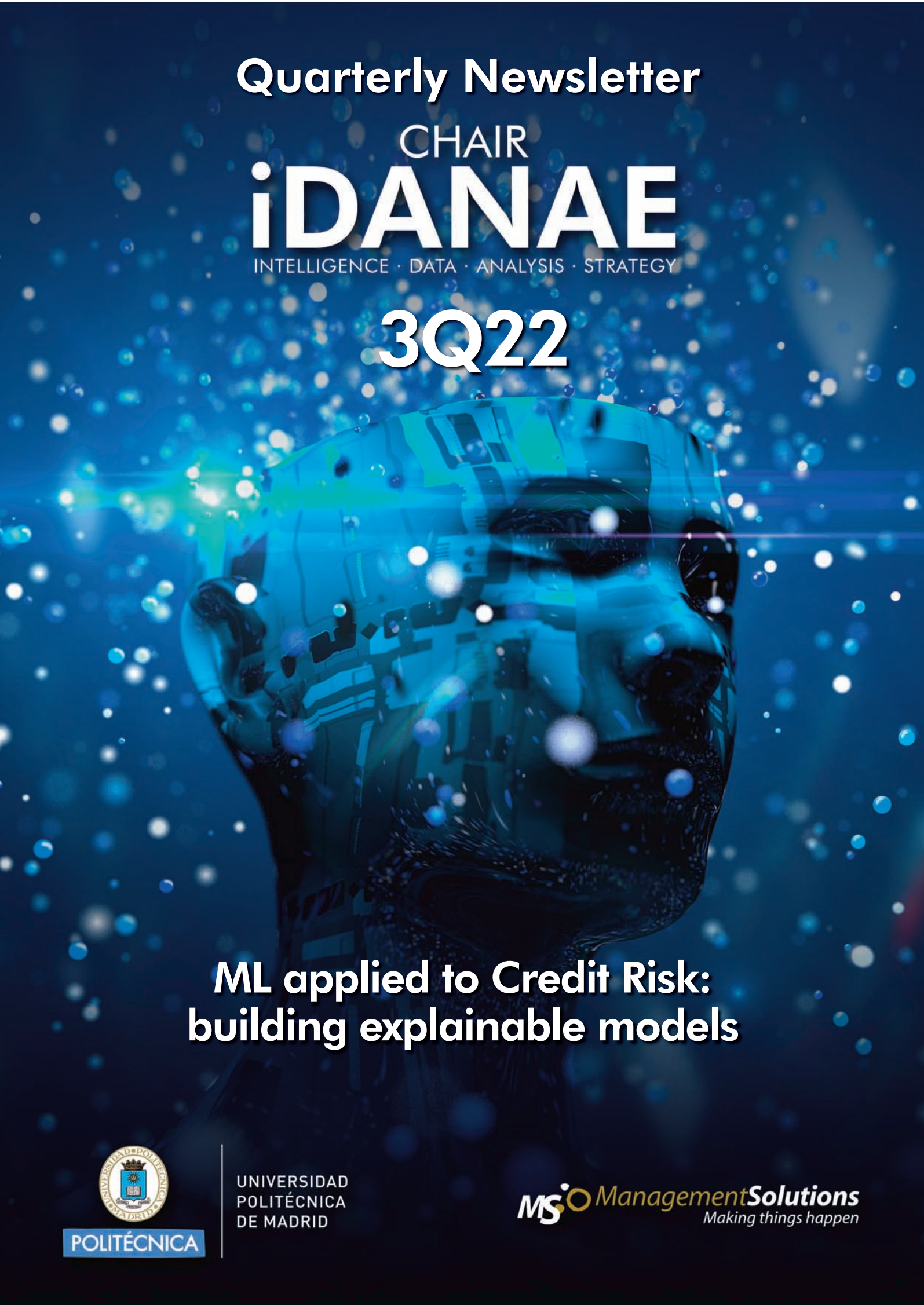
Quarterly Newsletter

CHAIR

iDANAE

INTELLIGENCE · DATA · ANALYSIS · STRATEGY

3Q22



ML applied to Credit Risk: building explainable models

POLITÉCNICA

UNIVERSIDAD POLITÉCNICA DE MADRID

ManagementSolutions
Making things happen

Introduction

In 2021 the EBA has issued a discussion paper on the use of machine learning (ML) models to calculate capital requirement for banking institutions [1]. In that publication, the EBA provides a consistent understanding of how new sophisticated ML models might coexist with prudential requirements, identify the main challenges and possible benefits of ML models, as well as provide a set of principle-based recommendations in the context of IRB models. The paper discusses four main aspects:

1. An exposition of different learning paradigms (supervised, unsupervised, and reinforcement learning) that can be used to train ML models. The use of each paradigm depends on the goal of the model and the data required.

2. A set of current practices in Credit Risk Modelling. For IRB models the use of ML techniques is limited, and said techniques are generally applied only as a complement to a standard model when used for regulatory purposes.

3. Some challenges and potential benefits. Depending on the context of their use, the complexity and interpretability of some ML models might pose additional challenges for the institutions to develop compliant IRB models. Indeed, the need for explainability is one of the main aspects that prevent institutions from fully applying ML techniques in the regulatory models. This issue will be discussed later in this document.

4. The identification of four main pillars that need to be present to support the rollout of advanced analytics, and which should be properly and sufficiently addressed:
    1. Data management
    2. Technological infrastructure
    3. Organization and governance
    4. Analytics methodology

In addition, the EBA includes some specific recommendations on the appropriate knowledge of the models, model interpretability, low complexity for model use, and adequate model validation techniques.

Other challenges may also arise, for example: (1) the integration of qualitative judgment with the models, which will require new methodologies; (2) data availability[1]; (3) model documentation requirements; (4) analysts training to understand and use these models; or (5) avoiding possible biases or discrimination issues due to race, religion, sex, etc.[2], among others. These and more challenges have also been described in detail by the Bank of Spain in a dedicated publication [2].

Despite all this challenges, the interest in the implementation of modern ML techniques for regulatory capital raises from the

[1]For regulatory models there is a minimum of 5 to 7 years (depending on the segment and parameter) in at least one of the data sources. See [7].
[2]This is also a requirement from GDPR (article 22.4).

possibility of reducing the capital requirements: according to Bank of Spain, the adoption of ML for such purposes could reduce the capital requirements by 12.4% to 17% [3].

In addition to the EBA's publication (which is addressed to financial institutions), the European Commission has also issued a proposal for regulation on Artificial Intelligence (AI), which also aims to enhance the use of AI (see Box 1) [4].

All this context opens the question for financial institutions about whether and how to consider ML models for regulatory purposes, despite the challenge of complying with all the requirements already stablished for the modeling process.

Amongst all challenges mentioned, the need for explainability in the models is one that has received special attention from the regulators. For this reason, the implications of this challenge for financial institutions are analyzed, as well as different approaches on how to tackle it.

Box 1. Proposal for a regulation laying down harmonised rules on AI [4]

In April 2021, the European Comission published the "Proposal for a regulation laying down harmonised rules on AI" to set harmonized rules for the development, placement on the market, and use of AI systems.

The proposal has the following specific objectives:

- Ensure that AI systems placed on the EU market and used are safe and respect existing law on fundamental rights and EU values.
- Ensure legal certainty to facilitate investment and innovation in AI.
- Enhance governance and effective enforcement of existing law on fundamental rights and safety requirements applicable to AI systems.
- Facilitate the development of a single market for lawful, safe and trustworthy AI applications and prevent market fragmentation.

This document prohibits certain AI practices such as the use of AIs that deploys subliminal techniques beyond a person's consciousness that could cause any kind of harm, or the use of "real-time" remote biometric identification systems in public spaces for the purpose of law enforcement (with a few exceptions).

The document also identifies certain AIs as High-Risk AI systems, such as safety components in the management and operation of essential public infrastructure networks, or AI systems intended to be used by law enforcement authorities for predicting the occurrence or reoccurrence of an actual or potential criminal offence based on profiling. These must follow extra legal requirements, and their providers have certain additional obligations.

The explainability challenge

The models constructed using ML techniques need to be understood by the users, which is one of the main challenges in adopting ML models for capital requirements calculation. The General Data Protection Regulation (GDPR) states[3] that "The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her". In other words, users have the right to an explanation on why the model has made a specific decision [5].

It is not clear to what extent these explanations must be accurate, reliable, complete, or true to the original model. The regulator has not stated which methodologies are appropriate for this new paradigm, whereas the academic world has been coming up with new ideas in the recent years under the label of explainability, or xAI (explainable artificial intelligence).

Anyhow, the explainability of both the model and the results must be ensured for the adoption of ML models for capital requirements. Following the principle of explainability from the document "Ethics guidelines for trustworthy AI" by the European Commission: "processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions – to the extent possible – explainable to those directly and indirectly affected" [6]. In addition, according to the CRR [7], the senior management must understand the structure of the models within the modeling approval process. Therefore, institutions must be able to explain to their clients in an easy and intuitive way those elements that determine the decision of a credit model.

[3]Article 22.1 of GDPR

Approaches to explainability

Two main approaches may be identified to solve the explainability problem: (1) to use an interpretable model (hence explainable) and enhance it using ML techniques; (2) to fit a ML model and use post-hoc techniques to explain its predictions.

Using interpretable models

In order to address the first approach, it is important to discuss the following question: what is an interpretable model? Although there could be many different answers, an interpretable model could be understood as a model that anybody could memorize and/or understand the relations between input and output. In other words, the relationship between variables does not necessarily have to be linear, but the relation between input and output must be simple.

This definition depends on human cognitive limitations. For example, a linear model is usually considered interpretable because the relation between input and output is simple; however, a linear model with 10.000 input variables should not be directly considered interpretable just because it is linear. Similarly, a decision tree is easy to visualize, but a decision tree with 10.000 splits is not directly interpretable [8]. As an example, figure 1 is an example of an interpretable model, for heart attack prediction.

1.	*Congestive Heart Failure*	1 point		⋯
2.	*Hypertension*	1 point	+	⋯
3.	*Age* $\geq$ *75*	1 point	+	⋯
4.	*Diabetes Mellitus*	1 point	+	⋯
5.	*Prior Stroke or Transient Ischemic Attack*	2 points	+	⋯
	ADD POINTS FROM ROWS 1–5	**SCORE**	=	⋯

SCORE	0	1	2	3	4	5	6
STROKE RISK	1.9%	2.8%	4.0%	5.9%	8.5%	12.5%	18.2%

Figure 1: Score table of a heart attack prediction model [9].

This is a particular case of a scoring model, widely used in the legal and medical fields in the US [9]. It has few variables, all variables are relevant according to expert judgement (e.g. stroke risk should not depend on income or colour of the eyes), and the relations between different inputs and output is simple to understand. The example above is an interpretable model, which implies explainability (i.e. given a prediction, it is easy to understand why the model reached that particular risk value).

The list of what could be considered interpretable models is not very extensive: generalized linear models (GLM), generalized additive models (GAM), decision trees, and scoring models.

However, even when these models are used, ML techniques could be included to enhance interpretable models by using ML to feature engineer new variables to feed into the model, segmenting the data, or variable selection (see Box 2).

Interpretable models have the advantages of explainability, that can be easily complemented with expert criteria, but the disadvantage of intellectual property: if a model is as simple and interpretable as in the stroke risk example, it can be memorized and even reverse engineered. This can cause that the model could end up in hands of the industry competitors, which could use it to improve their own.

The main difficulty of this approach is the need for expert knowledge, both in business and in quantitative disciplines, that results in optimal data pre-processing (or feature engineering). This feature engineering could make the performance of a traditional model comparable to more modern ML models.

Using post-hoc explainability techniques

Many authors have shown the improvements of ML techniques compared to the traditional logit for credit risk [13] [14], which increases the AUC of the model up to a 20% (with respect to the logit). However, there are authors who maintain that this difference can be compensated with feature engineering and knowledge of the data [15]. A particularly interesting case is Chen et al., 2018 [16] who won a ML competition on credit risk with an interpretable model. This model could be summarized as a two-layer neural network with sigmoid activation function: in their first layer they group variables with expert criteria to form risk subscores (e.g. the variables "number of total trades" and "number of total trades in the last year" are grouped to form the "trade frequency" subscore), and in the second layer they combine all different subscores into one final score prediction.

Nevertheless, ML techniques seem to be generally better than humans at extracting information or patterns from data. Since these models are not interpretable, there is a need to explain each prediction the model performs for this technology to be used.

There is a particular example of a ML model which is not interpretable (according to the definition in previous section) but easily explainable: the K nearest neighbor algorithm (KNN). In this model each prediction is based on the history of the closest data to the input (with proper processing and distance definition). In other words, for a credit scoring, a client can be

denied credit because in the historical data there are clients with very similar characteristics that defaulted.

For more complex models, like any tree-based ensemble method, a set of techniques have been developed in academia to explain any given prediction. The most popular example of these technique is LIME [17]. This method creates, for a given prediction, an interpretable model. This is a local approach, that is, the explanation for one prediction does not have to coincide with another. The authors propose another technique to choose certain observations which provide a global view of the model (SL-LIME). This method can also be used to choose the best between two models or for feature engineering. Other techniques have been developed in this field, such as SHApley Additive exPlanation (SHAP) [18] or permutation variance importance (VI) [19].

Both LIME and SHAP were investigated in many papers[4]. One of them uses these methods to explain models based on Random Forest, XGBoost, Logistic Regression, SVM and NN classifiers for credit risk management purposes [20]. The research found these methods useful due to their ability to provide explanations that are in line with financial logic, such as lower loan amounts are associated with a lower probability of default. Furthermore, they found consistency in the top 20 most important features for the two different methods studied, even when expanding their test dataset. Other authors highlight drawbacks that lead to unstable and unreliable evaluations [21]. These disadvantages make the methods easy to deceive, which was proven using real word data from criminal justice and credit scoring domains. It was also found that LIME is more vulnerable to deception than SHAP [22]. Another set of research papers tested these algorithms on simulated data to compare result of methods with a data generation process. It was found that LIME explanation cannot be considered stable around each data point in a tested case [23]. Although the quantity of papers enlisting limitations of these methods is significant, there is ongoing research to improve them.

One of the proposed improvements was to relax one of the Shapley Values assumptions – symmetry, to create Asymmetric Shapley Values that incorporate causality into model explainability. Not only the method works better as it takes correlation into account, but it might also work for a selection of important features without model retraining [24]. Other improvement to the SHAP method is to not analyze distinct features but to group them based on feature knowledge and dependence. This method is called groupShapley. It is important to know that the user of the method should have knowledge about the data, otherwise the method might not present meaningful results [25].

Bank of Spain [26] proposed a method to evaluate such interpretability techniques. They propose the creation of synthetic datasets where the importance of the variables is controlled. Once a model has been adjusted to these data, it is verified that the results of the interpretability techniques are aligned with the assumptions made when creating the dataset. The problem is that these synthetic datasets do not have to cover the entire spectrum of possible real datasets, which are generated by unknown stochastic models. Similarly, ensuring that an interpretability method works correctly 99% of the time is equivalent to assuming that 1% of the implemented models may be defective in the best case scenario or even unfair , with the consequences that these will have for people and for the local economy.

Bank of Spain [3] also measured the impact of using advanced ML models against a logistic regression on the calculation of capital for credit risk. They concluded that entities could save between 12.4% and 17% by implementing an XGBoost compared to a logit. The dataset used was provided by a

[4]Including a previous publication of this Chair [27]

financial institution, anonymizing the variables to make them unrecognizable, which prevented any attempt at feature engineering by the authors.

Therefore, the efficiency of the most modern ML methods in comparison with classic and interpretable models is widely accepted, in the absence of feature engineering. However, in the case of pursuing to make use of these techniques, methodologies have to be developed to be able to trust the model and find explanations for its predictions. As a drawback, the difficulty of incorporating expert judgment with these models must be solved, as it is one of the regulatory requirements for the development of credit models[5].

[5] Artículo 174e en [7]

Box 2. Different alternatives for using ML techniques with interpretable models

One of the most common approaches of using machine learning for feature engineering is clustering. It is a way of segregating groups with similar traits and assigning them into clusters to create new features for the final model. There are multiple ways of encoding clustering (e.g., distance to the center of each cluster, cluster membership probability, etc.). Such a new feature with proper encoding can improve the quality of classification for some classifiers. It is worth mentioning that the literature suggests not to get rid of old features but instead add new ones based on clustering results [10]. Although there are several clustering techniques, they mostly do not take into consideration correlation with the target variable. Other way of segmentation uses decision trees to find bins of numerical features that involve a higher correlation with the target variable. Because this feature is not met when performing a random clustering, using tree-based binning may lead to improve the final performance of the model [11].

Segmenting the data is a good approach, but it is also important to use only relevant features. Using random forest, an algorithm that is based on multiple decision trees, has proven to be a good approach for identifying features with high importance. The task is done by calculating how much each feature decreases the impurity and averaging the result across multiple trees. Having found important features, they can be used in the model, and reduce its complexity [12].

Conclusions

The EBA has addressed the discussion on the use of ML models for capital requirements calculation or credit risk scores. However, there are many challenges that financial institutions need to address to fully integrate ML techniques in regulatory models. Among them, the explainability is one of the main challenges to solve.

The implementation of traditional models (such as logit or lasso) for the calculation of regulatory capital is recommended, as well as evaluate the use of ML models. There are some reflections that need to be considered, such as the fact that ML model predictions may be more accurate but may not always be applicable due to lack of explainability in the prediction (amongst other possible problems). In such cases it should be always possible to rely back on the traditional model.

The regulator has not offered yet specific guidelines on how to face this explainability challenge, although it has highlighted some possible problems to consider. In any case, institutions must provide an explanation for the predictions of some models.

Whether an interpretable model with featuring engineering is desired, or a non-interpretable model together with an explainability tool, obtaining and training personnel for these tasks will continue to be a need in the industry.

Bibliography

[1] EBA, "Discussion paper on machine learning for IRB models," 2021.

[2] Banco de España, "Machine learning in credit risk: measuring the dilema between prediction and supervisory cost," 2020.

[3] Banco de España, "Understanding the performance of machine learning models to predict credit default: novel approach for supervisory evaluation," 2021.

[4] European Commission, "Laying down harmonised rules on artificial intelligence," 2021.

[5] B. Goodman and S. Flaxman, "European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"," AI Magazine, pp. 50-57, 2017.

[6] European Comission, "Ethics guidelines for trustworthy AI," 2019.

[7] European Commision and the Parliament, Capital Requirements Regulation, 2013.

[8] B. F. Gage, A. D. Waterman, W. Shannon, M. Boechler, M. W. Rich and M. J. Radford, "Validation of Clinical Classification Schemes for Predicting Stroke," JAMA, vol. 285, no. 22, p. 2864, 2001.

[9] B. Ustun and C. Rudin, "Learning optimized risk scores," Journal of Machine Learning Research, vol. 20, 2019.

[10] M. Piernik and T. Morzy, "A study on using data clustering for feature extraction to improve the quality of classification," Knowledge and Information Systems, vol. 63, pp. 1771-1805, 2021.

[11] S. Kumar, "Essential guide to perform Feature Binning using a Decision Tree Model," 2 September 2021. [Online]. Available: https://towardsdatascience.com/essential-guide-to-perform-feature-binning-using-a-decision-tree-model-90bcc66d61f9. [Accessed 1 August 2022].

[12] A. Dubey, "Feature Selection Using Random forest," 18 December 2018. [Online]. Available: https://towardsdatascience.com/feature-selection-using-random-forest-26d7b747597f. [Accessed 1 August 2022].

[13] J. Sirignano and R. Cont, "Universal Features of Price Formation in Financial Markets: Perspectives From Deep Learning," SSRN Electronic Journal, 2018.

[14] F. Sigrist and C. Hirnschall, "Grabit: Gradient tree-boosted Tobit models for default prediction," Journal of Banking & Finance, vol. 102, pp. 177-192, 2019.

[15] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, pp. 206-2015, 2019.

[16] C. Chen, K. Lin, C. Rudin, Y. Shaposhnik, S. Wang and T. Wang, "An Interpretable Model with Globally Consistent Explanations for Credit Risk," in Proceedings of NeurIPS 2018 Workshop on Challenges and Opportunities for AI inFinancial Services: the Impact of Fairness, Explainability, Accuracy, and Privacy, 2018.

[17] M. Ribeiro, S. Singh and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, 2016.

[18] S. M. Lundberg, P. G. Allen and S.-I. Lee, "Advances in Neural Information Processing Systems," 2017.

[19] A. Fisher, C. Rudin and F. Dominici, "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously," Journal of Machine Learning Research, pp. 1-81, 2019.

[20] B. H. Misheva, A. Hirsa, J. Osterrieder, O. Kulkarni and S. Fung Lin, "Explainable AI in Credit Risk Management," SSRN Electronic Journal, 2021.

[21] A. Gosiewska and P. Biecek, "Do Not Trust Additive Explanations," https://doi.org/10.48550/arXiv.1903.11420, 2019.

[22] D. Slack, S. Hilgard, E. Jia, S. Singh and H. Lakkaraju, "Fooling LIME and SHAP," in AIES '20: AAAI/ACM Conference on AI, Ethics, and Society, New York, 2020.

[23] Y. Zhang, K. Song, Y. Sun, S. Tan and M. Udell, ""Why Should You Trust My Explanation?" Understanding Uncertainty in LIME Explanations," https://doi.org/10.48550/arXiv.1904.12991, 2019.

[24] C. Frye, C. Rowat and C. Rowat, "Asymmetric shapley values: Incorporating causal knowledge into model-agnostic explainability," in Advances in Neural Information Processing Systems, 2020.

[25] M. Jullum, A. Redelmeier and K. Aas, "Efficient and simple prediction explanations with groupShapley: A practical perspective," in CEUR Workshop Proceedings, 2021.

[26] Banco de España, "Accuracy of explanations of machine learning models for credit decisions," 2022.

[27] iDanae Chair, "Interpretabilidad de los modelos de inteligencia artificial," 2019.

Autores

Ernestina Menasalvas (UPM)

Manuel Ángel Guzmán (Management Solutions)

Sergio Ruiz Bonilla (Management Solutions)

UNIVERSIDAD POLITÉCNICA DE MADRID

Management Solutions
Making things happen

The Universidad Politécnica de Madrid is a multi-sector and multi-disciplinary Public Law Entity, which carries out teaching, research and scientific and technological development activities.

www.upm.es

Management Solutions is an international consulting firm, focused on business, finance, risk, organization, technology and process consulting, operating in more than 40 countries and with a team of over 3,000 professionals working for more than 1,500 clients worldwide.

www.managementsolutions.com

For more information visit

blogs.upm.es/catedra-idanae/