Computational Economics (2021) 57:203–216
https://doi.org/10.1007/s10614-020-10042-0

Explainable Machine Learning in Credit Risk Management

Niklas Bussmann[1] · Paolo Giudici[1] · Dimitri Marinelli[2] · Jochen Papenbrock[3]

Accepted: 17 August 2020 / Published online: 25 September 2020

Abstract

The paper proposes an explainable Artificial Intelligence model that can be used in credit risk management and, in particular, in measuring the risks that arise when credit is borrowed employing peer to peer lending platforms. The model applies correlation networks to Shapley values so that Artificial Intelligence predictions are grouped according to the similarity in the underlying explanations. The empirical analysis of 15,000 small and medium companies asking for credit reveals that both risky and not risky borrowers can be grouped according to a set of similar financial characteristics, which can be employed to explain their credit score and, therefore, to predict their future behaviour.

Keywords Credit risk management · Explainable AI · Financial technologies · Similarity networks

1 Introduction

Black box Artificial Intelligence (AI) is not suitable in regulated financial services. To overcome this problem, Explainable AI models, which provide details or reasons to make the functioning of AI clear or easy to understand, are necessary.

✉ Paolo Giudici
giudici@unipv.it

Niklas Bussmann
Niklas.bussmann01@universitadipavia.it

Dimitri Marinelli
dm@financial-networks.eu

Jochen Papenbrock
jp@firamis.de

[1] University of Pavia, Pavia, Italy

[2] FinNet-Project, Frankfurt, Germany

[3] FIRAMIS, Frankfurt, Germany

To develop such models, we first need to understand what "Explainable" means. Recently, some important insitutional definitions have been provided. For example, Bracke et al. (2019) states that "Explanations can answer different kinds of questions about a model's operation depending on the stakeholder they are addressed to and Croxson et al. (2019)" 'interpretability' will be the focus will be the focus—generally taken to mean that an interested stakeholder can comprehend the main drivers of a model-driven decision".

Explainability means that an interested stakeholder can comprehend the main drivers of a model-driven decision; FSB (2017) suggests that "lack of interpretability and auditability of AI and Machine Learning (ML) methods could become a macro-level risk"; Croxson et al. (2019) establishes that "in some cases, the law itself may dictate a degree of explainability."

The European GDPR EU (2016) regulation states that "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." Under the GDPR regulation, the data subject is therefore, under certain circumstances, entitled to receive meaningful information about the logic of automated decision-making.

Finally, the European Commission High-Level Expert Group on AI presented the Ethics Guidelines for Trustworthy Artificial Intelligence in April 2019. Such guidelines put forward a set of seven key requirements that AI systems should meet in order to be deemed trustworthy. Among them three relate to the concept of "eXplainable Artificial Intelligence (XAI)" , and are the following.

- Human agency and oversight: decisions must be informed, and there must be a human-in-the-loop oversight.
- Transparency: AI systems and their decisions should be explained in a manner adapted to the concerned stakeholder. Humans need to be aware that they are interacting with an AI system.
- Accountability: AI systems should develop mechanisms for responsibility and accountability, auditability, assessment of algorithms, data and design processes.

Following the need to explain AI models, stated by legislators and regulators of different countries, many established and startup companies have started to embrace Explainable AI models. In addition, more and more people are searching information about what "Explainable Artificial Intelligence" means.

In this respect, Fig. 1 represents the evolution of Google searches for explainable AI related terms.

From a mathematical viewpoint, it is well known that "simple" statistical learning models, such as linear and logistic regression models, provide a high interpretability but, possibly, a limited predictive accuracy. On the other hand, "complex" machine learning models, such as neural networks and tree models, provide a high predictive accuracy at the expense of a limited interpretability.

To solve this trade-off, we propose to boost machine learning models, that are highly accurate, with a novel methodology, that can explain their predictive output.

Our proposed methodology acts in the post processing phase of the analysis, rather than in the preprocessing part. It is agnostic (technologically neutral) as it is applied to the predictive output, regardless of which model generated it: a linear regression, a classification tree or a neural network model.

The machine learning procedure proposed in the paper processes the outcomes of any other arbitrary machine learning model. It provides more insight, control and transparency to a trained, potentially black box machine learning model. It utilises a model-agnostic method aiming at identifying the decision-making criteria of an AI system in the form of variable importance (individual input variable contributions).

A key concept of our model is the Shapley value decomposition of a model, a pay-off concept from cooperative game theory. To the best of our knowledge this is the only explainable AI approach rooted in an economic foundation. It offers a breakdown of variable contributions so that every data point (e.g. a credit or loan customer in a portfolio) is not only represented by input features (the input of the machine learning model) but also by variable contributions to the prediction of the trained machine learning model.

More precisely, our proposed methodology is based on the combination of network analysis with Shapley values [see Lundberg and Lee (2017), Joseph (2019), and references therein]. Shapley values were originally introduced by Shapley (1953) as a solution concept in cooperative game theory. They correspond to the average of the marginal contributions of the players associated with all their possible orders. The advantage of Shapley values, over alternative XAI models, is that they can be exploited to measure the contribution of each explanatory variable for each point prediction of a machine learning model, regardless of the underlying model itself [see, e.g. Lundberg and Lee (2017)]. In other words, Shapley based XAI models combine generality of application (they are model agnostic) with the personalisation of their results (they can explain any single point prediction).

Our original contribution is to improve Shapley values, improving the interpretation of the predictive output of a machine learning model by means of correlation network models. To exemplify our proposal, we consider one area of the financial industry in which Artificial Intelligence methods are increasingly being applied: credit risk management [see for instance the review by Giudici (2018)].

Correlation networks, also known as similarity networks, have been introduced by Mantegna and Stanley (1999) to show how time series of asset prices can be clustered in groups on the basis of their correlation matrix. Correlation patterns between companies can similarly be extracted from cross-sectional features, based on balance sheet data, and they can be used in credit risk modelling. To account for such similarities we can rely on centrality measures, following Giudici et al. (2019) , who have shown that the inclusion of centrality measures in credit scoring models does improve their predictive utility. Here we propose a different use of similarity networks. Instead of applying network models in a pre-processing phase, as in Giudici et al. (2019) , who extract from them additional features to be included in a statistical learning model, we use them in a post-processing phase, to interpret the predictive output from a highly performing machine learning model. In this way we achieve both predictive accuracy and explainability.

We apply our proposed method to predict the credit risk of a large sample of small and medium enterprises. The obtained empirical evidence shows that, while improving the predictive accuracy with respect to a standard logistic regression model, we improve, the interpretability (explainability) of the results.

The rest of the paper is organized as follows: Sect. 2 introduces the proposed methodology. Section 3 shows the results of the analysis in the credit risk context. Section 4 concludes and presents possible future research developments.

2 Methodology

2.1 Statistical Learning of Credit Risk

Credit risk models are usually employed to estimate the expected financial loss that a credit institution (such as a bank or a peer-to-peer lender) suffers, if a borrower defaults to pay back a loan. The most important component of a credit risk model is the probability of default, which is usually estimated statistically employing credit scoring models.

Borrowers could be individuals, companies, or other credit institutions. Here we focus, without loss of generality, on small and medium enterprises, whose financial data are publicly available in the form of yearly balance sheets.

For each company, n, define a response variable Y_n to indicate whether it has defaulted on its loans or not, i.e. $Y_n = 1$ if company defaults, $Y_n = 0$ otherwise. And let X_n indicate a vector of explanatory variables. Credit scoring models assume that the response variable Y_n may be affected ("caused") by the explanatory variables X_n.

The most commonly employed model of credit scoring is the logistic regression model. It assumes that

$$ln\left(\frac{p_n}{1-p_n}\right) = \alpha + \sum_{j=1}^{J} \beta_j x_{nj} \tag{1}$$

where p_n is the probability of default for company n; $\mathbf{x}_n = (x_{i,1}, \dots, x_{i,J})$ is a J-dimensional vector containing the values that the J explanatory variables assume for company n; the parameter α represents an intercept; β_j is the jth regression coefficient.

Once the parameters α and β_j are estimated using the available data, It the probability of default can be estimated, inverting the logistic regression model, from:

$$p_n = \left(1 + exp\left(\alpha + \sum_{j=1}^{J} \beta_j x_{nj}\right)\right)^{-1} \tag{2}$$

2.2 Machine Learning of Credit Risk

Alternatively, credit risk can be measured with Machine Learning (ML) models, able to extract non-linear relations among the financial information contained in the balance sheets. In a standard data science life cycle, models are chosen to optimise the predictive accuracy. In highly regulated sectors, like finance or medicine, models should be chosen balancing accuracy with explainability (Murdoch et al. 2019). We improve the choice selecting models based on their predictive accuracy, and employing a posteriori an algorithm that achieves explanability. This does not limit the choice of the best performing models.

To exemplify our approach we consider, without loss of generality, the Extreme Gradient Boost model, one of the most popular and fast machine learning algorithms [see e.g. Chen and Guestrin (2016)].

Extreme Gradient Boosting (XGBoost) is a supervised model based on the combination of tree models with Gradient Boosting. Gradient Boosting is an optimisation technique able to support different learning tasks, such as classification, ranking and prediction. A tree model is a supervised classification model that searches for the partition of the explanatory variables that best classify a response (supervisor) variable. Extreme Gradient Boosting improves tree models strengthening their classification performance, as shown by Chen and Guestrin (2016). The same authors also show that XGBoost is faster than tree model algorithms.

In practice, a tree classification algorithm is applied successively to “training” samples of the data set. In each iteration, a sample of observations is drawn from the available data, using sampling weights which change over time, weighting more the observations with the worst fit. Once a sequence of trees is fit, and classifications made, a weighted majority vote is taken. For a more detailed description of the algorithm see, for instance (Friedman et al. 2000).

2.3 Learning Model Comparison

Once a default probability estimation model is chosen, it should be measured in terms of predictive accuracy, and compared with other models, so to select the best one. The most common approach to measure predictive accuracy of credit scoring models is to randomly split the available data in two parts: a “train” and a “test” set; build the model using data in the train set, and compare the predictions the model obtains on the test set, $\hat{Y}_n$, with the actual values of Y_n.

To obtain $\hat{Y}_n$ the estimated default probability is rounded into a “default” or “non default”, depending on whether a threshold is passed or not. For a given threshold T, one can then count the frequency of the four possible outputs, namely: False Positives (FP): companies predicted to default, that do not; True Positives (TP): companies predicted to default, which do; False Negatives (FN): companies predicted not to default, which do; True Negatives (TN): companies predicted not to default, which do not.

The misclassification rate of a model can be computed as:

$$\frac{FP + FN}{TP + TN + FP + FN} \tag{3}$$

and it characterizes the proportion of wrong predictions among the total number of cases.

The misclassification rate depends on the chosen threshold and it is not, therefore, a generally agreed measure of predictive accuracy. A common practice is to use the Receiver Operating Characteristics (ROC) curve, which plots the false positive rate (FPR) on the Y axis against the true positive rate (TPR) on the X axis, for a range of threshold values (usually percentile values). FPR and TPR are then calculated as follows:

$$FPR = \frac{FP}{FP + TN} \tag{4}$$

$$TPR = \frac{TP}{TP + FN} \tag{5}$$

The ideal ROC curve coincides with the Y axis, a situation which cannot be realistically achieved. The best model will be the one closest to it. The ROC curve is usually summarised with the Area Under the ROC curve value (AUROC), a number between 0 and 1. The higher the AUROC, the better the model.

2.4 Explaining Model Predictions

We now explain how to exploit the information contained in the explanatory variables to localise and cluster the position of each individual (company) in the sample. This information, coupled with the predicted default probabilities, allows a very insightful explanation of the determinant of each individual's creditworthiness. In our specific context, information on the explanatory variables is derived from the financial statements of borrowing companies, collected in a vector $\mathbf{x}_n$, representing the financial composition of the balance sheet of institution n.

We propose to calculate the Shapley value associated with each company. In this way we provide an agnostic tool that can interpret in a technologically neutral way the output from a highly accurate machine learning model. As suggested in Joseph (2019), the Shapley values of a model can be used as a tool to transfer predictive inferences into a linear space, opening a wide possibility of applying to them a variety of multivariate statistical methods.

We develop our Shapley approach using the SHAP Lundberg and Lee (2017) computational framework, which allows to estimate Shapley values expressing predictions as linear combinations of binary variables that describe whether each single variable is included or not in the model.

More formally, the explanation model $g(x')$ for the prediction $f(x)$ is constructed by an additive feature attribution method, which decomposes the prediction into a linear function of the binary variables $z' \in \{0, 1\}^M$ and the quantities $\phi_i \in \mathbb{R}$:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i. \tag{6}$$

In other terms, $g'(z') \approx f(h_x(z'))$ is a local approximation of the predictions where the local function $h_x(x') = x$ maps the simplified variables x' into x, $z' \approx x$ and M is the number of the selected input variables.

Indeed, Lundberg and Lee (2017) prove that the only additive feature attribution method that satisfies the properties of *local accuracy*, *missingness* and *consistency* is obtained attributing to each feature x'_i an effect ϕ_i called Shapley value, defined as

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} \left[f_x(z') - f_x(z' \backslash i)\right] \tag{7}$$

where f is the trained model, x the vector of inputs (features), x' the vector of the M selected input features. The quantity $f_x(z') - f_x(z' \backslash i)$ is the contribution of a variable i and expresses, for each single prediction, the deviation of Shapley values from their mean.

In other words, a Shapley value represents a unique quantity able to construct an explanatory model that locally linearly approximate the original model, for a specific input x,(*local accuracy*). With the property that, whenever a feature is locally zero, the Shapley value is zero (*missingness*) and if in a second model the contribution of a feature is higher, so will be its Shapley value (*consistency*).

Once Shapley values are calculated, we propose to employ similarity networks, defining a metric that provides the relative distance between companies by applying the Euclidean distance between each pair $(\mathbf{x}_i, \mathbf{x}_j)$ of company predicted vectors, as in Giudici et al. (2019).

We then derive the Minimal Spanning Tree (MST) representation of the companies, employing the correlation network method suggested by Mantegna and Stanley (1999). The MST is a tree without cycles of a complex network, that joins pairs of vertices with the minimum total "distance".

The choice is motivated by the consideration that, to represent all pairwise correlations between N companies in a graph, we need $N * (N - 1)/2$ edges, a number that quickly grows, making the corresponding graph not understandable. The Minimal Spanning Tree simplifies the graph into a tree of $N - 1$ edges, which takes $N - 1$ steps to be completed. At each step, it joins the two companies that are closest, in terms of the Euclidean distance between the corresponding explanatory variables.

In our Shapley value context, the similarity of variable contributions is expressed as a symmetric matrix of dimension $n \times n$, where n Is the number of data points in the (train) data set. Each entry of the matrix measures how similar or distant a pair of data points is in terms of variable contributions. The MST representation associates to each point its closest neighbour. To generate the MST we have used the EMST Dual-Tree Boruvka algorithm, and its implementation in the R package "emstreeR".

The same matrix can also be used, in a second step, for a further merging of the nodes, through cluster analysis. This extra step can reveal segmentations of data points with very similar variable contributions, corresponding to similar credit scoring decision making.

3 Application

3.1 Data

We test our proposed model to data supplied by European External Credit Assessment Institution (ECAI) that specializes in credit scoring for P2P platforms focused on SME commercial lending. The data is described by Giudici et al. (2019) to which we refer for further details. In summary, the analysis relies on a dataset composed of official financial information (balance-sheet variables) on 15,045 SMEs, mostly based in Southern Europe, for the year 2015. The information about the status (0 = active, 1 = defaulted) of each company one year later (2016) is also provided. The proportion of defaulted companies within this dataset is 10.9%.

Using this data, Giudici et al. (2019) have constructed logistic regression scoring models that aim at estimating the probability of default of each company, using the available financial data from the balance sheets and, in addition, network centrality measures that are obtained from similarity networks.

Here we aim to improve the predictive performance of the model and, for this purpose, we run an XGBoost tree algorithm [see e.g. Chen and Guestrin (2016)]. To explain the results from the model, typically highly predictive, we employ similarity network models, in a post-processing step. In particular, we employ the cluster dendrogram representation that corresponds to the application of the Minimum Spanning Tree algorithm.

3.2 Results

We first split the data in a training set (80%) and a test set (20%), using random sampling without replacement.

We then estimate the XGBoost model on the training set, apply the obtained model to the test set and compare it with the best logistic regression model. The ROC curves of the two models are contained in Fig. 1 below.

From Fig. 1 note that the XGBoost clearly improves predictive accuracy. Indeed the comparison of the Area Under the ROC curve (AUROC) for the two models indicate an increase from 0.81 (best logistic regression model) to 0.93 (best XGBoost model).

We then calculate the Shapley value explanations of the companies in the test set, using the values of their explanatory variables. In particular, we use TreeSHAP method (Lundberg et al. 2020) [see e.g. Murdoch et al. (2019); Molnar 2019)] in combination with XGBoost. The Minimal Spanning Tree (a single linkage cluster) is used to simplify and interpret the structure present among Shapley values. We can also "colour" the MST graph in terms of the associated response variables values: default, not default.

Figures 2 and 3 present the MST representation. While in Fig. 3 company nodes are colored according to the cluster to which they belong, in Fig. 4 they are colored according to their status: not defaulted (grey); defaulted (red).

In Fig. 2, nodes are colored according to the cluster in which they are classified. The figure shows that clusters are quite scattered along the correlation network.

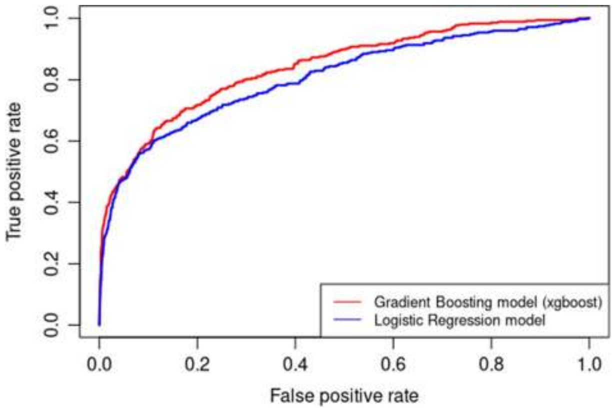

Fig. 1 Receiver Operating Characteristic (ROC) curves for the logistic credit risk model and for the XGBoost model. In blue, we show the results related to the logistic models while in red we show the results related to the XGBoost model

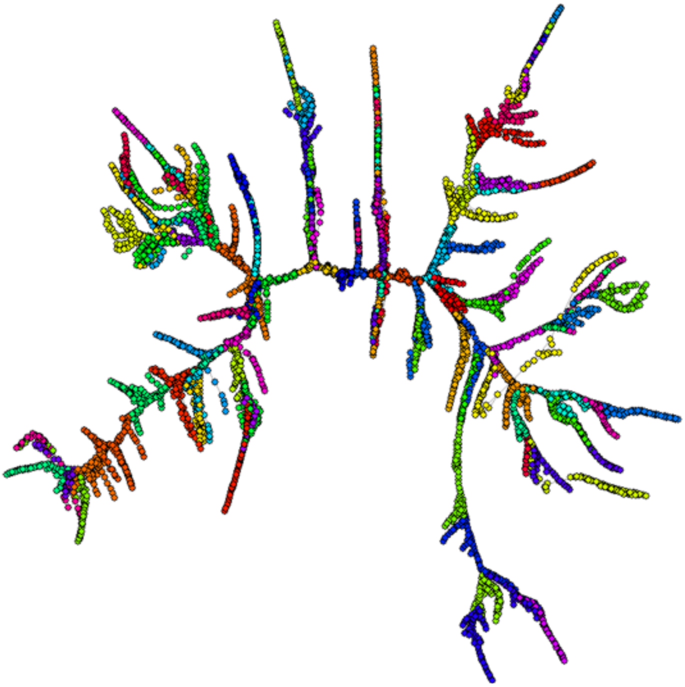

Fig. 2 Minimal Spanning Tree representation of the borrowing companies. Companies are colored according to their cluster of belonging

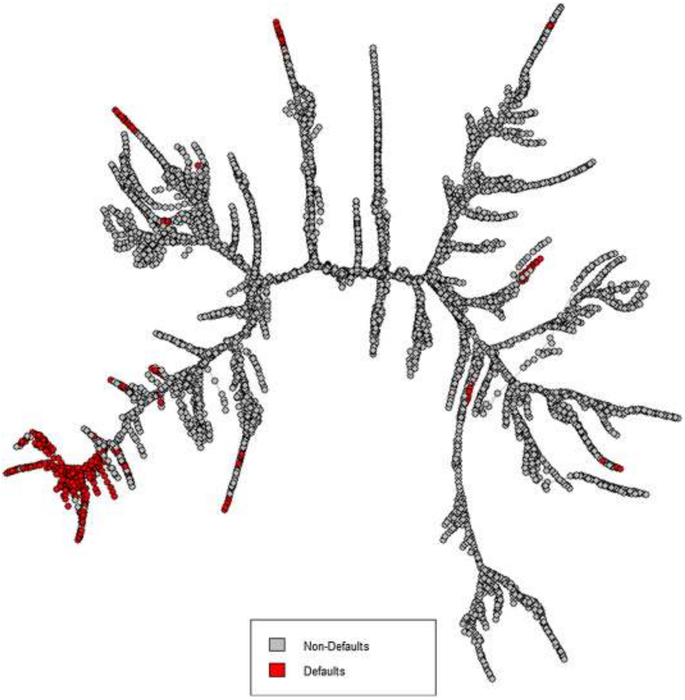

Fig. 3 Minimal Spanning Tree representation of the borrowing companies. Clustering has been performed using the standardized Euclidean distance between institutions. Companies are colored according to their default status: red = defaulted; grey = not defaulted

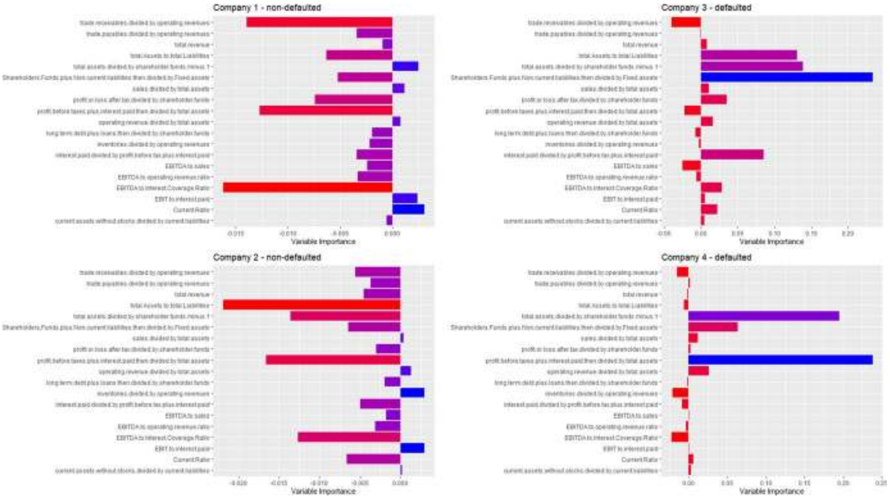

Fig. 4 Contribution of each explanatory variable to the Shapley's decomposition of four predicted default probabilities, for two defaulted and two non defaulted companies. The more red the color the higher the negative importance, and the more blue the color the higher the positive importance

To construct the colored communities in Fig. 2, we used the algorithm implemented in the R package “igraph” that directly optimizes a modularity score. The algorithm is very efficient and easily scales to very large networks (Clauset et al. 2004).

In Fig. 3, nodes are colored in a simpler binary way: whether the corresponding company has defaulted or not.

From Fig. 3 note that default nodes appear grouped together in the MST representation, particularly along the bottom left branch. In general, defaulted institutions occupy precise portion of the network, usually to the leafs of the tree, and form clusters. This suggests that those companies form communities, characterised by similar predictor variables’ importances. It also suggests that not defaulted companies that are close to default ones have a high risk of becoming defaulted as well, being the importance of their predictor variables very similar to those of the defaulted companies.

To better explain the explainability of our results, in Fig. 4 we provide the interpretation of the estimated credit scoring of four companies: two that actually defaulted and two that did not.

Figure 4 clearly shows the advantage of our explainable model. It can indicate which variables contribute more to the prediction of default. Not only in general, as is typically done by statistical and machine learning models, but differently and specifically for each company in the test set. Indeed, Fig. 4 clearly shows how the explanations are different (“personalised”) for each of the four considered companies.

The most important variables, for the two non defaulted companies (left boxes) regard: profits before taxes plus interests paid, and earnings before income tax and depreciation (EBITDA), which are common to both; trade receivables, for company 1; total assets, for company 2.

Economically, a high proficiency decreases the probability of default, for both companies; whereas a high stock of outstanding invoices, not yet paid, or a large stock of assets, helps reducing the same probability.

On the other hand, Fig. 4 shows that the most important variables, for the two defaulted companies (right boxes) concern: total assets, for both companies; shareholders funds plus non current liabilities, for company 3; profits before taxes plus interests paid, for company 4.

In other words, lower total assets coupled, in one case, with limited shareholder funds and, in the other, with low proficiency, increase the probability of default of these two companies.

The above results are consistent with previous analysis of the same data: both Giudici et al. (2019) select, as most important variables in several models, the return on equity, related to both EBITDA and profit before taxes plus interests paid; the leverage, related to total assets and shareholders’ funds; and the solvency ratio, related to trade payables.

We remark that Fig. 4 contains a “local” explanation of the predictive power of the explanatory variables, and it is the most important contribution of Shapley value theory. If we average Shapley values across all observations we get an “overall” or “global” explanation, similar to what already available in the

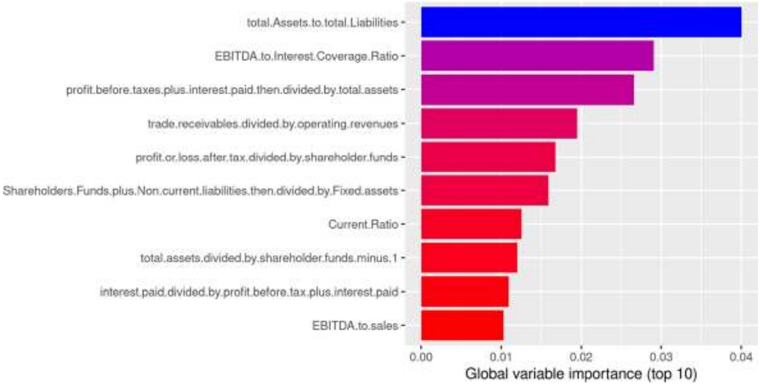

Fig. 5 Mean contribution of each explanatory variable to the Shapley's decomposition. The more red the color the higher the negative importance, and the more blue the color the higher the positive importance

statistical and machine learning literature. Figure 5 below provides the global explanation in our context: the ten most important explanatory variables, over the whole sample.

From Fig. 5 note that total assets to total liabilities (the leverage) is the most important variable, followed by the EBITDA, along with profit before taxes plus interest paid, measures of operational efficiency; and by trade receivables, related to solvency, in line with the previous comments.

4 Conclusions and Future Research

The need to leverage the high predictive accuracy brought by sophisticated machine learning models, making them interpretable, has motivated us to introduce an agnostic, post-processing methodology, based on correlation network models. The model can explain, from a substantial viewpoint, any single prediction in terms of the Shapley value contribution of each explanatory variables.

For the implementation of our model, we have used TreeSHAP, a consistent and accurate method, available in open-source packages. TreeSHAP is a fast algorithm that can compute SHapley Additive exPlanation for trees in polynomial time instead of the classical exponential runtime. For the xgboost part of our model we have used NVIDIA GPUs to considerably speed up the computations. In this way, the TreeSHAP method can quickly extract the information from the xgboost model.

Our research has important policy implications for policy makers and regulators who are in their attempt to protect the consumers of artificial intelligence services. While artificial intelligence effectively improve the convenience and accessibility of financial services, they also trigger new risks. Our research suggests that network based explainable AI models can effectively advance the understanding of the determinants of financial risks and, specifically, of credit risks.

The same models can be applied to forecast the probability of default, which is critical for risk monitoring and prevention.

Future research should extend the proposed methodology to other datasets and, in particular, to imbalanced ones, for which the occurrence of defaults tends to be rare, even more than what observed for the analysed data. The presence of rare events may inflate the predictive accuracy of such events [as shown in Bracke et al. (2019)]. Indeed, Thomas and Crook (1997) suggests to deal with this problem via oversampling and it would be interesting to see what this implies in the proposed correlation network Shapley value context.

Acknowledgements This research has received funding from the European Union's Horizon 2020 research and innovation program "FIN-TECH: A Financial supervision and Technology compliance training programme" under the Grant Agreement No 825215 (Topic: ICT-35-2018, Type of action: CSA), and from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No.750961. Firamis acknowledges the NVIDIA Inception DACH program for the computational GPU resources. In addition, the Authors thank ModeFinance, a European ECAI, for the data; the partners of the FIN-TECH European project, for useful comments and discussions. The authors also thank the Guest Editor, and two anonymous referees, for the useful comments and suggestions.

Funding Open access funding provided by Università degli Studi di Pavia within the CRUI-CARE Agreement.

Compliance with Ethical Standards

Conflicts of interest Niklas Bussmann, Dimitri Marinelli and Jochen Papenbrock have been, or are, employed by the company FIRAMIS. The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The paper is the result of a close collaboration between all four authors. However, JP is the main reference for use case identification, method and process ideation and conception as well as fast and controllable implementation, whereas PG is the main reference for statistical modelling, literature benchmarking and paper writing.

References

Bracke, P., Datta, A., Jung, C., & Sen, S. (2019). Machine learning explainability in finance: an application to default risk analysis. Bank of England staff working paper no. 816.

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* (pp. 785–794). ACM.

Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical review E*, *70*(6), 066111.

Croxson, K., Bracke, P., & Jung, C. (2019). Explaining why the computer says ‘no’. FCA-Insight.

EU. (2016). Regulation (EU) 2016/679—general data protection regulation (GDPR). *Official Journal of the European Union*.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). *Annals of Statistics*, *28*(2), 337–407.

FSB. (2017). *Artificial intelligence and machine learning in financial services—market developments and financial stability implication*. Technical report, Financial Stability Board.

Giudici, P. (2018). Financial data science. *Statistics and Probability Letters*, *136*, 160–164.

Giudici, P., Hadji-Misheva, B., & Spelta, A. (2019). Network based credit risk models. *Quality Engineering*, *32*(2), 1–13.

Joseph, A. (2019). *Shapley regressions: a framework for statistical inference on machine learning models*. Research report 784, Bank of England.

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30* (pp. 4765–4774). Curran Associates, Inc.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., & Nair, B., et al. (2020). From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, *2*(1), 2522–5839.

Mantegna, R. N., & Stanley, H. E. (1999). *Introduction to econophysics: Correlations and complexity in finance*. Cambridge: Cambridge University Press.

Molnar, C. (2019). *Interpretable machine learning: A guide for making black box models explainable*.

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

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

Thomas, L., & Crook, J. (1997). Credit scoring and its applications. *SIAM Monographs*.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.