Evaluation of the paper “Explainable AI in credit risk management”

Table of Contents

A. Introduction

1 V29 Legal is a boutique law firm specializing in international dispute resolution as well as legal matters related to modern technologies.[1] With regards to artificial intelligence (AI), V29 Legal works with innovators to drive change and shape the development and deployment of technology which is and will continue changing our life tremendously.

2 Firamis GmbH (***Firamis***), is a B2B FinTech company that is funded by the EU Research and Innovation programme Horizon 2020.[2]

3 This paper is a brief evaluation of the paper "Explainable AI in credit risk management"[3] (Paper), which has been co-authored by Niklas Bussmann (Firamis and University of Pavia), Paolo Giudici (Fintech laboratory, University of Pavia), Dimitri Marinelli (Firamis), Jochen Papenbrock (Firamis) within the EU Horizon2020 FIN-TECH project[4], a program for financial supervision and technology compliance training.

B. Findings

4 Artificial Intelligence (***AI***) plays a vital role in many industries. While deployment of AI brings about tremendous potential, it also bears great risks. In order to address potential risks and to increase transparency and acceptance of this technology, various initiatives as well as legislative proposals, *inter alia*, calls for more transparency, explainability and accountability of AI have been introduced. Explainable AI (***XAI***) models can partially address these concerns, providing its user with a possibility to fulfil existing or future legal requirements.

5 The Paper introduces an agnostic method aiming at identifying the decision-making criteria of an AI system and variable importance (***Model***). In addition, the Model enables a visualisation of the results on a dashboard. While the Model can be applied to various AI systems, the Paper introduced a use case for credit risk assessment when loan decisions are made based on the credit scoring platforms. The Model identifies the logic behind a credit risk assessment with a possibility to reveal the decision-making criteria, so that the actual score can be explained, understood and, accordingly, a future forecast for lenders behaviour can be made.

6 Based on our analysis, the model can be of great assistance, when minimizing the risks related to opacity of AI systems. Firstly, it enables the possibility to monitor the decision-making process to ensure accountability and traceability. Secondly, it helps to expose potentially biased or erroneous decisions. Thirdly, the individuals affected by the decision

[1] V29 Legal – Duve Hamama Rechtsanwälte PartG mbB, https://www.v29-legal.com/.

[2] Firamis GmbH, https://firamis.de; *Horizon 2020, Research and Innovation programme*, https://ec.europa.eu/programmes/horizon2020/en.

[3] Bussmann, Niklas/Giudici, Paolo/Marinelli, Dimitri/Papenbrock, Jochen, *Explainable AI in Credit Risk Management*, December 18, 2019, https://ssrn.com/abstract=3506274.

[4] EU Horizon 2020 FIN-Tech project, https://firamis.de/ai-fintech-riskmanagement-regulation/#the-eu-horizon2020-fin-tech-project.

would have the opportunity to not only understand, but also to contest and take further action.

7 In the following, we summarize our understanding of the factual matrix of the Paper (**I.**) and follow up with an overview of questions which we have asked ourselves in order to conduct evaluation and believe to be of interest for potential users of the Model (**II.**). The answers which have been kindly provided by Firamis are integrated into this section.

C. Factual Summary

8 The Paper proposes an agnostic method that can be used in credit risk management. Its purpose is to measure the risks of default that arise when issuing a credit by using credit scoring platforms.

9 The Model has been tested with a dataset of publicly available financial information from 2015 and 2016 (balance-sheets) of over 15.000 SMEs from the south of Europe.

10 The data was supplied by the European External Credit Assessment Institution (***ECAI***) that specialises in credit scoring for P2P platforms with a focus on SMEs and was split in a training (80%) and a test set (20%). Assuming that each of the SMEs is a borrower and by comparing their financial status in 2015 with the following year, the analysis revealed that the borrowers can be divided into two risk groups - risky and not risky. Furthermore, by applying correlation network models to Shapley values, a predictive output of a machine learning model can be interpreted.

D. Questions for the Evaluation of the Model

I. Who is the target group of the proposed Model?

11 *Based on the mentioned dataset of SMEs, that play the role of borrowers, one could assume that the proposed Model was developed for lenders, such as banks and other financial institutions. However, it may also be relevant for the so-called credit scoring platforms or any other participant that is partaking in credit business.*

12 **Firamis:** Functions for risk management, assessment and scoring of credit portfolios in traditional banks as well as in 'fintech' platforms for p2p lending/crowdfunding.

II. What problem is the Model trying to solve?

13 *The proposed Model is providing an opportunity to interpret the output of Machine Learning models. It could be useful to precisely identify a specific problem which one can address by implementation of the Model for a credit risk assessment.*

14 **Firamis**: Here are some use case scenarios:

- Most obvious use case: a loan application is rejected. As a result, the bank has to provide the main reasons for this decision. What are the most important features leading to the rejection? This should be in line with GDPR requirements.
- Groups or clusters of data points express similar decision-making of the machine learning model. These clusters summarise the mechanics of the machine learning model and they represent types of decision-making of the model. Users get a better understanding of what the model has learned in order to verify it and for checking the plausibility.
- Data points at the intersect between clusters point to fuzzy decision-making which can be further investigated.
- A cluster with almost equal amount of predictions for default and non-default could point to bugs or problems in the machine-learning model. It could be checked if those data points are really indifferent or if the model exhibits a buggy decision making in this region of the model.
- Customer segmentation: the data points could not only be clustered by their input variables (representing clusters of similarity of the customers) but also by their variable contributions in the decision-making. This new clustering incorporates the 'intelligence' of the machine learning model (the mapping of input variables to the default labels). Customers in the same cluster exhibit similar decision making, e.g. how and whether they default or not.

III. Why is Black Box AI not suitable in regulated financial services?

15 *The paper draws attention to the problem of a Black Box AI, which does not provide a possibility to identify input and decision-making criteria of such systems. The Model could be useful to identify specific problems resulting from a Black Box AI usage in the financial sector.*

16 **Firamis**: Let me answer with two quotations standing for many similar statements by other national und supra-national supervisors/banks:

- 'It is the responsibility of supervised institutions to ensure the explainability and traceability of BDAI-based decisions. At least some insight can be gained into how models work and the reasons behind decisions, even in the case of highly complex models, and there is no need to categorise models as black boxes. For this reason, supervisory authorities will not accept any models presented as an unexplainable black box. Due to the complexity of the applications, it should be considered whether process results, in addition to documentation requirements, should also be examined in the future.'[5]

- 'It is often difficult to know (i) how reliable the inferred relationship between input and output is and (ii) which causality exists between them. This is called the explanatory gap of AI. [...] Supervisors have to adjust their approaches and skills to escort the introduction of AI/ML in banking. Banks have to give supervisors sound explanations of what their AI/ML systems actually do, as well as to what end.'[6]

IV. What is the exact result of the Model?

17 *The paper concludes, that the proposed Model provides the opportunity to "find and visualise hidden relationships like segmentations in diverse resolutions, trends, anomalies, hot spots, emergent effects and tipping points". It could be useful to depict a specific example for a positive outcome resulting from the application of the Model in credit risk assessment.*

18 **Firamis**: Let me introduce the idea, the model and the application a bit more details:

19 The machine-learning-based and visual procedure proposed in the paper processes the outcomes of another arbitrary machine learning model. It provides more insight, control and transparency to a trained, potentially black box machine learning model.

20 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) with applications in credit risk assessment and management as well as in other (financial) areas.

21 A key concept is the Shapley value decomposition of a model, a pay-off concept from cooperative game theory. To the best of our knowledge it is so far the only XAI (explainable AI) approach rooted in an economic foundation.

22 The approach offers a breakdown of variable contributions to the forecast probability. That means 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

[5] Bartels, Jörn/Deckers, Thomas, *Big data meets artificial intelligence – results of the consultation on BaFin's report*, March 21, 2019, https://www.bafin.de/SharedDocs/Veroeffentlichungen/EN/BaFinPerspektiven/2019_01/bp_19-1_Beitrag_SR3_en.html;jsessionid=DDFA1EBC7153FD5A23B710F25C5DAFA0.1_cid393.

[6] On the 'explanatory gap of AI': Wuermeling, Joachmi, *Is AI growing in the financial sector? – How can AI change change banking and what will this mean for supervision?*, Views Magazine, April 2020, https://www.eurofi.net/wp-content/uploads/2020/04/views-the-eurofi-magazine_zagreb_april-2020.pdf, p. 158.

learning model) but also by variable contributions to the prediction of the trained machine learning model.

23 The new contribution of the paper:

24 Similar combinations of variable contributions are mapped into spatial neighbourhoods on a map. This means that the data points are arranged on a map such that neighbouring data points exhibit similar decision-making mechanisms of the trained machine learning model.

25 The similarity/proximity of variable contributions (the paper uses an Euclidean Distance matrix) is expressed as a symmetric matrix of dimension **nxn** 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 their profile or combination of variable contributions.

26 This matrix can be used for visual mapping based on dimensionality reduction techniques (like PCA, MDS, t-SNE), or for representation learning like clustering and graph-analytics (like community detection).

27 Those data-driven, learned representations reveal segmentations of data points (customers) where each of those clusters contains very similar decision making whereas data points in other clusters exhibit very different decision-making.

28 (Hierarchical) Clustering and especially graph theory and network analytics are very well suited to study complex systems. Those systems are characterized by emergent, self-organizing properties. Our approach treats the variable contribution outcome of a (black box) machine learning model as a complex system and further analyses its properties by graph theory and cluster analysis. This way the user gets a better and deeper understanding of what exactly a black box machine learning has learned. The following phenomena inside the black box model can be analysed and understood: trends, anomalies, hot spots, emergent effects and tipping points. Since our methodology is model agnostic it can be applied to any machine learning model. This also enables a comparison of several machine learning models trained on the same data. The complex system of decision-making mechanisms that belong to a series of machine learning models can be compared to each other.

29 XAI and Graph Analytics are some of the most trending approaches, currently becoming very relevant in the financial service industry, both for regulatory and economic reasons. This is underlined by the conclusion of Gartner that graph analytics and XAI will be some of the most trending technologies in the next years.[7]

[7] Moore, Susan, *Gartner Top 10 Data and Analytics Trends*, Smarter with Gartner, November 5, 2019, https://www.gartner.com/smarterwithgartner/gartner-top-10-data-analytics-trends/.

30 Paper uses TreeSHAP, a consistent and accurate method, available in open-source packages. Tree SHAP is a fast algorithm that can exactly compute SHAP values for trees in polynomial time instead of the classical exponential runtime.[8]

31 Our Paper does not follow a causal approach. However, it brings transparency to the statistical inference of a potential black box machine learning in mapping and learning representations of the model's decision-making mechanisms which in turn are based on the variable contributions produced by the black box model.

32 For the xgboost part of our model we use NVIDIA GPUs to considerably speed up the computations. The TreeSHAP method can quickly extract the information from the xgboost model.

33 The paper is based on the SHAP concept:

Lundberg, Scott/Lee, Sung-In, *A unified approach to interpreting model predictions Advances in Neural Information Processing Systems*, May 22, 2017, https://arxiv.org/abs/1705.07874, Volume 30, p. 4765 - 4774.

34 The paper Is related to:

Bracke, Philippe/Datta, Anupam et al., *Staff Working Paper No. 816: Machine learning explainability in finance: an application to default risk analysis*, August 9, 2019, https://www.bankofengland.co.uk/working-paper/2019/machine-learning-explainability-in-finance-an-application-to-default-risk-analysis.

Joseph, Andreas, *Staff Working Paper No. 784: Shapley regressions: a framework for statistical inference on machine learning models*, March 8, 2019, https://www.bankofengland.co.uk/working-paper/2019/shapley-regressions-a-framework-for-statistical-inference-on-machine-learning-models.

35 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. Firamis acknowledges the NVIDIA Inception DACH program for the computational GPU resources.

V. Does the Model help to implement or comply with existing or upcoming regulations?

36 The paper links XAI to three out of seven key requirements of the Ethics Guidelines for Trustworthy AI (***Guidelines***): Human agency and oversight, Transparency and Accountability.

37 Following this approach and given that the proposed Model is a “tool” to identify and visualize the decision-making process of an AI system, it raises the question whether, and

[8] Lundberg, Scott/Erion, Gabriel/Lee, Sun-In, *Consistent Individualized Feature Attribution for Tree Ensembles*, February 12, 2018, https://arxiv.org/abs/1802.03888.

if so, for compliance with which of the 7 key requirements provided in the Guidelines the Model could be helpful to comply with upcoming EU framework:

- Human agency and oversight (**1.**)
- Technical robustness and safety (**2.**)
- Privacy and data governance (**3.**)
- Transparency (**4.**)
- Diversity, non-discrimination and fairness (**5.**)
- Societal and environmental wellbeing (**6.**)
- Accountability (**7.**)

1. Does the Model help to ensure human agency and oversight?

38 *Besides of considering whether the AI system has an impact on fundamental rights, AI practitioners have to ensure an appropriate level of human control for the particular system and use case. It could be useful to show to what extent the proposed Model helps to ensure human oversight and/or control within credit risk management.*

39 **Firamis**: Decisions must be informed, and there must be a human-in-the-loop oversight. The SHAP approach enables the user to understand why the decision was made. The 'why' is not causal but expressed as numeric contributions of input variables. The user can look at a specific data point and see the input variables, the variables contributions to the prediction as well as the prediction itself. A more human-based plausible explanation can arise to reconcile the machine-based decision with a human narrative 'that makes sense'. The model can be better controlled as it delivers a feedback how it comes to all decisions both on global level (global variable importance) and well as local levels (data points). The clustering step even delivers the variable contributions for the members of that specific cluster, so for a group of customers. The user could identify properties of this group of customers based on the input variables in order to understand how the decision-making works for this group of customers. All these analytics capabilities and tools plus the interactive visual exploration enable the user to much better understand the outcome of an entirely black box model. Better understanding leads to more effective control.

2. Does the Model help to ensure technical robustness and safety?

40 *According to the Guidelines, technical robustness requires that AI systems be developed with a preventative approach to risks and in a manner such that they reliably behave as intended while minimising unintentional and unexpected harm and preventing unacceptable harm. This entails the protection of human dignity as well as mental and physical integrity. Assuming that the proposed Model helps to ensure safety and robustness, it could be useful to depict a specific example in credit risk assessment.*

41 **Firamis**: Our approach can NOT directly prevent risk, harm and protect human dignity as well as mental and physical integrity. However, a reliable credit

assessment should minimise errors of first and second order: no customer should be rejected although a 'good' customers and 'bad' customers should be locked out. Explainability helps to debug the model, to find errors in the model. In this context the model could become more robust regarding wrong decision-making.

3. Does the Model help to ensure privacy and data governance?

42 *The Guidelines stipulate that AI systems must guarantee privacy and data protection throughout a system's entire lifecycle.*

43 **Firamis**: Our approach only processes the data given so this point seems to be more a data-related issues.

4. Does the Model help to ensure transparency?

44 In order to comply with the requirement of transparency, an AI system needs to encompass traceability (**a.**), explainability (**b.**) and communication (**c.**).

(a) Does the Model help to ensure traceability?

45 *To ensure traceability, a mechanism of documentation to the best possible standard should be incorporated. Be it, for example, the documentation of data sets, data labelling or the decisions made by the AI system.*

46 **Firamis:** the SHAP approach allows to trace back and document the variable contributions to decision making. The clustering of SHAP information is one of the new pieces of information added by our approach so this can be used to enrich the traceability and documentation. Also, the steps to improve the model based on the new information could be documented.

(b) Does the Model help to ensure explainability?

47 *According to the Guidelines, explainability concerns the ability to explain both the <u>technical processes</u> of an AI system and the <u>related human decisions</u>. The explanation must be understood by <u>all</u> users.*

48 **Firamis:** What is explainability? Every human being will come up with an own interpretation when a model is fully explained. Irrespective of this fact we try to come up with some definitions:

- The mode is explained when you can't ask 'why' any more
- ability to explain AI-made decisions

- make its functioning clear or easy to understand.
- XAI 'proves the work'. It is to trust AI and to accelerate adoption.
- ensuring compliance with expanding regulatory and public expectations and
- in fostering trust."

49 It is true that the SHAP approach delivers an explanation of global and local decision making of a black box machine learning model, so it explains the AI model on all levels. Our extension allows to further analyse the explanations, for example in context to the other local decision mechanisms and data points. Thus, a richer set of information about the decision-making mechanism is given. This could lead to a situation where a human narrative arises and a 'story that is plausible' can be delivered. In this way the machine decision making is connected to human decision making.

50 However, as mentioned before, it's not an explanation in terms of a 'causal why'.

51 An important aspect of explainability is to clarify the audience. The audience of a model explanation is manifold: model builder, model checker, compliance and governance officer, risk manager, product owner, senior manager, supervisors, customer. The SHAP information can be understood by the data science people and most other people in a bank or fintech company would understand it with training. The same applies to the supervisors. For customers/and clients it may be sufficient to mention which variables are most important (the client should probably be informed about the reason for a decision/rejection according to GDPR) or what a client could do to improve certain variables to get a positive decision.

52 The SHAP information delivers a consistent and accurate view and language to describe an AI model. The next step would be standardisation and generally accepted definitions of explainability.

(c) Does the Model help to ensure communication?

53 *AI systems should not represent themselves as humans to users; humans have the right to be informed that they are interacting with an AI system. This entails that AI systems must be identifiable as such.*

54 **Firamis:** not really. Our model only delivers explanations. Those who are responsible to run an AI system have to make clear which information is human and which is machine-based.

5. Does the Model help to ensure diversity, non-discrimination and fairness?

55 *To achieve Trustworthy AI, unfair bias must be avoided, accessibility and universal design ensured, and stakeholder participation included.*

56 *According to the Guidelines, Data sets used by AI systems (both for training and operation) may suffer from the inclusion of inadvertent historic bias, incompleteness and bad governance models. It could be useful to precisely identify a specific problem that may occur in credit risk management, which can then be addressed through implementation of the Model.*

57 **Firamis:** The SHAP information and our cluster/network extensions can help to address these issues. For example, there could be a combination of variable contributions that basically replaces a discriminatory variable that had been excluded.

6. Does the Model help to ensure societal and environmental wellbeing?

58 *The Guidelines encourage sustainability and ecological responsibility of AI systems.*

59 **Firamis:** by running an AI model in more controlled, robust, traceable, accurate and explainable way, the model should be more sustainable. The credit and lending business should profit from those XAI-based human-centric models and become more trustworthy. Both consumers and the banking and fintech industry should profit from a more honest approach in the long run.

7. Does the Model help to ensure accountability?

60 *According to the Guidelines, mechanisms must be put in place to ensure responsibility and accountability for AI systems and their outcomes, both before and after their development, deployment and use. It could be useful to depict a specific example on how the proposed Model helps to ensure accountability in credit risk management.*

61 **Firamis**: As outlined in '(a) Does the Model help to ensure human agency and oversight' a XAI-driven model is more accountable, and responsibility can be taken for such a non-black box model. It also reduced the risk of failure and loss of reputation for a financial company running such a model.

62 However, from a risk management perspective, an AI model (using the XAI approach or not) should be treated as a very complex model where adequate risk management and governance need to be in place. Our XAI approach can just be a contribution to that. However, a system that can be measured and explained can always be better risk managed.

63 Also, our approach is just a single building block towards many other measures and tool to create trustworthy, responsible and human-centric AI.