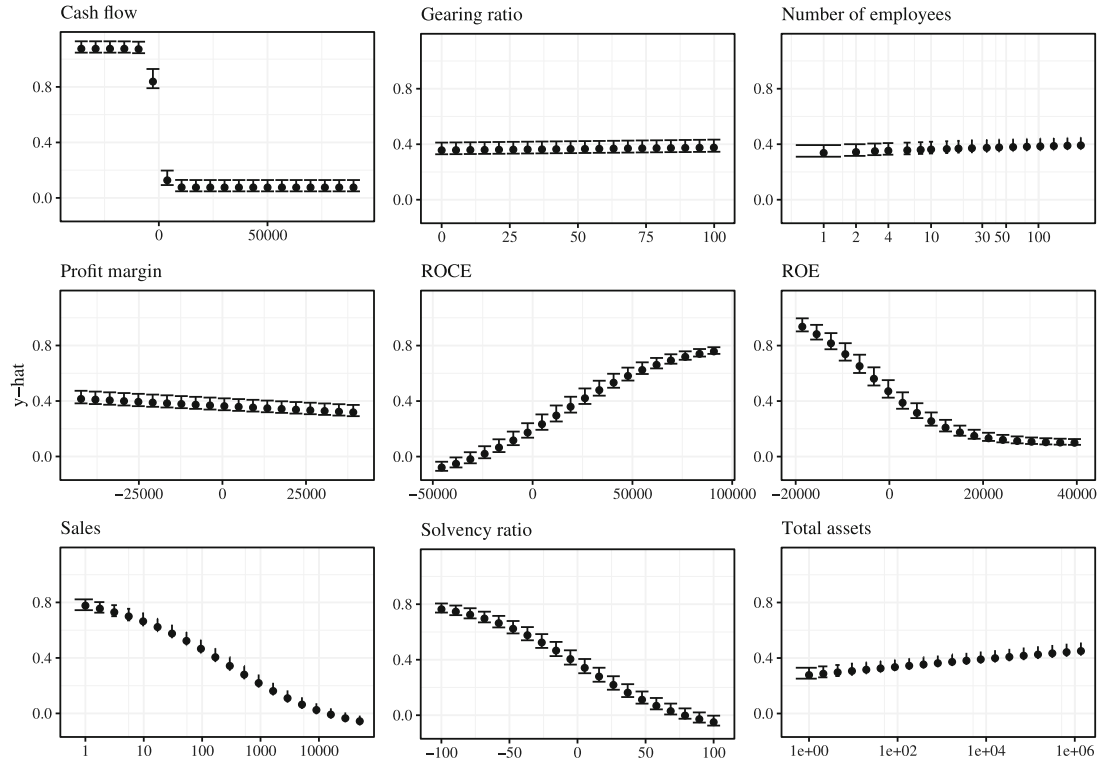
Most of the variables have non-significant effects on the probabilities of default estimated by white-box models, as long as these effects are ascertained by p-values (Table 3). Three variables display a significant and non-null coefficient, no matter the model: Sales[§], the Solvency Ratio and the Cash flow, all with an adverse effect on the probability of default.

The negative impact exerted by Sales on default, recurrent in many works, is not surprising since Sales is one of the main proxies of a company's size and largest firms tend to overcome demand shocks better than smaller firms,[33,82] which is also consistent with the means reported in Table 1 for the two groups of firms. Apparently, the size effect is captured exclusively by the output-side variable since the other size proxies, the Number of employees and the Total Assets, both highly correlated with Sales,[43] do not have instead significant effects.

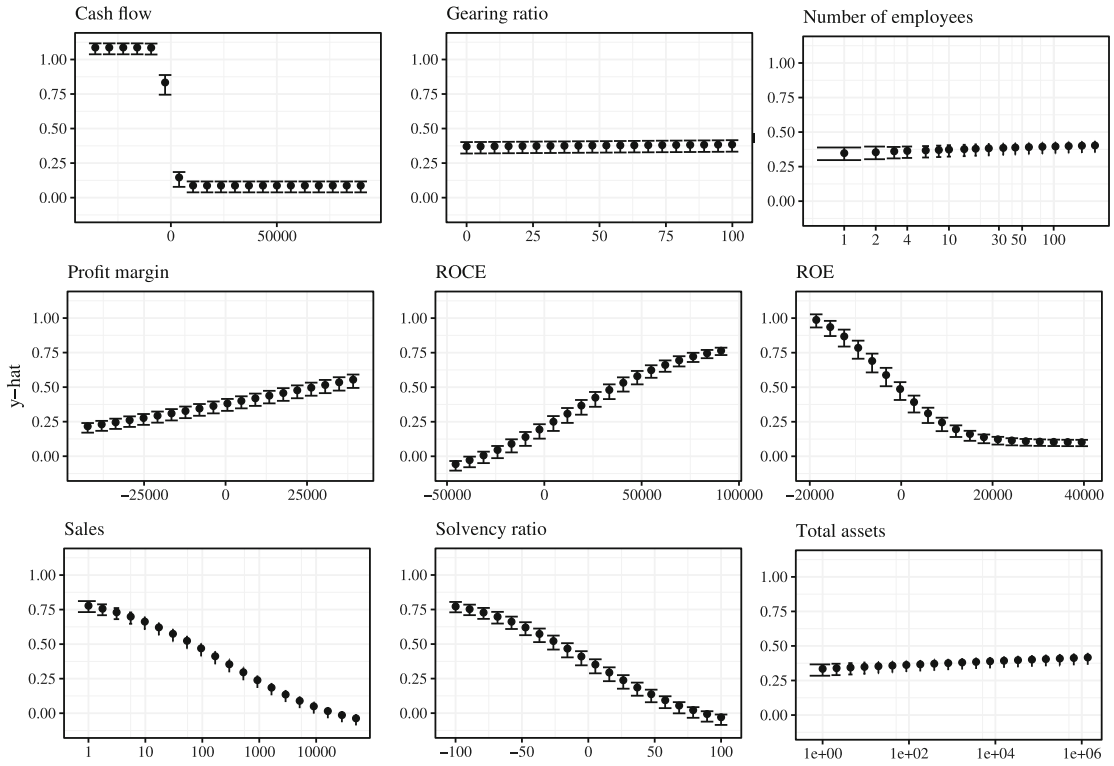
As expected, firms with a strongest leverage (Solvency ratio) and higher liquidity (Cash flow) are less likely to default.[55,56]

Notice that profitability measures, rather unexpectedly, do not impact on the probability of default according to significance criteria. BGEVA signals a significant ROCE but the estimated coefficient is zero. To gain additional insights, we can turn to the ALEs: the three common significant variables can be interpreted likewise since they all follow a non-flat path. However, while the models' coefficients for the Solvency ratio and Cash flow describe almost neutral effects on the outcome (with an odds-ratio of 1 for the Cash flow in the Probit model, see Table 3), post-hoc interpretation reveals a marked decreasing effect for the former and a clear non-linear pattern for the latter. On the other hand, and contrary to the p-value reading, we can observe that Profit margin and ROE do reduce the probability of default, whereas ROCE increases it according to the LR, Probit (see Figure 1, panels (a) and (b) respectively) and to the BGEVA model (Figure 2).

[§]In the text we refer to sales, total assets and number of employees for readability reasons. However, we have transformed them through logarithms as it is common in the literature.[54,81,82]



(A)



(B)

FIGURE 1 Accumulated local effects of the LR (A) and probit (B) models with superimposed bootstrap 5%–95% confidence intervals. The ALEs for sales, total assets and number of employees are calculated on log-transformed variables but depicted on anti-log values to enhance readability.

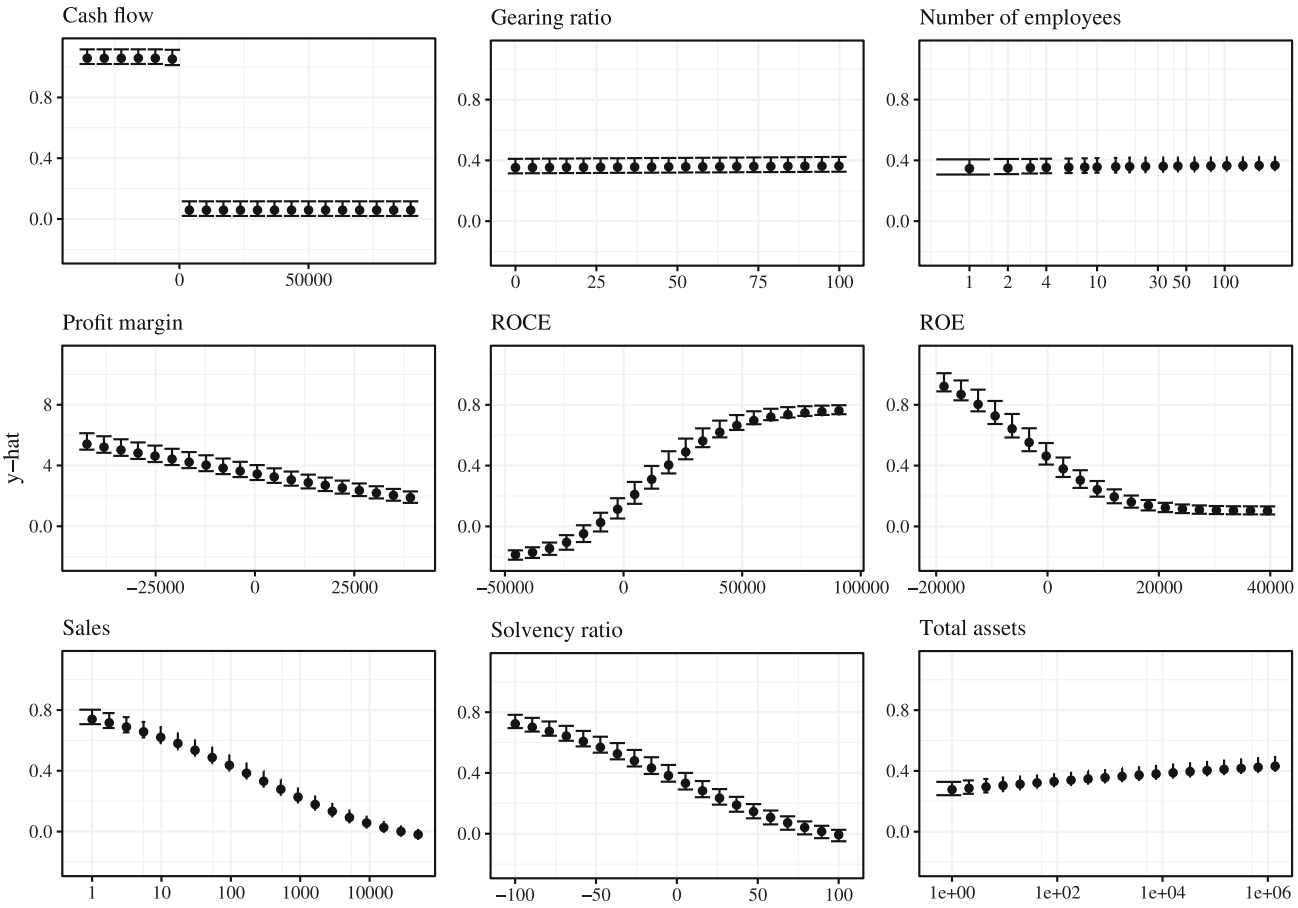


FIGURE 2 Accumulated local effects of the BGEVA model with superimposed bootstrap 5%–95% confidence intervals. The ALEs for sales, total assets and number of employees are calculated on log-transformed variables but depicted on anti-log values to enhance readability.

Another counterintuitive effect is revealed by the ALEs plot of the Profit margin for the Probit (Figure 1, panel (b)), which could partially explain the suboptimal classification performance of the same model.

The picture changes when it comes to black-box models. Global Shapley values indicate (Figure 3) that both FANN and XGBoost predictions are influenced mainly by Profit margin. This outcome is further clarified by the average change in the model output corresponding to increasing values of the variable, represented by ALEs (Figure 4).

The ALEs of either model show a downward sharp jump in the probability of default when moving from negative to positive values of Profit margin, with no further decrease in the probability of default as the ratio increases, revealing a clearly decreasing effect of this ratio on the probability of default, as previously found by References 54, 55, and 60.

The negative impact of Sales, already emerged in the white-box models, is confirmed to a minor extent by both FANN and XGBoost (second and third important variable respectively according to Shapley values). However, the pattern of the estimated default probabilities for Sales is unlike: a smooth path with no evident plateauing effect in FANN and a first sudden decrease around 100.000 euros and a second drop around 316.000 euros in XGBoost.

A remarkable difference with respect to the white-box models are the sways of Total assets and the Number of employees. Total assets is the third important variable for FANN according to the Shapley values and seems to increase the probability of default judging from ALEs. On the contrary, the variable shows no importance in the prediction by XGBoost (Shapley value close to 0 and flat ALE). A positive impact of Total assets on the probability of default is anomalous, though shared by other scholars,[55] in the light of our descriptive statistics and referring to the literature on firm demography, where exit is usually associated to less tangible assets.[56] This effect could be associated to the same found by other authors in the credit scoring literature. In that case a non-linear behaviour could be accounted to the fact that creditors do pursue firms with larger assets with the hope to get back the money they have lent, whereas firms with low tangible assets are less worth being pursued.[36,54]

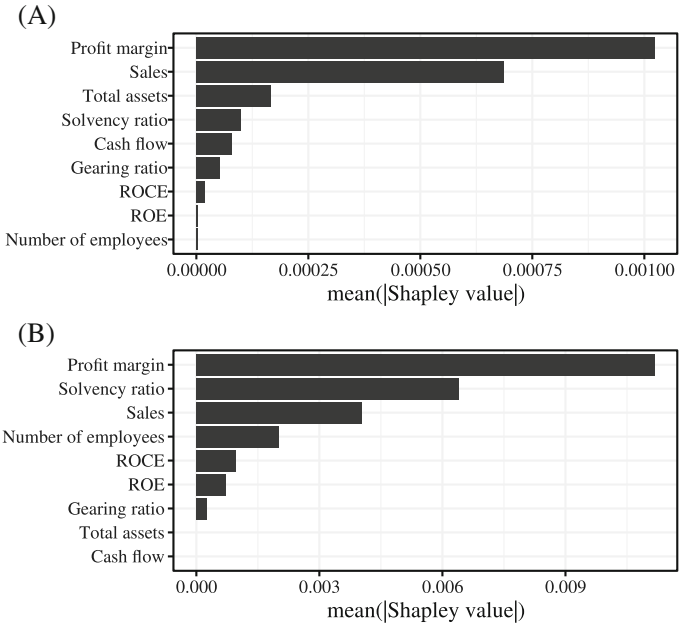


FIGURE 3 Global Shapley values for the Feedforward artificial neural network model (A) and the XGBoost model (B).

A somewhat opposite situation regards the Number of employees: FANN attributes scarce weight to this variable whereas XGBoost highlights its moderate impact (fourth important variable in the Shapley values) and a decrease in the probability of default around five employees. The XGBoost algorithm seems therefore able to capture separate and concordant effects of two firm size variables, respectively on the input and the output side, in decreasing the probability of default, contrary to other empirical applications.[55]

The Solvency ratio behaves similarly to Sales, for which the XGBoost shows a plateauing effect after 0 that the FANN does not point out. However, its importance, measured by the Shapley values, differs between the two algorithms since it is the second most relevant variable for XGBoost and the fourth relevant variable in the FANN.

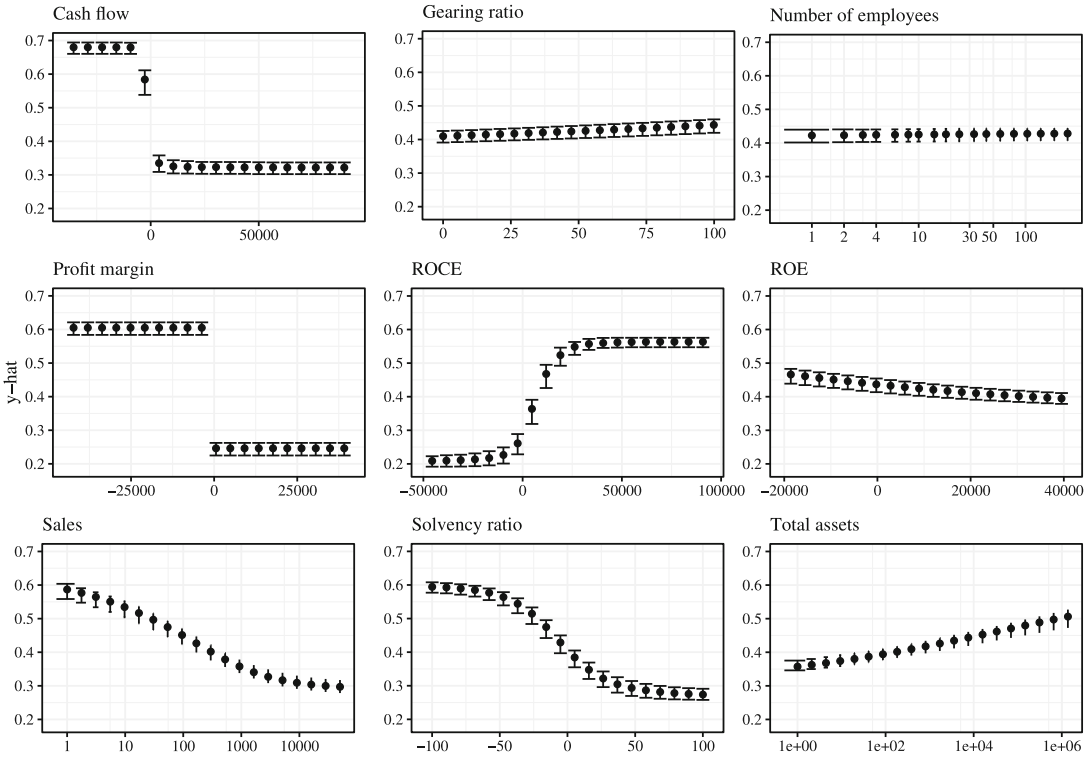
The Cash flow, the third variable impacting on default according to white-box models, maintains a negative sign also in FANN, while it is not relevant in the XGBoost model (as in Reference 56). The Gearing ratio, ROCE and ROE are of little consequence for XGBoost output and even less for the FANN according to the Shapley values and to overlapping bootstrap confidence intervals in Figure 4, except for the FANN's ALEs plot that displays ROCE (however small its importance) as enhancing the probability of default, which is in line with part of the literature (Reference 37 pointed out ROCE's positive effect). Another part of the literature instead found it non-significant.[59]

To summarize, blurry effects of one or more variables are encountered for the FANN model (Total assets and ROCE) and for all the white-box models (ROCE for all of them, Profit margin only for the Probit). Considering the prominent roles assigned by FANN to both Sales and Total assets, it seems that these two variables compensate each another in the wrong way, resulting in a the lowest correct classification of defaulted firms among the competing models.

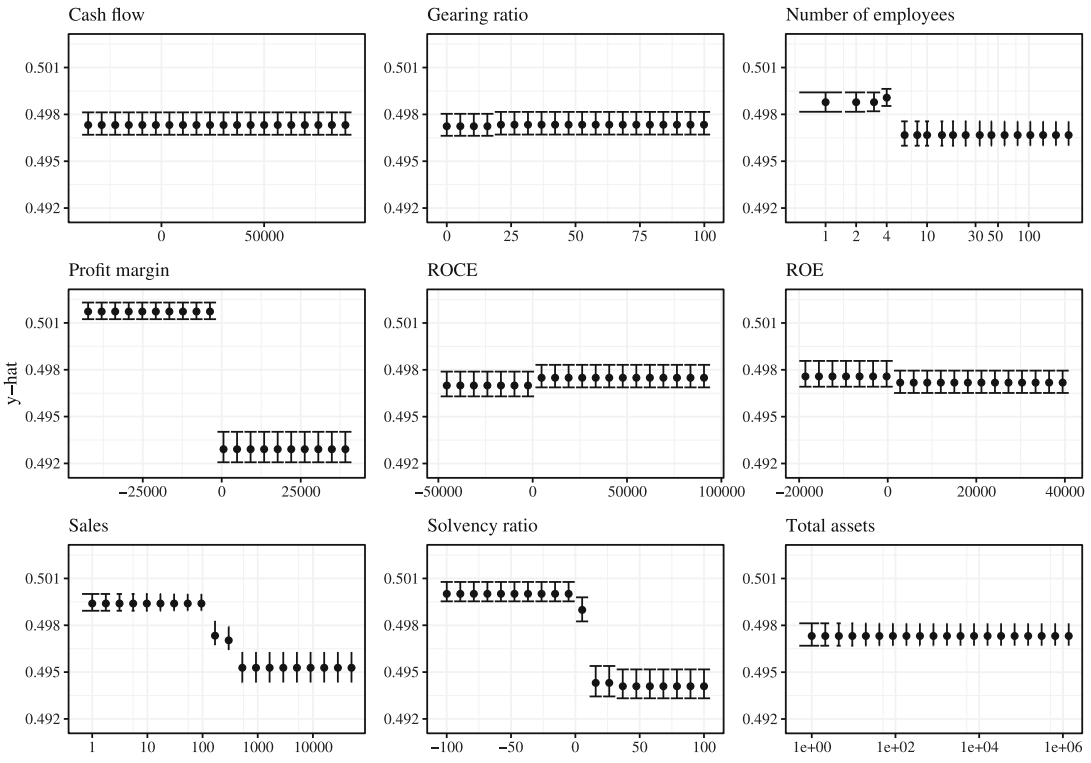
An interesting puzzle remains regarding the completely different ranking in the importance of variables according to white versus black-box models. Keeping performance in mind, we should consider what emerges from the interpretation of the XGBoost output, attributing the highest sensitivity achieved to an evaluation of the interplay among the variables which results more effective in predicting default.

6 | CONCLUSIONS

Making an AI system interpretable allows external observers to understand its working and meaning, with the non-negligible consequence of making it usable in practice: when a firm (or a customer) applies for a credit line, it has the right to be informed about the possible reasons for a refusal. AI driven decisions must be explained–as much as possible to and understood by those directly and indirectly affected, in order to allow the contesting of such decisions. This issue has become extremely relevant since both academicians and practitioners have progressively embraced ML modelling of



(A)



(B)

FIGURE 4 Accumulated local effects of the FANN (A) and XGBoost (B) with related bootstrap 5%–95% confidence intervals. The ALEs for sales, total assets and number of employees are calculated on log-transformed variables but depicted on anti-log values to enhance interpretability.

firm default due to excellent performances[2] and, concurrently, Institutions have started to question the trustworthiness of–and set boundaries for–a safe use of AI in the interest of all involved.[4] At the same time, using AI methods might grant larger amounts of credit and result in lower default rates.[83]

Here we contribute to the literature on SMEs default by showing that the good performances in classification tasks obtained through ML models can and should be accompanied by a clear interpretation of the role and type of effect played by the variables involved. We also contribute to the literature on global post-hoc interpretability showing that, differently from the ante-hoc techniques, they enable the comparison among white and black-boxes on a common ground.

Using a collection of relevant accounting indicators, widely employed in the literature, for all the Italian SMEs available in the BvD-Orbis dataset 2016, we have supplied an accurate prediction of default in 2017. Thanks to our research design, caring for imbalance among classes and cross-validation to select the most performing rules, we have achieved fair rates of correct classifications for all the models involved. However, focusing in particular on the correct rate of default classification, the XGBoost algorithm prevails over three white-box models and over the alternative ML model FANN.

Interpretability was provided by means of Shapley values and ALEs, two recent model-agnostic techniques which measure the relative importance of the predictors and shape the predictor-outcome relationship respectively. The analysis of the XGBoost ALEs reveals that such complex models capture highly non-linear patterns as the effects of sales on the probability of default, account for separate effects of correlated measures and suggest also non-trivial risky thresholds: a thing that was not completely grasped by any standard discriminant rule.

We think that the examination of ALEs for models which are already ante-hoc interpretable in the traditional scheme of statistical significance is quite revealing, both methodologically and empirically speaking. The latter models' ALEs permits to add different shades to the variables' effects with respect to the standard parameter-pvalues' paradigm, paradoxically uplifting their a-priori interpretability. Finally, the assessment of ALEs' variability is fundamental to check the output robustness and to evaluate the soundness of results.

With this paper we have shown that, under the assumption that interpretability is crucial to building and maintaining the users' trust in AI systems, their potential superiority in classification tasks does no longer need to be an alibi to hide the underlying mechanisms in black-boxes.

The relevancy of this approach could become definitely more important for default prediction based on alternative sources of data, such as web-scraped information,[84,85] whose dimensionality and complexity require the power of ML models and whose interpretability is even more puzzling. This, as well as applications to a more extensive basket of traditional predictors, might represent a good ground for further research.

This study has some limitations revolving around three main aspects. The first is given by the post-hoc nature of ALEs: post-hoc methods restrict the possibility to address any biases and impose some sort of regularization on the interpretations.[86] On the user's side, they require some basic knowledge of the methodology to interpret its outcomes. Second, the cross-sectional nature of the data prevented us from including in the analysis standard non-firm specific predictors, such as regional GDP growth, industry-level value added or business confidence indicators, which could have helped to reduce classification errors. Third, our findings, regarding Italian SMEs evaluated in a specific year, should be generalized with caution and would surely benefit from a cross-country comparison and a longitudinal follow-up.

ACKNOWLEDGMENTS

We thank the Italian Ministry of Education, University and Research (MIUR) for sponsoring this work under the 'Departments of Excellence 2018–2022' funding schema. We greatly acknowledge the DEMS Data Science Lab of the University of Milano–Bicocca for supporting this work by providing computational resources. We also thank two anonymous referees for helpful comments.

DATA AVAILABILITY STATEMENT

The data that support the findings will be available in Orbis-BvD at https://www.bvdinfo.com/en-gb/our-products/data/international/orbis following an embargo from the date of publication to allow for commercialization of research findings.

ORCID

Lisa Crosato https://orcid.org/0000-0002-3415-656X
Caterina Liberati https://orcid.org/0000-0001-9910-4018
Marco Repetto https://orcid.org/0000-0002-2606-0664

REFERENCES

1. European Commission. Annual Report on European SMEs 2018/2019, Tech. rep. 2019.
2. Ciampi F, Giannozzi A, Marzi G, Altman EI. Rethinking SME default prediction: a systematic literature review and future perspectives. *Scientometrics*. 2021;126:2141-2188.
3. Cornille D, Rycx F, Tojerow I. Heterogeneous effects of credit constraints on SMEs' employment: evidence from the European sovereign debt crisis. *J Financ Stab*. 2019;41:1-13.
4. European Commission. Ethics guidelines for trustworthy AI, Tech. rep. 2019.
5. Coussement K, Benoit DF. Interpretable data science for decision making. *Decis Support Syst*. 2021;150:113664.
6. High-Level Expert Group on Artificial Intelligence. European Commission, The Assessment List for Trustworthy Artificial Intelligence, Tech. rep. 2020.
7. Haykin SS. *Neural Networks: A Comprehensive Foundation*. 2nd ed. Prentice Hall; 1999.
8. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco California USA. 2016 785–794.
9. Apley DW, Zhu J. Visualizing the effects of predictor variables in black box supervised learning models. *J R Stat Soc Series B Stat Methodology*. 2020;82(4):1059-1086.
10. Bussmann N, Giudici P, Marinelli D, Papenbrock J. Explainable machine learning in credit risk management. *Comput Econ*. 2021;57(1):203-216.
11. Berg D. Bankruptcy prediction by generalized additive models. *Appl Stochast Models Bus Ind*. 2007;23(2):129-143.
12. Bank of England. Machine learning in UK financial services, Tech. rep. 2019.
13. Alonso A, Carbó JM. On the risk-adjusted performance of machine learning models in credit default prediction. *SUERF Policy Note*. 2020;210:1-10.
14. Institute of International Finance. *Machine Learning Governance Summary Report, Summary Report*. Institute of International Finance; 2020.
15. Institute of International Finance. *Machine Learning in Credit Risk, Tech. Rep*. Institute of International Finance; 2019.
16. Mai F, Tian S, Lee C, Ma L. Deep learning models for bankruptcy prediction using textual disclosures. *Eur J Oper Res*. 2019;274(2):743-758.
17. Gordini N. A genetic algorithm approach for SMEs bankruptcy prediction: empirical evidence from Italy. *Expert Syst Appl*. 2014;41(14):6433-6445.
18. Zhang L, Hu H, Zhang D. A credit risk assessment model based on SVM for small and medium enterprises in supply chain finance. *Financ Innov*. 2015;1(1):1-21.
19. Ciampi F, Gordini N. Small enterprise default prediction modeling through artificial neural networks: an empirical analysis of Italian small enterprises. *J Small Bus Manag*. 2013;51(1):23-45.
20. De Leonardis D, Rocci R. Assessing the default risk by means of a discrete-time survival analysis approach. *Appl Stochast Models Bus Ind*. 2008;24(4):291-306.
21. De Leonardis D, Rocci R. Default risk analysis via a discrete-time cure rate model. *Appl Stochast Models Bus Ind*. 2014;30(5):529-543.
22. European Banking Authority. EBA Report on Big Data and Advanced Analytics, Tech. rep. 2020.
23. Doran D, Schulz S, Besold TR. What does explainable AI really mean? A new conceptualization of perspectives. 2017 arXiv:1710.00794.
24. Lipton ZC. The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. *Queue*. 2018;16(3):31-57.
25. Montavon G, Samek W, Müller K-R. Methods for interpreting and understanding deep neural networks. *Digit Signal Process*. 2018; 73:1-15.
26. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. 2017 arXiv:1702.08608.
27. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. *ACM Comput Surv*. 2018;51(5):1-42.
28. Arrieta AB, Díaz-Rodríguez N, Del Ser J, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion*. 2020;58:82-115.
29. Ribeiro MT, Singh S, Guestrin C. Model-agnostic interpretability of machine learning. 2016 arXiv:1606.05386.
30. Du M, Liu N, Hu X. Techniques for interpretable machine learning. *Commun ACM*. 2019;63(1):68-77.
31. Lin SM, Ansell J, Andreeva G. Predicting default of a small business using different definitions of financial distress. *J Oper Res Soc*. 2012;63(4):539-548.
32. Modina M, Pietrovito F. A default prediction model for Italian SMEs: the relevance of the capital structure. *Appl Financ Econ*. 2014;24(23):1537-1554.
33. Ciampi F. Corporate governance characteristics and default prediction modeling for small enterprises. An empirical analysis of Italian firms. *J Bus Res*. 2015;68(5):1012-1025.
34. Holmes P, Hunt A, Stone I. An analysis of new firm survival using a hazard function. *Appl Econ*. 2010;42(2):185-195.
35. Gupta J, Gregoriou A, Ebrahimi T. Empirical comparison of hazard models in predicting SMEs failure. *Quant Finance*. 2018;18(3): 437-466.
36. El Kalak I, Hudson R. The effect of size on the failure probabilities of SMEs: an empirical study on the US market using discrete hazard model. *Int Rev Financ Anal*. 2016;43:135-145.
37. Calabrese R, Marra G, Angela Osmetti S. Bankruptcy prediction of small and medium enterprises using a flexible binary generalized extreme value model. *J Oper Res Soc*. 2016;67(4):604-615.

38. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat*. 2001;29(5):1189-1232.
39. Ribeiro MT, Singh S, Guestrin C. Why should I trust you? Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 1135–1144.
40. Lundberg S, Lee S-I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*. Curran Associates Inc; 2017:4765-4774.
41. Jones S, Wang T. Predicting private company failure: a multi-class analysis. *J Int Financ Markets Inst Money*. 2019;61:161-188.
42. Sigrist F, Hirnschall C. Grabit: gradient tree-boosted Tobit models for default prediction. *J Bank Financ*. 2019;102:177-192.
43. Jabeur SB, Gharib C, Mefteh-Wali S, Arfi WB. Catboost model and artificial intelligence techniques for corporate failure prediction. *Technol Forecast Soc Change*. 2021;166:120658.
44. Bücker M, Szepannek G, Gosiewska A, Biecek P. Transparency, auditability, and explainability of machine learning models in credit scoring. *J Oper Res Soc*. 2022;73(1):70-90.
45. Stevenson M, Mues C, Bravo C. The value of text for small business default prediction: a deep learning approach. *Eur J Oper Res*. 2021;295:758-771.
46. Yıldırım M, Okay FY, Özdemir S. Big data analytics for default prediction using graph theory. *Expert Syst Appl*. 2021;176:114840.
47. Liberati C, Camillo F, Saporta G. Advances in credit scoring: combining performance and interpretation in kernel discriminant analysis. *Adv Data Anal Classification*. 2017;11(1):121-138.
48. Glynn C. Learning low-dimensional structure in house price indices. *Appl Stochast Models Bus Ind*. 2022;38(1):151-168.
49. Gosiewska A, Kozak A, Biecek P. Simpler is better: lifting interpretability-performance trade-off via automated feature engineering. *Decis Support Syst*. 2021;150:113556.
50. Chen C, Lin K, Rudin C, Shaposhnik Y, Wang S, Wang T. A holistic approach to interpretability in financial lending: models, visualizations, and summary-explanations. *Decis Support Syst*. 2022;152:113647.
51. Bellandi M, Lombardi S, Santini E. Traditional manufacturing areas and the emergence of product-service systems: the case of Italy. *J Ind Bus Econ*. 2020;47:311-331.
52. Eurostat. Relative importance of Manufacturing (NACE Section C), EU. 2018.
53. EU. Commission Recommendation of 6 May 2003 concerning the definition of micro, small and medium-sized enterprises (Text with EEA relevance) (notified under document number C(2003) 1422), Tech. Rep. 32003H0361. 2003.
54. Altman EI, Sabato G, Wilson N. The value of non-financial information in small and medium-sized enterprise risk management. *J Credit Risk*. 2010;6(2):1-33.
55. Andreeva G, Calabrese R, Osmetti SA. A comparative analysis of the UK and Italian small businesses using generalised extreme value models. *Eur J Oper Res*. 2016;249(2):506-516.
56. Michala D, Grammatikos T, Filipe SF. Forecasting distress in European SME portfolios. *EIF Working Paper 2013/17*. European Investment Fund (EIF); 2013.
57. Succurro M, Mannarino L. The impact of financial structure on firms' probability of bankruptcy: a comparison across Western Europe convergence regions. *Region Sector Econ Stud*. 2014;14(1):81-94.
58. Jones S, Johnstone D, Wilson R. An empirical evaluation of the performance of binary classifiers in the prediction of credit ratings changes. *J Bank Financ*. 2015;56:72-85.
59. Giudici P, Hadji-Misheva B, Spelta A. Network based credit risk models. *Qual Eng*. 2020;32(2):199-211.
60. Petropoulos A, Siakoulis V, Stavroulakis E, Klamargias A. A robust machine learning approach for credit risk analysis of large loan-level datasets using deep learning and extreme gradient boosting. *IFC Bulletins Chapters*. Vol 50. Bank for International Settlements; 2019.
61. West D. Neural network credit scoring models. *Comput Oper Res*. 2000;27(11-12):1131-1152.
62. Baesens B, Van Gestel T, Viaene S, Stepanova M, Suykens J, Vanthienen J. Benchmarking state-of-the-art classification algorithms for credit scoring. *J Oper Res Soc*. 2003;54(6):627-635.
63. West D, Dellana S, Qian J. Neural network ensemble strategies for financial decision applications. *Comput Oper Res*. 2005;32(10):2543-2559.
64. Arifovic J, Gencay R. Using genetic algorithms to select architecture of a feedforward artificial neural network. *Phys A: Stat Mech Appl*. 2001;289(3-4):574-594.
65. Shapley LS. A value for n-person games. In: Kuhn HW, Tucker AW, eds. *Contributions to the Theory of Games (AM-28)*. Vol II. Princeton University Press; 1953:307-318.
66. Covert I, Lundberg S, Lee S-I. Understanding global feature contributions with additive importance measures. 2020. doi:10.48550/ARXIV.2004.00668
67. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. 2017. doi:10.48550/ARXIV.1703.01365
68. Friedman JH. Multivariate adaptive regression splines. *Ann Stat*. 1991;19(1):1-67.
69. Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. *J Comput Graph Stat*. 2015;24(1):44-65.
70. Ozgur O, Karagol ET, Ozbugday FC. Machine learning approach to drivers of bank lending: evidence from an emerging economy. *Financ Innov*. 2021;7(1):1-29.
71. Hechtlinger Y. Interpretation of Prediction Models Using the Input Gradient. 2016 arXiv:1611.07634.
72. Ribeiro MT, Singh S, Guestrin C. Anchors: high-precision model-agnostic explanations. Paper presented at: Aaai, Vol. 18, Association for the Advancement of Artificial Intelligence. 2018 1527–1535.

73. Lessmann S, Baesens B, Seow H-V, Thomas LC. Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research. *Eur J Oper Res*. 2015;247(1):124-136.
74. Xu Q-S, Liang Y-Z. Monte Carlo cross validation. *Chemom Intel Lab Syst*. 2001;56(1):1-11.
75. Davison AC, Kuonen D. An introduction to the bootstrap with applications in R. *Stat Comput Graph Newsletter*. 2002;13(1):6-11.
76. Baesens B, Höppner S, Ortner I, Verdonck T. robROSE: a robust approach for dealing with imbalanced data in fraud detection. *Stat Methods Appl*. 2021;3(30):841-861.
77. Veganzones D, Séverin E. An investigation of bankruptcy prediction in imbalanced datasets. *Decis Support Syst*. 2018;112:111-124.
78. Gong J, Kim H. RHSBoost: improving classification performance in imbalance data. *Comput Stat Data Anal*. 2017;111:1-13.
79. Santos MS, Soares JP, Abreu PH, Araujo H, Santos J. Cross-validation for imbalanced datasets: avoiding overoptimistic and overfitting approaches [research frontier]. *IEEE Comput Intell Mag*. 2018;13(4):59-76.
80. Breeden JL. Survey of Machine Learning in Credit Risk (May 30, 2020). 2020. doi:10.2139/ssrn.3616342
81. Altman EI, Sabato G. Modelling credit risk for SMEs: evidence from the US market. *Abacus*. 2007;43(3):332-357.
82. Psillaki M, Tsolas IE, Margaritis D. Evaluation of credit risk based on firm performance. *Eur J Oper Res*. 2010;201(3):873-881.
83. Moscatelli M, Parlapiano F, Narizzano S, Viggiano G. Corporate default forecasting with machine learning. *Expert Syst Appl*. 2020;161:113567.
84. Crosato L, Domenech J, Liberati C. Predicting SME's default: are their websites informative? *Econ Lett*. 2021;204:109888.
85. Crosato L, Domenech J, Liberati C. Websites' data: a new asset for enhancing credit risk modeling. *Ann Oper Res*. 2023;1-16. doi:10.1007/s10479-023-05306-5
86. Repetto M. Multicriteria interpretability driven deep learning. *Ann Oper Res*. 2022;1-15. doi:10.1007/s10479-022-04692-6
87. Blume JD. Bounding sample size projections for the area under a roc curve. *J Stat Plan Inference*. 2009;139:711-721.
88. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321-357.

How to cite this article: Crosato L, Liberati C, Repetto M. Lost in a black-box? Interpretable machine learning for assessing Italian SMEs default. *Appl Stochastic Models Bus Ind*. 2023;39(6):829-846. doi: 10.1002/asmb.2803

APPENDIX A. ALTERNATIVE TRAINING AND TEST SPLITTING

As a robustness check, we have re-run the models using two alternative data splits: 80% training −20% test and 90% training − 10% test, evaluating the outcomes via AUC and H performance metrics computed on the test sets (Table A1). According to the results, the XGboost turns out to be the best solution in all the split cases, showing the highest value for both the performance indicators. Following Reference 87, we do not compare solutions across splits because the empirical AUC depends on the sample size.

TABLE A1 Test set results for alternative training and test splits.

	Split 80%–20%		Split 90%–10%	
Model	**AUC**	**H**	**AUC**	**H**
FANN	0.798	0.345	0.827	0.364
XGBoost	0.867	0.444	0.872	0.440
BGEVA	0.820	0.343	0.820	0.333
LR	0.815	0.322	0.814	0.321
Probit	0.813	0.323	0.815	0.321

The different splits in training (tr) and test (ts) sets allocate data as follows:

- 70%–30%:
  - tr 1285 (failed), 73,540 (survived)
  - ts 522 (failed), 31,518 (survived)
- 80%–20%:
  - tr 1462 (failed), 84,046 (survived)
  - ts 345 (failed), 21,012 (survived)

- 90%–10%:
  - tr 1646 (failed), 94,552 (survived)
  - ts 161 (failed), 10,506 (survived)

As can be seen, the more we feed the training sample the less the test set is able of giving a variegated representation of the minority class. Considering this and following some relevant studies in the literature[17,37,62] we believe that the split 70%–30%, combined with the cross-validation scheme, guarantees enough variability for the estimation process, preserving on the same time a sufficient diversity among the defaulted firms in the test set.

APPENDIX B. CLASSIFICATION RESULTS UNDER DIFFERENT RESAMPLING SCHEMES

Before opting for undersampling, we have tried other sampling schemes such as SMOTE[88] and RobRose.[76] To evaluate which scheme was the most suitable to accomplish our goal, we have considered two aspects, equally important:

1. A fair rate of correct classification in both groups of firms;
2. A good overall performance in terms of H-measure, AUC, Brier Score and Kolmogorov-Smirnov statistic

In order to guarantee a sensible comparison across models, as well as sound estimates, we have selected the scheme that satisfied the above aspects across models. Table B1 reports also the metrics obtained over the whole dataset (about

TABLE B1 Classification results according to different resampling schemes (undersampling, SMOTE, robROSE) and for the whole dataset.

Model	Sampling scheme	Sensitivity	Specificity	H	AUC	BS	KS
FANN	*whole dataset*	*0.000*	*1.000*	*0.385*	*0.830*	*0.081*	*0.469*
FANN	Undersampling	0.694	0.829	0.391	0.827	0.187	0.501
FANN	SMOTE	0.625	0.872	0.373	0.837	0.167	0.523
FANN	robROSE	0.770	0.692	0.309	0.793	0.112	0.362
XGBoost	*whole dataset*	*0.390*	*0.999*	*0.406*	*0.803*	*0.019*	*0.551*
XGBoost	Undersampling	0.821	0.719	0.383	0.843	0.146	0.552
XGBoost	SMOTE	0.559	0.930	0.418	0.852	0.086	0.548
XGBoost	robROSE	0.613	0.842	0.335	0.771	0.024	0.521
BGEVA	*whole dataset*	*0.002*	*1.000*	*0.287*	*0.799*	*0.021*	*0.437*
BGEVA	Undersampling	0.752	0.727	0.331	0.819	0.178	0.481
BGEVA	SMOTE	0.657	0.810	0.309	0.813	0.157	0.463
BGEVA	robROSE	0.809	0.634	0.298	0.807	0.191	0.451
LR	*whole dataset*	*0.010*	*0.999*	*0.281*	*0.796*	*0.022*	*0.418*
LR	Undersampling	0.745	0.736	0.303	0.809	0.151	0.483
LR	SMOTE	0.662	0.808	0.306	0.811	0.158	0.461
LR	robROSE	0.824	0.638	0.310	0.814	0.179	0.452
Probit	*whole dataset*	*0.003*	*1.000*	*0.280*	*0.795*	*0.021*	*0.430*
Probit	Undersampling	0.738	0.737	0.299	0.809	0.190	0.448
Probit	SMOTE	0.627	0.809	0.282	0.799	0.420	0.331
Probit	robROSE	0.120	0.987	0.074	0.554	0.192	0.459

105,000 firms, of which 1.72% defaulted). Looking at the metrics obtained through the different resampling schemes, we can notice that the only scheme that guarantees a minimum rate of correct classification of 70% in both classes and across models is the undersampling, with the exception of the robROSE in the case of the FANN model (see Table B1). Furthermore, the undersampling scheme is the most frequently selected across measures and models (10 times vs. 5 each for the other schemes).