Explainable AI in fintech risk management

Niklas Bussmann

FIRAMIS and University of Pavia
Frankfurt, Germany and Pavia, Italy

Paolo Giudici

Fintech laboratory, University of Pavia, Italy

Dimitri Marinelli

FIRAMIS,
Frankfurt, Germany

Jochen Papenbrock

FIRAMIS,
Frankfurt, Germany

Abstract

The paper proposes an explainable AI model that can be used in fintech risk management and, in particular, in measuring the risks that arise when credit is borrowed employing peer to peer lending platforms. The model employs Shapley values, so that AI predictions are interpreted according to the underlying explanatory variables. The empirical analysis of 15,000 small and medium companies asking for peer to peer lending 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 and understand their credit score and, therefore, to predict their future behaviour.

Keywords: Credit risk management, Explainable AI, Financial Technologies, Peer to peer lending

Email addresses: nb@firamis.de (Niklas Bussmann), giudici@unipv.it (Paolo Giudici), dm@firamis.de (Dimitri Marinelli), jp@firamis.de (Jochen Papenbrock)

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.

To develop such models, we first need to understand what "Explainable" means. During this year, some important benchkmark definitions have been provided, at the institutional level. We report some of them, in the context of the European Union.

For example, the Bank of England (BoE, 2019) states that "Explainability means that an interested stakeholder can comprehend the main drivers of a model-driven decision". The Financial Stability Board [4] suggests that "lack of interpretability and auditability of AI and ML methods could become a macro-level risk". Finally, the UK Financial Conduct Authority [2] establishes that "In some cases, the law itself may dictate a degree of explainability."

The European GDPR [3] 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 related to 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 (XAI) models.

From a mathematical viewpoint, it is well known that, while "simpler" statistical learning models, such as linear and logistic regression models, provide a high interpretability but, possibly, a limited predictive accuracy, "more 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.

More precisely, our proposed methodology is based on Shapley values (see [8] and references therein). We consider a relevant application of AI in financial technology: peer to peer lending.

We employ Shapley values to predict the credit risk of a large sample of small and medium enterprises which apply for credit to a peer to peer lending platform. The obtained empirical evidence shows that, while improving the predictive accuracy with respect to a standard logistic regression model, we maintain and, possibly, improve, the interpretability (explainability) of the results.

In other words, our results confirm the validity of this approach in discriminating between defaulted and sound institutions, and it shows the power of explainable AI in both prediction accuracy and in the interpretation of the results.

The rest of the paper is organized as follows: Section 2 introduces the proposed methodology. Section 3 shows the results of the analysis in the credit risk context. Section 4 concludes.

2. Methodology

2.1. Credit risk in peer to peer lending

Credit risk models are useful tools for modelling and predicting individual firm default. Such models are usually grounded on regression techniques or machine learning approaches often employed for financial analysis and decision-making tasks.

Consider N firms having observation regarding T different variables (usually balance-sheet measures or financial ratios). For each institution n define a variable γ_n to indicate whether such institution has defaulted on its loans or not, i.e. $\gamma_n = 1$ if company defaults, $\gamma_n = 0$ otherwise. Credit risk models develop relationships between the explanatory variables embedded in T and the dependent variable γ.

The logistic regression model is one of the most widely used method for credit scoring. The model aims at classifying the dependent variable into two groups, characterized by different status (defaulted v.s. active) by the following model:

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

where p_n is the probability of default for institution n, $\mathbf{x}_i = (x_{i,1}, ..., x_{i,T})$ is the T-dimensional vector of borrower specific explanatory variables, the parameter α is the model intercept while β_t is the t-th regression coefficient. It follows that the probability of default can be found as:

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

2.2. Machine learning of credit risk

Credit risk can be measured with very different Machine Learning (ML) models, able to extract non-linear relations among the financial information 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 ([9]). We improve the choice selecting models based on their predictive accuracy, and employing a posteriori an explanations algorithm. This does not limit the choice of the best performing models.

To exemplify our approach we consider, without loss of generality, the XGBoost model, one of the most popular and fast algorithm [1], that implements gradient tree boosting learning models.

2.3. Learning model comparison

For evaluating the performance of each learning model, we employ, as a reference measure, the indicator $\gamma \in \{0, 1\}$, a binary variable which takes value one whenever the institutions has defaulted and value zero otherwise. For detecting default events represented in γ, we need a continuous measurement $p \in [0, 1]$ to be turned into a binary prediction B assuming value one if p exceeds a specified threshold $\tau \in [0, 1]$ and value zero otherwise. The correspondence between the prediction B and the ideal leading indicator γ can then be summarized in a so-called confusion matrix.

From the confusion matrix we can easy illustrate the performance capabilities of a binary classifier system. To this aim, we compute the receiver operating characteristic (ROC) curve and the corresponding area under the curve (AUC). The ROC curve plots the false positive rate (FPR) against the true positive rate (TPR), as follows:

$$FPR = \frac{FP}{FP + TN} \tag{3}$$

$$TPR = \frac{TP}{TP + FN} \tag{4}$$

The overall accuracy of each model can be computed as:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

and it characterizes the proportion of true results (both true positives and true negatives) among the total number of cases.

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 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 [7], 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 using the toolbox of econometrics, hypothesis testing, and network analysis.

We develop our Shapley approach using the SHAP[8] computational framework, which allows to express each single prediction as a sum of the contributions of the different explanatory variables.

More formally, the Shapley explanation model for each prediction $\phi(\hat{f}(x_i))$ is obtained by an additive feature attribution method, which decomposes them as:

$$\phi(\hat{f}(\mathbf{x}_i)) = \phi_0 + \sum_{k=1}^{M} \phi_k(x_i). \tag{6}$$

where M is the number of available explanatory variables, $\phi \in \mathbb{R}^{\mathbb{M}}$, $\phi_k \in \mathbb{R}$. The local functions $\phi_k(x_i)$ are called Shapley values.

Indeed, [8] prove that the only additive feature attribution method that satisfies the properties of *local accuracy*, *missingness* and *consistency* is obtained attributing to each feature $x_k, k = 1, \ldots, M$, a SHapley Additive exPlanation (SHAP) defined by

$$\phi_k(x_i) = \sum_{x^{'} \subseteq \mathcal{C}(x) \setminus x_k} \frac{|x^{'}|!(M - |x^{'}| - 1)!}{M!} [\hat{f}(x^{'} \cup x_k) - \hat{f}(x^{'})], \tag{7}$$

where $\mathcal{C}(x) \setminus x_k$ is the set of all the possible models excluding variable x_k (with $m = 1, \ldots, M$), $|x^{'}|$ denotes the number of variables included in model x', M is the number of the available variables, $\hat{f}(x^{'} \cup x_k)$ and $\hat{f}(x^{'})$ are the predictions associated with all the possible model configurations including variable x_k and excluding variable x_k, respectively.

The quantity $\hat{f}(x^{'} \cup x_k) - \hat{f}(x^{'})$ defines the contribution of variable x_k to each individual prediction.

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 [5] 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. Using this data, [6] and [5] 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. [1]). To explain the results from the model, typically highly predictive, we employ Shapley values.

The proportion of defaulted companies within this dataset is 10.9%.

3.2. Results

We first split the data in a training set (80%) and a test set (20%).

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

Figure 1 about here

From Table 2 note that the XGBoost clearly improves predictive accuracy. Indeed the calculation of the AUROC of the two curves indicate an increase from 0.81 (best logistic regression model) to 0.93 (best XGBoost model).

We then calculate the Shapley values for the companies in the test set.

To exemplify our results, Figure 2 we provide the interpretation of the estimated credit scoring of four companies: two that default and two that do not default.

Figure 2 about here

Figure 2 clearly shows the advantage of our explainable model. It can indicate which variables contribute more to the prediction. Not only in general, as is typically done by feature selection models, but differently and specifically for each company in the test set. Note how the explanations are rather different ("personalised") for each of the four considered companies.

4. Conclusions

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 Shapley values. This allows to explain any single prediction in terms of the potential contribution of each explanatory variable.

Future research should include a better understanding of the predictions through clustering of the Shapley values. This can be achieved, for example, using correlation network models. A second direction would be to extend the approach developing model selection procedures based on Shapley values, which would require appropriate statistical testing. A last extension would be to develop a Shapley like measure that applies also to ordinal reponse variables.

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, and among them is the lack of model interpretability. Our empirical findings suggest that explainable AI models can effectively advance our understanding and interpretation of credit risks in peer to peer lending.

Future research may involve further experimentation and the application to other case studies.

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 N.750961. 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.

References

[1] Chen, T. and C. Guestrin (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.

[2] Croxson, K., P. Bracke, and C. Jung (2019, May). Explaining why the computer says ‘no’. *FCA - Insight*.

[3] EU (2016). Regulation (eu) 2016/679 - general data protection regulation (gdpr). *Official Journal of the European Union*.

[4] FSB (2017, November). Artificial intelligence and machine learning in financial services - market developments and financial stability implication. Technical report, Financial Stability Board.

[5] Giudici, P., B. Hadji-Misheva, and A. Spelta (2019a). Correlation network models to improve p2p credit risk management. *Artificial Intelligence in Finance*.

[6] Giudici, P., B. Hadji-Misheva, and A. Spelta (2019b). Network based credit risk models. *Quality Engineering*, 1–13.

[7] Joseph, A. (2019, March). Shapley regressions: a framework for statistical inference on machine learning models. resreport 784, Bank of England.

[8] Lundberg, S. and S.-I. Lee (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30*, pp. 4765–4774. Curran Associates, Inc.

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

Figures

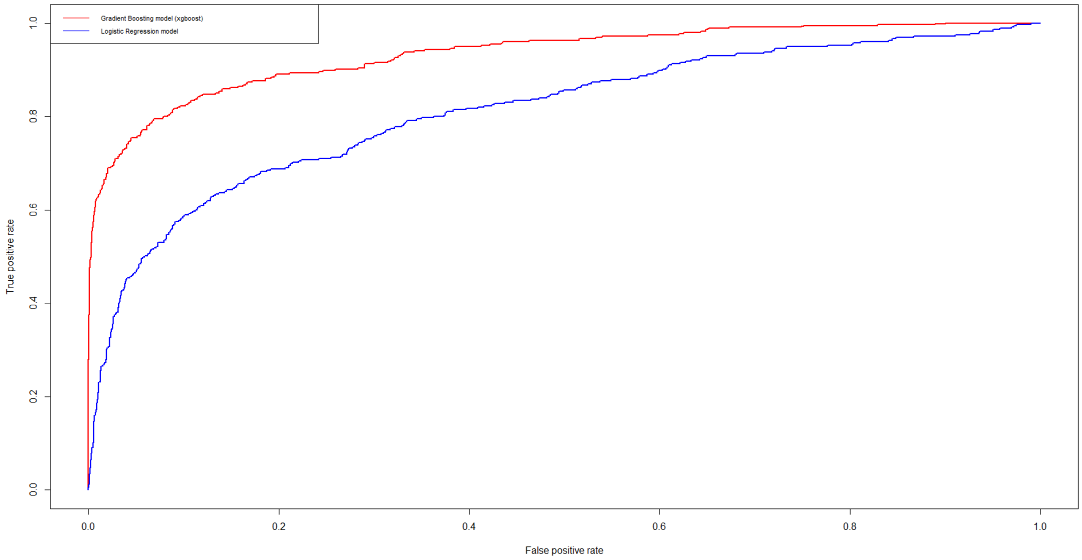

Figure 1: Receiver Operating Characteristic (ROC) curves for the logistic credit risk model and for the XGBoost model. In blue, we show the results related to the logistic models while in red we show the results related to the XGBoost model.

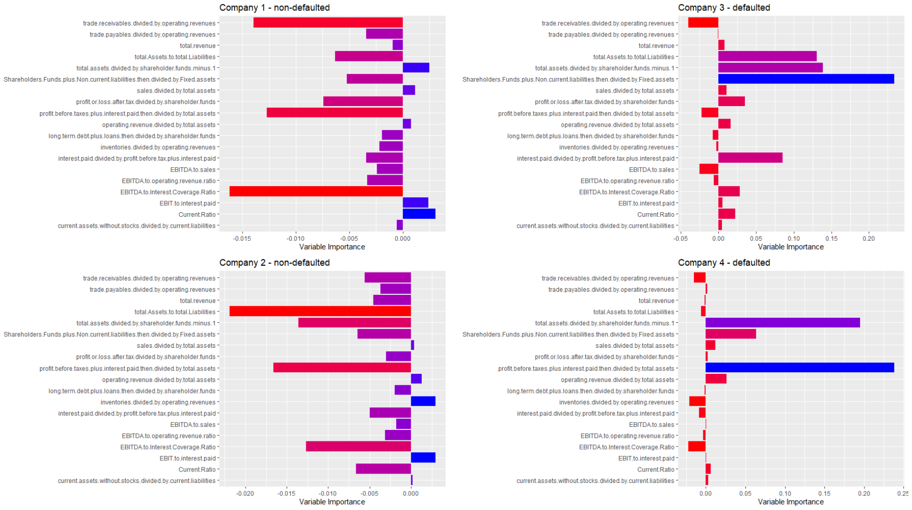

Figure 2: Contribution of each explanatory variable to the Shapley's decomposition of four predicted default probabilities, for two defaulted and two non defaulted companies. A red color indicates a low variable importance, and a blue color a high variable importance.