ACCURACY OF EXPLANATIONS OF MACHINE LEARNING MODELS FOR CREDIT DECISIONS

2022

BANCO DE ESPAÑA
Eurosistema

Documentos de Trabajo
N.º 2222

Andrés Alonso and José Manuel Carbó

ACCURACY OF EXPLANATIONS OF MACHINE LEARNING MODELS FOR CREDIT DECISIONS

ACCURACY OF EXPLANATIONS OF MACHINE LEARNING MODELS FOR CREDIT DECISIONS [*]

Andrés Alonso

BANCO DE ESPAÑA

José Manuel Carbó

BANCO DE ESPAÑA

[*] The authors work as senior economists in the Financial Innovation Division of Banco de España, and appreciate the comments received from José Manuel Marqués, Ana Fernández and Sergio Gorjón.

Documentos de Trabajo. N.º 2222

June 2022

ISSN: 1579-8666 (on line)

Abstract

One of the biggest challenges for the application of machine learning (ML) models in finance is how to explain their results. In recent years, different interpretability techniques have appeared to assist in this task, although their usefulness is still a matter of debate. In this article we contribute to the debate by creating a framework to assess the accuracy of these interpretability techniques. We start from the generation of synthetic data sets, following an approach that allows us to control the importance of each explanatory variable (feature) in our target variable. By defining the importance of features ourselves, we can then calculate to what extent the explanations given by the interpretability techniques match the underlying truth. Therefore, if in our synthetic dataset we define a feature as relevant to the target variable, the interpretability technique should also identify it as a relevant feature. We run an empirical example in which we generate synthetic datasets intended to resemble underwriting and credit rating datasets, where the target variable is a binary variable representing applicant default. We then use non-interpretable ML models, such as deep learning, to predict default, and then explain their results using two popular interpretability techniques, SHAP and permutation Feature Importance (FI). Our results using the proposed framework suggest that SHAP is better at interpreting relevant features as such, although the results may vary significantly depending on the characteristics of the dataset and the ML model used. We conclude that generating synthetic datasets shows potential as a useful approach for supervisors and practitioners looking for solutions to assess the interpretability tools available for ML models in the financial sector.

Keywords: synthetic datasets, artificial intelligence, interpretability, machine learning, credit assessment.

JEL classification: C55, C63, G17.

Resumen

Uno de los principales retos en el uso de modelos de aprendizaje automático, o machine learning en inglés (ML), en finanzas es cómo explicar sus resultados. Recientemente han aparecido técnicas de interpretabilidad con este objetivo, pero existe discusión sobre su fiabilidad. En este documento contribuimos al debate proponiendo una metodología para evaluar la precisión de estas técnicas de interpretabilidad. Partimos de la generación de conjuntos de datos sintéticos, siguiendo un enfoque que nos permite controlar la importancia de cada variable explicativa (feature) en nuestra variable objetivo. Al definir nosotros la importancia de las features, podemos posteriormente calcular en qué medida las explicaciones dadas por las técnicas de interpretabilidad coinciden con la verdad subyacente. Por lo tanto, si en nuestro conjunto de datos sintéticos definimos una feature como relevante para la variable objetivo, la técnica de interpretabilidad también debería identificarla como una feature relevante. Desarrollamos un ejemplo empírico en el que generamos conjuntos de datos sintéticos de manera que se parezcan a datos de suscripción y calificación crediticia, donde la variable objetivo es una variable binaria que representa el incumplimiento del solicitante. Usamos modelos de ML no interpretables, como redes neuronales, para predecir el incumplimiento, y luego explicamos sus resultados usando dos técnicas populares de interpretabilidad, SHAP y permutation Feature Importance (FI). Nuestros resultados usando la metodología propuesta sugieren que SHAP identifica mejor las variables relevantes como tales, aunque los resultados pueden variar significativamente según las características del conjunto de datos y el modelo ML utilizado. Concluimos que el recurso a la generación sintética de bases de datos muestra un elevado potencial para supervisores y entidades financieras que precisen evaluar la fidelidad de estas técnicas.

Palabras clave: datos sintéticos, inteligencia artificial, interpretabilidad, aprendizaje automático, evaluación de crédito.

Códigos JEL: C55, C63, G17.

1. Introduction

The use of Machine Learning (ML) models is gaining traction in finance due to their better predictive capacity compared to traditional statistical techniques (see a survey by Königstorfer and Thalmann 2020, or Goodell et al. 2021). One of the use cases with greater potential is its application to credit underwriting and scoring, since by having better predictive capacity, ML models allow better estimates of the probability of default and therefore credit scores could be more accurate (e.g.: Bono et al. 2021). But this improvement in predictive performance does not come without risk. ML models can potentially be much more complex than traditional econometric ones, and this implies new challenges for both users and supervisors in terms of new model risk factors like biases, data quality, dependencies on third-party providers, etc. (EBA 2020, Dupont et al. 2020, BaFin, 2021). Importantly, one of the main challenges for using ML in credit scoring is the interpretability of the outcome of the models (IIF 2018, IIF 2019). While traditional statistical techniques are inherently interpretable[1] and therefore easy to explain their outcome through reasoning, the interpretation of the result of complex ML models could be a much more difficult task. This is why in recent years a field of study is flourishing in data science which brings together different methods and processes capable of explaining the influence of the explanatory variables on the outcome of ML models. We will focus on a strand of the literature that encompasses what are known as *post hoc* interpretability techniques, or model agnostic techniques, because they are applied after the model is trained. While they can represent a valuable tool for the challenge of interpretability in the use of ML models for credit decisions, there are currently doubts about their reliability (e.g.: Rudin 2019, or Ghorbani et al. 2019). Financial supervisors and regulators are currently looking into how to properly evaluate the fidelity of these techniques (e.g.: BaFin 2022, Dupont et al. 2020, EBA 2021), which motivates our work to build a novel approach to fulfill this task.

We start our study wondering why people ask for explanations. We identify two main areas of concern, model risk governance and fair lending, and provide a description of the current regulatory framework regarding the need for explanations in credit decisions both in USA and Europe. We then propose a framework to analyze the accuracy of different *post hoc* interpretability techniques for binary classification problems, such as credit underwriting and scoring. We do so by generating synthetic datasets, following an approach that allows us to control the importance of each feature on our target variable (default of applicants), while being completely agnostic about the underlying data generating process of the explanatory variables (features). This way we can generate a wide range of situations when choosing the number of features, instances, distribution classes, and percentage of zeros and ones in the target variable. As an empirical exercise, we apply two non-interpretable ML models to our synthetic datasets: XGBoost, and Deep Learning. We perform proper training and cross-validation for both ML models, to ensure good test performance, since reliable explanations require the

[1] Traditional statistical models are inherently interpretable because they rely on the assumption that the "data generation process" (DGP) is known at any moment. Therefore, the estimation of the parameters of said DGP directly provides us with the interpretability of the model.

ML models to perform properly (Blattner et al. 2021). Afterwards, we use two of the main global *post hoc* techniques, SHAP and permutation Feature Importance, which provide the relevance of the variables in the outcome of the two ML models. Since we have created the dataset ourselves, we can define the ground truth or importance between the features and the target, so we can compute the accuracy of the explanations given by the interpretability techniques. For this we first use the metric Ranking Based Ordering (RBO) which allows to compare how similar two rankings are, in this case the real ranking of our generated dataset and the rankings obtained from SHAP and permutation Feature Importance. We evaluate as well the absolute magnitude of importance obtained from these techniques with our ground truth. Our results suggest that the reliability of SHAP and permutation Feature Importance could vary significantly depending on the dataset characteristics and ML model used. Notwithstanding this, SHAP seems to have a better accuracy than permutation Feature Importance, particularly for XGBoost. We include a sensitivity analysis to understand to what extent a higher predictive performance of the original ML model can influence how accurate the explanation of an interpretability technique can be.

Our framework contributes to the literature on ML interpretability methods with an application to credit risk governance and regulation. To the best of our knowledge, this is the first study that proposes a methodology to evaluate how accurate are the results of global interpretability techniques. Our methodology allows us to calculate the performance of any post hoc interpretability technique, fully controlling the importance of the variables by artificially creating our own datasets, being completely agnostic about any underlying data generating process. This allows us to overcome the limitations of previous comparisons which rely on real data (e.g.: Krishna et al. 2022). The results are particularly relevant in credit underwriting, where transparency of the algorithms is essential. Our work offers a framework that lenders, regulators, and researchers can use to better assess the fidelity of the explanations of ML models in accordance with applicable regulatory requirements in the context of credit underwriting.

Notably, the use of synthetic data is still a developing field, and we acknowledge that our approach for generating the datasets does not cover all possible scenarios. For instance, due to the agnostic way in which we have generated our data at inception, the resulting correlation between variables is low. Similarly, both SHAP and permutation Feature Importance undertake permutations of features' values assuming that they are independent of each other, so their performance will vary in the presence of more correlated data (Kumar et al. 2020, Hall et al. 2021, Aas et al. 2021, Jullum et al. 2021). We assume that practitioners usually perform feature engineering to raw datasets. For instance, "grouping" involves treating a group of correlated features (with strong correlations between features in the group and weak correlations with features outside of the group) as a single feature from an explanation standpoint. Reducing the number of input features through "grouping" would result in lower correlation in empirical data sets, so this could help produce better explanations when the original data set has high dependencies. In any case, we are aware of these shortcomings, but our work reinforces the potential of using synthetic data to get a true benchmark[2] to assess the accuracy of the explanations of machine learning models. The fact that ML interpretability has become a priority research area for supervisors (see Blattner et al. 2021, or Akinwumi et al. 2021) motivates our work to investigate the potential of synthetic data sets as a tool for professionals and supervisors to examine the reliability of machine learning explanations.

While it is true that both SHAP and FI have limitations, we have chosen them to test our framework because they remain among the most popular at the moment for global interpretability and they are currently under the scrutiny of financial regulators (see for instance the discussion paper by EBA, 2021, on ML for IRB models, where they refer to Shapley values as a widely used technique). Additionally, they are also used in the industry (FinRegLab, 2022) and its use is extending as well in academic work related to the use of machine learning for credit scoring, like for example: Albanesi and Vamossy (2019), Ariza-Garzon (2021), Misheva et al (2021), Cascarino et al (2022), or Bücker et al (2022). Therefore, due to their popularity and widespread use, we believe that, as of today, they are good benchmarks to put under test. Last but not least, as there exist public open-source implementations of these tools, it makes our exercise more transparent.

The paper is organized as follows. Section 2 provides a literature review on the interpretability of ML techniques. Section 3 explains the need for explanations in credit decisions. Section 4 explains the two interpretability techniques that we will evaluate: SHAP and permutation Feature Importance. Section 5 dives into the data generation and ML models, and in Section 6 we show our results on the accuracy of explanations. Section 7 concludes.

2. Literature review

There is an extensive and growing literature on the applications of ML in finance. Königstorfer and Thalmann (2020) and Goodell et al. (2021) provide an overview of AI and ML research and note that credit risk is one of the key topics studied (e.g. Liu and Schumann, 2005 who look into feature selection for credit scoring using several ML methods, Shen et al. 2019 who use neural networks in imbalanced credit risk evaluation, or Xia et al. 2020 who propose a novel tree-based ensemble model applied to credit scoring). The list of topics is heterogeneous, including asset pricing (as in Gu et al. 2020 whose auto-encoder asset pricing model delivers out-of-sample pricing errors that are far smaller compared to other leading factor models, or Avramov et al. 2021 show that investments based on deep learning signals extract profitability from difficult-to-arbitrage stocks), market risk management (Arimond et al. 2020 investigate if ML can advance the process of estimating Value at Risk), corporate failure prediction (both Sheng et al. 2019, and Lee et al. 2020 propose a graph convolutional network based credit default prediction model), derivative pricing (Ye and Zhang 2019 combine techniques of ML with regression

[2] We propose one particular way to control the importance of the data, but there could be others, for instance, assuming a causal model.

analysis and apply the new methodologies on financial derivatives), forecasting foreign exchange rates (Nag and Mitra, 2002 use genetically optimized neural networks), or volatility forecasting (Arroyo et al., 2011 study different approaches, including ML, to forecast interval financial time series).

Focusing on Explainable AI (xAI) in general, and interpretability of ML in particular, it is noted that this field is advancing as ML models get more popular. Leaving apart efforts on building replicable, white-box models, the main approach to interpret complex ML models is to rely on model agnostic techniques, or *post hoc* evaluation techniques, that aim to explain the outcome of any non-interpretable model. These techniques area designed for local (explaining at an individual level) or global explainability (explaining the whole dataset)[3]. LIME (Ribeiro et al. 2016) is probably the most popular local technique (recently "upgraded" to Anchors in Ribeiro et al. 2018). The main global *post hoc* interpretation techniques are permutation Feature Importance (Breinman 2001, Fisher et al. 2019) and SHAP (Lundberg and Lee, 2017, Lundberg et al. 2020), which can be also used as local interpretability[4].

The importance of xAI in credit underwriting is represented by recent work on explainability in credit risk management (see Misheva et al. 2021), and focused on fair lending, or racial discrimination, like Barlett et al. (2022) who find that racial discrimination between Fintech and non-Fintech lenders in the US mortgage market, or Fuster et al. (2022) who find that ML increases disparity in rates between and within protected groups, with these changes attributable primarily to greater flexibility of better statistical technology.

There is a recent strand of the literature that tries to understand the suitability and reliability of *post hoc* evaluation techniques. Despite their initial success and popularity in the industry and in academia, there are several papers that highlight the shortcomings of these techniques. Rudin (2019) says that *post hoc* explanations are not reliable because they provide correlations with no informative content, and therefore they are not truly representative of the model they try to interpret. Ghorbani et al. (2019) focus on the application of LIME and SHAP for interpretation of neural networks and they claim that the output of the explanation techniques is highly sensible to small perturbations in the data, even when those perturbations do not change the predictions of the classifier that these techniques try to explain. Mittelstadt et al. (2020) also focus on the application of LIME and SHAP and argue that perturbation points created by these methods are not at all intuitive, especially for structured data. Slack et al. (2020) also criticize LIME and SHAP because according to them is relatively easy to deceive techniques like SHAP even when

[3] For example, imagine we have a credit default dataset and we apply a ML model, where the target variable is binary (default or not), and the features are the person's income, loan size, and loan type. A local interpretability technique will show us how the different features contributed to the predicted probability of default for a particular individual. In this way, the local interpretability of the models allows us to know the reason why the ML model predicts that a person repays their loan or not. Global interpretability, on the other hand, refers to how features affect the predictions of ML models in general. Following the example above, global interpretation techniques could help us to understand which of the three features of the data set (income, loan size, loan type) most influences the prediction in the entire data set, and therefore we could determine which is the most important feature globally.

[4] We will explain in Section 4 all these techniques in detail.

using biased (racist) classifiers, although Vres et al (2020) proved that this statement might be too pessimistic. Other papers that look at the lack of stability of LIME are: Alvarez Melis and Jaakkola (2020), Visani et al. (2020), and Gosiewka and Biecek (2019), and all found that LIME can give very unstable *post hoc* evaluations.

Another set of the literature analyzes the reliability of *post hoc* explanations methods using synthetic data. Since the truth about the data generation process cannot be known when using real data, the use of simulated data can help to understand the reliability of these techniques. Barr et al. (2020) created datasets using copulas, and showed that the correlation of redundant variables can affect the explanation given by SHAP. In a similar fashion, Aas et al. (2021) showed that correlation of the features could affect the explanations of SHAP. Zhang et al. (2019) also used synthetic generated data and showed that there are several sources of uncertainty and instability for LIME. In fact, (Hall et al. 2021) share the concern on inconsistent explanations, as different configurations of the same ML model, or refreshing the same ML model with new data can result in different explanations for the same consumer, if not controlled. On top of that, the success of an explainable ML model will rely on the human comprehension of model behavior by less technical audiences (Kumar et al. 2020).

While certainly most of these papers find limitations and shortcomings to *post hoc* evaluation techniques like LIME and SHAP, on the other hand there is an ongoing effort in xAI literature that tries to improve these methods. Papers like Frye et al. (2019), Heskes et al. (2020) and Janzig et al. (2020) are incorporating new elements to the SHAP methodology in order to incorporate notions like causality. Also Miroshnikov et al. (2021) propose a feature grouping technique to design appropriate groups explainers offering consistency guarantees and more stability of the results, working in the same direction as Jullum et al. (2021), who propose a new method called *groupShapley*, while Aas et al. (2021) try to extend the kernel SHAP (Lundberg and Lee, 2017) to account for features' dependencies. This shows a promising future path to include more sophisticated elements to evaluate these techniques. In the context of adverse action notices, grouping features can assist on the usability of these techniques to produce information that is valuable to a rejected applicant looking for a feasible path to credit acceptance within a time period, as usually an isolated change in just one feature (or a limited set of) does not produce significant changes in the estimated probability of default. We leave for further research to assess the impact of features' grouping on the accuracy of the explanations in our framework.

As stated in the introduction, we contribute to the literature by proposing a methodology to evaluate how accurate are the results of global interpretability techniques, using synthetic datasets. Despite the existence of extensions of SHAP and new interpretability techniques that could handle the presence of correlation in datasets (Aas et al. 2021) or define richer objective outcomes to explain (Giudici and Raffinetti 2021), we prefer to test the basic versions of SHAP and FI, since their used is widespread. At this stage, the purpose of our article is not to determine which of the existing interpretability techniques is better overall, but rather to outline a practical framework on how to test the accuracy of these techniques using a novel approach that complements prior work on this field by creating a highly curated and fully controlled experiment to evaluate the real accuracy of these tools. This includes potential mis-prediction sources, and is broader than previous analysis that only compared these tools in relative terms.

3. Why people ask for explanations

Curiosity is an element intrinsic to human nature. People tend to ask questions about events or observations that they consider abnormal or unexpected from their own point of view. This way, a primary function of explanation is to facilitate learning. In fact, a good explanation will create a shared understanding, and therefore a sense of trustworthiness (Miller, 2018). This is key in decision making and human-machine interaction. It is therefore not surprising that there is a desire to explain the results of so called black-box ML models, particularly on such important issues as credit decisions. But in addition to curiosity, there are legal requirements that demand, to greater or lesser degree, the explanation of the results of the use of ML for credit concession. Mainly, lenders must detail to the borrower in writing the main reasons for taking an adverse action on a loan application, following consumer law requirements. Additionally, regulators will oversight any potential discrimination of protected classes (like gender or race), and supervisors will analyze the sustainability and transparency of the models aiming to mitigate potential operational risks.

In this section we will revise the prudential expectations in two main domains: model risk governance and fair lending, addressing current doctrines both in US and EU.

3.1 Model risk management

US regulators have issued extensive guidance outlining their expectations for steps that banks should take in developing, monitoring, and validating models throughout the lifecycle (Blatter and Spiess, 2021). This broadly applies to all use cases that might lead to unexpected losses, or other negative outcomes and requires risk management processes and controls[5]. Banks subject to prudential oversight are generally required to conduct a comprehensive review and monitoring of credit models, especially those engaged in retail banking, due to consumer protection laws. The prudential model risk management expectations emphasize various aspects of model transparency, some of which can prove to be challenging in the context of ML underwriting models. At a broad level, the guidance requires documenting on a transparent way how the learning algorithm produced the prediction. Though, as a condition precedent, developers must evaluate whether models are based on relationships between variables that are intuitive and make economic sense. Banks must also perform appropriate sensitivity analyses to establish the sustainability of the model in business-as-usual situations, and prepare processes that identify and mitigate operational risks. See for example Unceta and

[5] The Federal Reserve Board’s Supervisory & Regulation Letter 11-7 is often used to refer to all three US agencies’ guidance.

Nin (2020) for a theory on copying the behavior of complex ML classifiers, aiming to avoid production bottlenecks and having to retrain tailor-made solutions if unexpected risks arise.

In the European front, article 144(1)(b) CRR prescribes that banks' models for internal ratings and probability of default (PD) and loss given default (LGD) used in the calculation of regulatory capital requirements are aligned for internal purposes like risk management, credit approval and decision-making processes. To this purpose, this regulation establishes a "use test", which rationale is to prevent banks using proprietary models only to reduce capital requirements, but also use their models for other internal purposes (EBA, 2021). This requirement may hamper the introduction of ML models for credit decisions or early warning systems, due to the challenges that banks may encounter in complying with strict CRR requirements.

Finally, EBA (2021) remembers that the pivotal challenge towards ML requires banks to (i) interpret their results, (ii) ensure their adequate understanding by the management, and (iii) justify their results to supervisors. In fact, as stated by the European Commission (2019, 2021), the results of a ML model need to be interpretable for all the people who participate in the process, including clients, since the decision that entails the concession, denial or refinancing of a loan can have a significant economic impact on people's lives.

This is consistent with our previous results found in Alonso and Carbó (2020) where we concluded that for credit scoring as a use case a key risk factor is the interpretability of models' output. But, how can potential ML users evaluate the reliability and usefulness of information produced by currently available *post hoc* interpretability techniques?

3.2 Fair lending

Banks are subject to broad anti-discrimination requirements regardless of the type of model they might use to predict the default of borrowers.

For instance, the US has prohibited illegal discrimination and establish a strong framework since long time ago[6]. In this jurisdiction, the legal requirements give rise to two fair lending principles: disparate treatment and disparate impact. Disparate treatment focuses on whether lenders have treated applicants differently based on protected characteristics, like race or gender. Disparate impact addresses lenders' use of underwriting practices that have a disproportionately negative effect on protected classes, unless there is a legitimate business need that cannot reasonably be achieved through alternative means with a smaller adverse impact[7]. Both

[6] The Equal Credit Opportunity Act (ECOA) prohibits discrimination in "any aspect of a credit transaction" for both consumer and commercial credit on the basis of race, colour, national origin, religion, sex, marital status, age, or certain other protected characteristics, and the Fair Housing Act (FHA) prohibits discrimination on many of the same bases in connection with residential mortgage lending.

[7] Historically, regulators have looked at whether particular variables have an "understandable relationship to an individual applicant's creditworthiness" as well as a statistical relationship to loan performance in determining whether they meet a legitimate business need (Blatter and Spiess, 2021).

principles rely on statistical tests and analyses of data inputs that can be more challenging to implement in the context of complex ML models. For example, the identification and management of variables that may proxy for protected class status under both disparate treatment and disparate impact theories of discrimination requires a high degree of transparency into how the models are built and how they make predictions. However, the fact that ML models can better unravel patterns in consumer data has raised concerns about whether they might be unintentionally using sensitive information to generate the predictions. At the end, the fear is that these models may result in more accurate but harm fairness in their predictions (Bono et al. 2021).

Under the US Equal Credit Opportunity Act (ECOA) and the Fair Credit Reporting Act (FCRA), many credit decisions that are adverse (either rejection or offering less favorable terms) to the applicant must be summarized through a predefined set of written explanations known as “adverse action notices". Banks must indicate the principal reasons for the adverse action and accurately describe the features actually considered, but it is not required to state how or why a given characteristic contributed to an adverse outcome. At this stage banks face the challenge of how to evaluate the statistical and economic relevance of explanatory variables in a way that meets current regulatory requirements.[8]

Actually, these concerns transcend jurisdictions, as regulatory expectations are calibrated to the degree of risk posed by the particular use case, and credit underwriting is often considered to be among the highest risk activities; see for instance. The EU-wide legislative proposal on artificial intelligence (“AI act”) includes among the high-risk use cases the evaluation of the creditworthiness of natural persons. European Commission's Guide to Ethical Principles of AI (2019) cites the principle of explicability of algorithms as one of the critical elements, and in accordance with the European General Data Protection Regulation (GDPR) Article 22 on automated individual decision-making, including profiling, *the data subject shall have the right not to be subject to a decision based solely on automated processing*, implying that *decisions [...] shall not be based on special categories of personal data* and pointing to the need to include human judgement in any decision-making process (i.e.: data controller).

All in all, these two issues together of model risk governance and fairness, motivate the development of different methodologies to evaluate the compliance with current regulation. For example, counterfactual explanations may assist on assessing potential discriminatory issues (Wachter et al. 2017), while *post hoc* techniques are usually being used to help on model governance and the provision of adverse action notices (Hall et al. 2021). In this study we will investigate these latter tools, particularly by computing the accuracy of their explanations. In Section 4 we will

[8] Apart from local factors regarding his/her own case, it is important for a customer to understand what global factors resulted in an adverse action on their credit decision (e.g.: length of credit history), while also understanding what are the local factors that are within their control to achieve a favourable outcome in the near future (e.g.: lower utilization of credit limit). Therefore, here the challenge for ML goes into both dimensions, global and local interpretability of the predictions.

describe two of the most cited ones in the literature, SHAP and permutation Feature Importance, which we will put to the test in Section 5.

4. Two techniques for explanations

In the ML literature, early work on explanation often focused on producing visualizations of the predictions in order to assist ML experts in evaluating the correctness of the model. Beyond visualization, some researchers nowadays try to create interpretable models, devising surrogate models that can be explained through reasoning (e.g.: Unceta and Nin, 2020). This is based on the assumption that in order to fully interpret a ML algorithm it needs to be transparent and entirely replicable (Hoepner et al. 2021). This would mean that a human actor could interpret each (relevant) analytical decision step. However, the quantitative models with the greatest predictive power usually are ML models that are not intrinsically interpretable (Bono et al. 2021), requiring *post hoc* techniques to assist on the task of explaining their outcome.

In this article we are going to focus on two of the most popular techniques: SHAP (Lunbdberg and Lee, 2017) and permutation Feature Importance (Breiman, 2001). The former will allow the user of a complex ML model to know which are the features that most influence a certain prediction, and the latter will inform about which are the features that most influence the error that the model is making. Both methods are model agnostic, meaning that can be used for any ML model (e.g.: regressions, tree-based models, or deep learning). They are applied after the ML model has been trained, and are based on perturbing or permuting the input data on the test sample to determine the relevance of the variables, measuring how those changes affect the ML model's output.

It is important to highlight that both techniques can explain how each feature affects all individuals in the entire dataset, which is called global interpretability[9]. Now we will briefly explain in the following Section how both SHAP and permutation Feature Importance work.

4.1 Permutation Feature Importance

Permutation Feature Importance (also known and referred to in this article sometimes as Feature Importance) is a *post hoc* evaluation technique that measures the impact of each feature in the dataset based on the impact on a given performance metric. As mentioned before it was introduced by Breiman (2001) for Random Forest, but Fisher, Rudin and Dominici (2018) provided a model agnostic version, so we will be able to use it on both XGBoost and Deep Learning in our exercise.

[9] SHAP that stands for Shapley Additive Explanations, is initially designed for local interpretability, letting the user get a justification about how the model works for an individual prediction or for a set of predictions. But it allows to add the local values in order to obtain a global interpretation as well. Other techniques that allow local interpretability are LIME (Local interpretable model-agnostic explanations), PDP (Partial Dependent Plots), ICE (Individual Conditional Expectation), ALE (Accumulated Local Effects)

The method computes the importance of a given feature j as follows. We need a trained model $\hat{f}$ with an original set of features X to predict our target variable y. First, we estimate the original model error $e_{orig} = L\left(y, \hat{f}(X)\right)$ for each feature j of the dataset. As we are in a classification problem, then y is a binary variable and our measurable error will be $(1 - AUC)$[10]. We repeat $k = 10$ times the following process to make sure that a specific shuffling of the feature is not biasing our results[11]. For the chosen feature j, we randomly shuffle its value to j' and we estimate the new error $e_{perm} = L\left(y, \hat{f}(X_{j'})\right)$. Finally, we compare both error values to compute the feature importance of j as $FI_j = e_{perm} - e_{orig}$. After 10 iterations, we sort features by descending value of its arithmetic average $\overline{FI_j}$. If the feature is important, then we should observe a considerable impact in AUC. If the feature is not important, the AUC should be similar when using shuffled values instead of the original ones. The method is simple but time consuming. It implies repeating the process for each feature at least several times, so the higher the number of features or the more complicated is to train the original ML model, then the more time it will take to compute FI. One of the drawbacks, as it happens with SHAP, is that it assumes feature independence something which might be unrealistic in credit decisions, as we will comment later on. Additionally, this technique cannot indicate us the direction of the effect of a given feature. For example, it cannot indicate whether increases or decreases in the value of feature j are related to increases or decreases in the value of target variable y (the probability of default in our exercise).

4.2 Shapley Additive Explanations (SHAP)

The SHAP method relies on game theory. In a cooperative game with M players (features) and a function (model) that values how much total output is generated if all the players contribute together, SHAP is a method that attempts to measure the individual contribution of each player to the output generated by the cooperation of all players. From an economic standpoint, it can be interpreted as a weighted average of a feature's marginal contribution (Shap value) to every possible subset of grouped features (coalitions).

SHAP explanation of a feature j for a given instance x could be computed as:

$$g(S') = \emptyset_0 + \sum_{j=1}^{M} \emptyset_j S'_j$$

Where g is the explanation model and $\emptyset_j \in \mathbb{R}$ is the Shapley value of feature j. We start by considering all possible coalitions of features that exclude the feature of interest, including the empty set $\emptyset_0$. For all different coalition vectors $S' \in \{0,1\}^M$,

[10] AUC refers always to ROC-AUC or Area Under the Curve of the Receiving Operating Characteristic is a common measure to evaluate the classification power of the models. AUC is a probability curve and the area under it will give as a value that ranges between 0 and 1. The higher, then the better is the model at distinguishing between classes. A value of 0.5 indicates that the model has no discrimination capacity. We have tried with other metrics like recall or F1, and the main results of our paper do not change.

[11] After repeating the process 10 times, the results of the Feature Importance do not seem to change.

where M is the maximum coalition size, and an entry of one means that the corresponding feature value is "present" and zero that it is "absent", we compute the difference in the predicted outcome with and without the feature of interest. The Shapley value will be calculated as the weighted average of the differences in the predictions among all coalitions.

Mathematically, the Shapley value or contribution $\emptyset_j$ of a given feature j in a prediction p is summarized by the following formula:

$$\emptyset_j = \sum_{S \in \frac{M}{j}} \frac{|S|!\,(M - |S| - 1)!}{M!} \cdot [p(S \cup j) - p(S)]$$

Where M is the total number of features, $S \in M/j$represents all possible coalitions of features excluding feature j, considering all possible orders, and $p(S \cup j) - p(S)$ represents the difference in the predicted outcome p when we consider a particular coalition of features and feature j minus the predicted outcome when we consider the coalition of features without feature j. The term $\frac{|S|!(M-|S|-1)}{M!}$ assigns different weights to the differences, depending on the features that are in the set $|S|!$, the features that have to be added$(M - |S| - 1)$, and all normalized by the features that we have in total.

For illustrative purposes, let's consider the following example. Imagine a consumer loan dataset that includes the features "income," "loan size," and "monthly credit card payments". The target is a binary variable that indicates whether or not the loan has defaulted. Imagine that we want to know the importance of the variable "income" on the probability of default of an individual x. We have the following possible coalitions without including "income":

- No features
- Size of loan
- Credit card
- Size of loan and credit card

For all these four coalitions we compute the predicted probability of default of individual x with and without "income" in order to get the marginal contribution of "income" for the four coalitions. The Shapley value of "income" for the predicted probability of default of individual x is the weighted average of those marginal contributions. To get the global contribution SHAP of "income" to the probability of default in the whole dataset, we repeat the process for all individuals in our dataset, and average over the absolute Shapley values. Features with large absolute Shapley values are considered as important local features (and correspondingly, SHAP values for a global explanation).

Interestingly, the Shapley value is the only attribution method that can achieve the following desired properties: efficiency, as the sum of Shapley values of all features equals the value of the total coalition; symmetry, because if two features contributed the same across all possible coalitions, their Shapley value should be the same; dummy, as if a feature does not change the predicted outcome, regardless of the coalition of features, then its Shapley value should be zero; and finally additivity, as for any pair of games x and z it follows $\emptyset_{\mathrm{i}}(\mathrm{x}+\mathrm{z}) = \emptyset_{\mathrm{i}}(\mathrm{x}) + \emptyset_{\mathrm{i}}(\mathrm{z})$. However, there are two key drawbacks which are important to highlight:

- Feature independence: SHAP makes the unrealistic assumption that features are uncorrelated. This assumption strikes over real-world datasets in finance where variables are strongly correlated, as usually happens in credit portfolios. This is common in many *post hoc* explainability techniques. In fact, permutation Feature Importance relies on this property as well, and it calls into question the utility of these techniques in applied settings.

- Convergence: when the number of features is high, the number of coalitions can be almost impossible to manage, and this is why there are several numerical methods for approximating the results, usually through sampling when permuting the input data. The resulting explanation can change as more samples are used, creating another source of potential instability in the results, as different configurations of the same ML model, or refreshing the same ML model with new data can result in different explanations for the same consumer (Hall et al. 2021). This is a special concern in financial services, especially for the generation of adverse action notices for credit decisions, where two similar models giving different explanations to the same applicant may raise questions (e.g.: see the consistency test in FinrRegLab, 2022).

  To approximate the results in our article we will use Tree explainer (Lundberg, 2018) for XGBoost, and for Deep Learning we will use Deep SHAP (Shrikumar et al, 2017).

To tackle these weaknesses, new promising research is ongoing. For instance, Miroshnikov et al. (2021) propose a feature grouping technique that employs an information-theoretic measure of dependence to design appropriate groups of features. Achieving independent groups of features allows to reduce the dimensionality of the problem and consequently the computational complexity of generating explanations, while increasing the accuracy of the partitioning and therefore the reliability and stability of the results. In this line Jullum et al. (2021) presents an adaptation called *groupShapley*. Following a different approach, Aas et al. (2021), extend the Kernel SHAP method (Lundberg and Lee, 2017) to handle dependent features, though at a high computational cost.

5. Data and models

While *post hoc* evaluation techniques are helpful in understanding the relevance of the input variables on a model's prediction, how can we know when these explanations are reliable? To answer this question, we use a framework based on creating several synthetic datasets to test the goodness of fit of *post hoc* explanation techniques. Our synthetic datasets contain a binary variable that would correspond to the target variable. The approach we follow to obtain the synthetic datasets allows us to choose flexibly the characteristics of the features.

Subsequently, we estimate two non-interpretable ML models, XGBoost and Deep Learning, to predict the binary variable based on the features. These two ML models are among the most used ones in the literature on credit default prediction (Königstorfer and Thalmann, 2020). Once we have estimated them, we will apply Feature Importance and SHAP to reveal which are the most important features for the ML models' predictions. Since the datasets have been created by us, we know the ground truth, in particular which are the features that most influence the target variable. By comparing the real importance of the features with the importance given to the features by the interpretability techniques, we can assess the reliability of the interpretability techniques. Before we begin our discussion, we explain in the following Section how the dataset generation process works.

5.1 Synthetic datasets

The importance of synthetic datasets to evaluate the performance of *post hoc* interpretation techniques is paramount. Without knowing the ground truth of the data it is not possible to understand up to which degree is the explanation given by the interpretability technique correct. Despite the existence of a new and growing literature on the use of synthetic data for ML interpretability, there is not a standardized procedure on how to create these synthetic datasets. As mentioned in the literature review, some papers have used Gaussian Copulas (Barr et al. 2020) or Monte Carlo linear models (Aas et al. 2021) to generate the datasets. In this paper we want to create synthetic datasets by being completely agnostic about the underlying data generating process. We will take no assumption or knowledge on the data we wish to create, this way we can be as flexible as possible. With our methodology we can choose randomly, for each dataset, how many observations, how many features, the percentage of positives in the binary target variable, the statistical distribution of the features (in our case we allow Normal, Beta, Gamma, Cauchy, Uniform distributions), and the number of categorical variables. More importantly, we can control the importance of all features with the binary target variable through the following four characteristics of each feature: Overlap between positives and negatives in the target variable conditional on the distribution of the feature, the percentage of noise, sparsity or missing values (#NAs)[12] and outliers or extreme values in some features (corruption). We will explain how these four characteristics control the importance of the feature on the Target in section 5.1.1.

There are two advantages of creating datasets with this methodology. First, we generate very granular levels of importance, feature by feature. This allows for a richer analysis than having just identified a cluster of important or redundant features, as in Barr et al. (2020). Second, we control for different circumstances, like different sample sizes, distribution of features, or number of #NAs, giving robustness to our results. On the other hand, one of the drawbacks is that the resulting correlation among variables in our dataset is low as we will explain later.

5.1.1 Step-by-step creation of the synthetic datasets

The steps to create a single dataset are as follows:

[12] Not Available observations or missing values.

1. We first decide the number of observations as a random integer from 50,000 to 150,000.
2. We select the percentage of zeros and ones in the target variable.
3. We select the number of features of each class as a random integer between a minimum and a maximum number.
   a. For each class, we specify a minimum and a maximum value for its mean.
   b. For each class, we specify a minimum and a maximum value for the following properties: overlap, noise, sparsity and corruption.

We will now define each of these four characteristics through which we control the influence of the feature on the variable Target.

Overlap: It is a parameter that takes values between zero and one and refers to the amount of separation between positives and negatives in the target variable conditional on the distribution of a given feature. For each feature of the dataset, we create a distribution for the observations with Target equal to one, and a different distribution for the observations with target equal to zero. If a feature has overlap one, it means that its distribution associated with Target equal to one overlaps completely with its distribution associated with Target equal to zero, and it has no discriminatory power. On the other hand, if a feature has an overlap of zero, then its effect on the Target is maximum, see **Figure 1** for a normal distribution with low overlap, and **Figure 2** for a normal distribution with high overlap.

Once we decide the desired degree of overlap, the way to create the final values of the feature is as follows: we select randomly a mean within the desired interval for the distribution of observations with Target one, and a mean for the observations with Target zero. The standard deviation for the observations with Target one is calculated as the absolute difference between those two means. And the standard deviation for the observations with Target zero is calculated as the standard deviation of the observations with Target one multiplied by a value such that both distributions have the desired overlap. Here we depict an example for a variable with a Normal distribution[13]:

$$X \sim N(\mu_1 \in [\mu_{min}, \mu_{max}], \sigma_1 = |\mu_1 - \mu_2|) \; if \; Target = 1$$

$$X \sim N(\mu_2 \in [\mu_{min}, \mu_{max}], \sigma_2 = \alpha * \sigma_1) \; if \; Target = 0$$

Where α is the parameter with a value such that we can achieve the desired overlap, μ_1 and σ_1 are the mean and the standard deviation of the Normal distribution for the observations with Target equal to one, and μ_2 and σ_2 are the mean and the standard deviation of the Normal distribution for the observations with Target equal to zero.

Noise: refers to the percentage of values of the feature that will have a random noise created. It follows a Bernoulli, with probability $p_{noise} \in [p_{noiseMIN}, p_{noiseMAX}]$, multiplied by a uniform variable that controls the extent of the noise. The random noise will be created as follows:

[13] The creation of variables belonging to the other distributions follow a similar process. We explain in the Appendix how to proceed with them.

Figure 1. Low overlap in normal distributions.

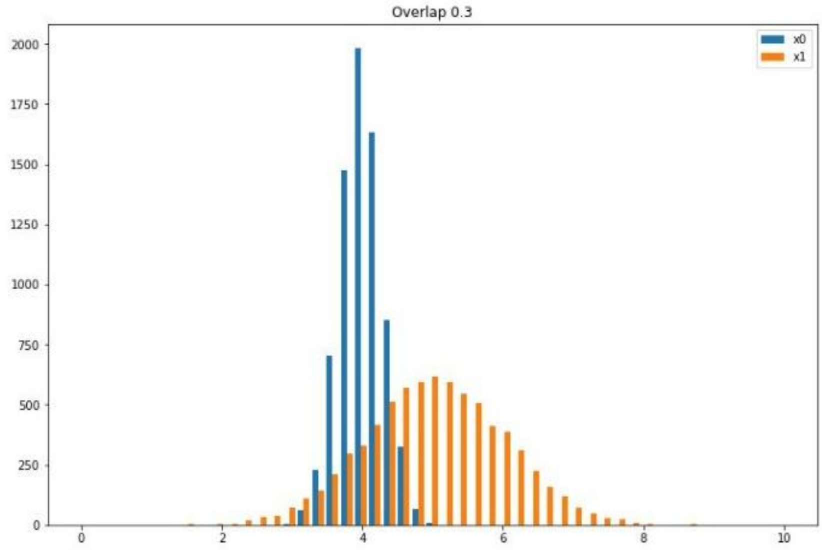
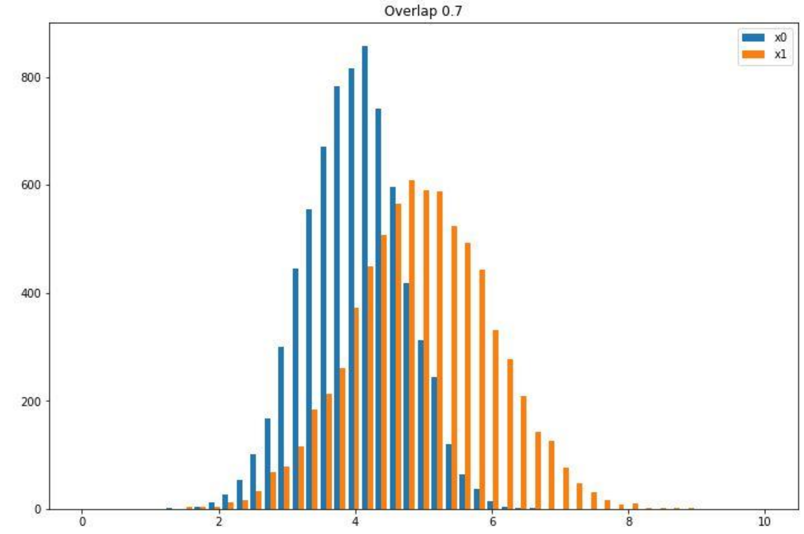


Figure 2. High overlap in normal distributions.



$$Noise \sim Bernoulli(p_{noise} \in [p_{noiseMIN}, p_{noiseMAX}]) * Uniform(-1.5 * IQR(X), 1.5 * IQR(X))$$

In case of being one, the Noise variable changes the value of the original value with an extent from minus 1.5 times the interquartile range of the original variable *IQR(x)* to plus 1.5 times the interquartile range of the original variable.

Corruption: it is a parameter that we use to control the percentage of elements of each generated feature that will change its position without taking into account the Target values giving rise to outliers.

And finally, **sparsity** is a parameter that controls the percentage of variables which values will be replaced by empty or null value (#NAs).

Giving different values to the parameters of Overlap, Noise, Corruption and Sparsity we can control the importance of a given feature on the Target. Therefore, in order to create the ranking, since all four characteristics are values between zero and one, for each feature we sum the four characteristics, and we order the features in descending order. The less Overlap, Corruption, Noise and Sparsity, the more important the feature. Following this, we create 250 datasets, with (bounded) random values for the number of instances and features and different parameters. In **Table 1** we summarize the values that we have used to create the datasets.

Table 1. Parameters for the generation of synthetic datasets

Parameter	Min	Max
Number of observations	50,000	150,000
Percentage of ones in Target	3%	7%
Number of Normal variables	10	25
Number of Uniform variables	10	25
Number of Cauchy variables	10	25
Number of Beta variables	10	25
Number of Gamma variables	10	25
Number of categorical	10	60
Probability sparsity	30%	75%
Probability corruption	50%	95%
Probability noise	50%	95%
Probability overlap	40%	90%
Mean Normal	-1,500	15,000
Mean Uniform	-1,500	15,000
Mean Cauchy	3	40,000
Mean Gamma	10	400,000

Table 2. Parameters for the generation of synthetic datasets

	Number of features	Number of rows	Target
Mean	71.21	98,661.71	0.049
Standard deviation	9.37	28,298	0.011
Minimum	45	50,217	0.030
25%	64	75,184	0.039
50%	71	97,453	0.050
75%	77	122,022	0.060
Maximum	100	149,801	0.069

We acknowledge that the accuracy of the interpretability techniques will depend on the particular parameters of **Table 1**. For instance, had we chosen a set of relaxed parameters (like minimum probability of sparsity 1% and maximum probability of sparsity 5%), the accuracy of the *post hoc* evaluation techniques would be higher. We assume that by using these ranges of values we control realistically for situations we could encounter in real applied settings, just by averaging out our results for each parameter. The parameters regarding the number of features, number of rows and target have been chosen so that the resulting dataset is similar to other credit datasets, like Alonso and Carbo (2020), or "Give me some credit", available in Kaggle.com. We acknowledge that more work is needed to generate datasets which resulting distributions and correlation structure could be similar to real credit datasets.

Notwithstanding this, since we draw the features one by one without explicitly imposing any correlation structure, our dataset present low levels of cross correlation, between -10% to 10% at maximum. Two features with similar overlap would be correlated, but since we are using random overlaps for every feature, and we are adding a significant amount of noise, sparsity, and corruption, the resulting correlations end being low. We will not be able to recreate situations in which the data displays higher correlations among variables, something that might be characteristic in real life credit datasets. While it is potentially feasible to adapt the methodology to accommodate a particular correlation structure, we would need to abandon the assumption of being fully agnostic on the underlying processes. Consequently, we find it insightful to start the analysis using this idealized dataset, due to the advantages mentioned before, and leaving for further research how to impose a higher level of correlation in the synthetic dataset. All in all, we believe that this methodology contributes to the literature by showing a way to make a fine grain comparison of the two rankings of relevance of the variables, controlling for a broad scope of real life situations (while not exhaustively) in credit portfolios.

5.2 Machine learning models

Before examining the interpretability techniques, we will need to estimate our Target variable using two (non-interpretable) ML models: XGBoost and Deep Learning. Ideally, the data should be split into three samples, train (60%), test (20%), and validation (20%). The validation test sample should be used to choose the hyper parameters and the proper architecture of the ML models. Since our aim is to create potentially hundreds of datasets, we could not afford cross validate the ML models in all of them, since that would have increase the computational time exponentially. Therefore, we have decided to realize a proper 5-k fold cross validation with only 20 sets. We select the architecture for XGBoost and Deep Learning that achieves the best AUC-ROC on average on the validation test for those 20 sets, and we consider those architectures as the baseline XGBoost and the baseline Deep Learning[14] for the rest of the paper. Therefore, for the reminder of the paper, we use

[14] We show in section 5.2.1 the baseline XGBoost and in section 5.2.2 the baseline Deep Learning model.

80% of the data to train the ML models, and 20% of the data to test the out-of-sample performance.

5.2.1 XGBoost

Gradient-boosted decision trees (XGB) are an ensemble ML method that consists on initially estimating a very simple tree with the training dataset and then by looking at the residuals it fits another tree by giving a higher weight to the observations erroneously classified and repeats this process subsequently with more trees until a stopping criteria is met. The boosting method works as a committee method to aggregate the decision of the individual trees. The committee evolves over time and the members cast a weighted vote. The final prediction is obtained by taking a weighted majority vote on the sequence of generated trees (Hastie, Friedman and Tibshirani, 2009).

Our baseline XGBoost is composed of 80 trees, estimated using a logistic regression as loss function, a learning rate of 0.1, and a maximum depth of 3. All the sub-samples are used for fitting the individual decision trees, and we use the minimum square error with improvement score by Friedman as a function to measure the quality of the split.

5.2.2 Deep Learning

Artificial neural networks (NN) learn non-linear relationships between features and the target variable through several inner layers (Bengio and Lecun, 2007). The first layer consists of the features of the input data, which are used to generate latent features that make up the nodes in the second layer. The evaluation is conducted by weighting the sum of inputs, based on an activation function which combines several features into a single number (usually between zero and one). This process repeats until the final layer, where predictions for the target variable are generated.

In our study we will estimate a Deep Neural Network (those with three or more intermediate or hidden layers) and we will use the ReLU activation function at the hidden nodes, as defined by:

$$\mathrm{ReLU}(\mathrm{x_i}) = \max(0, \mathrm{x_i})$$

The main advantage of using the ReLU function over other activation functions is that it does not activate all the neurons at the same time. In particular, the neurons will only be deactivated if the output of the linear transformation is less than 0.

For the output layer, we use a sigmoid function, as defined by:

$$\mathrm{Sigmoid}(\mathrm{x_i}) = \frac{1}{1 + \mathrm{e}^{-\mathrm{x_i}}}$$

This function is widely used for binary classification problems, as it returns values between 0 and 1, which can be treated as probabilities of a data point belonging to a particular class (in our case a default or not).

There are many choices to make when structuring a neural network, including the number of hidden layers, the number of neurons in each layer, and the activation functions. For our analysis the number of layers and the number of neurons in each layer, along with other hyper-parameters of the model, are chosen using the open-source libraries Keras and Talos in Python. Our baseline architecture consists of four hidden layers with 512, 256, 128 and 64 neurons respectively, and in its implementation, to improve the convergence of the weights, input data have been normalized to have a mean of 0 and standard deviation of 1. As neural networks tend to be low-bias, high-variance models, which gives them a tendency to over-fit the data, we apply dropout of 20% to each of the layers, limiting the complexity of the fitted model.

6. Empirical results

6.1 Predictive performance

We have created 250 sets, and in all of them we have used XGB and NN to predict the target variable in the test sample. The model with the highest AUC-ROC has been always XGBoost. As shown in **Table 3** the minimum AUC achieved by XGBoost has been 0.768, while the maximum has been 0.955. The average AUC across the 250 datasets has been 0.881. On the other hand, the performance of Deep Learning has been worse, with an average AUC of 0.831, maximum AUC of 0.899 and minimum AUC of 0.667. Overall both models have a decent predictive performance, and the fact that XGBoost is better that Deep Learning in a classification problem with a low percentage of positives is consistent with the literature (Alonso, Carbo 2020).

Table 3. Model performance.

Model	Average AUC	Maximum AUC	Minimum AUC
XGBoost	0.881	0.955	0.768
Deep Learning	0.831	0.899	0.667

6.2 Accuracy of the explanations. Comparing rankings

How do we determine the goodness of fit of *post hoc* explanations? We know the ordered ranking of importance for each dataset by adding the Overlap, Noise, Corruption and Sparsity of each feature. And we can compute the ordered ranking from both SHAP and Feature Importance. In order to compare the real ordered ranking with the one from the *post hoc* interpretability techniques, we need a quantitative metric. Two of the most used metrics to compare rankings are Kendal Tau and Ranked Based Overlap (RBO). We will use RBO since Kendall Tau has some serious drawbacks, like the fact the it needs the ranking list to be conjoint (and then it prevents us from focusing on the top elements of a list), and it is unweighted, meaning that disagreements in the top of the ranking are as important as disagreements at the bottom of the ranking. However, presumably a potential financial supervisor would be more interested in putting more emphasis on the most important features than on the least relevant ones. The RBO metric on the other hand allows us to decide how much weight the metric should assign to top of the ranking. We will now explain briefly how the RBO works, as proposed by Webber et al (2010).

Let *A* and *B* be two possibly infinity ranking lists, where A_i represents element *i* in list *A.* Let $A_{c:d}$ be the set of elements from position *c* to position *d*.

At each depth *d*, we are interested in the intersection of lists A and B. Such intersection can be defined as:

$$Int_{A,B,d} = A_{1:d} \cap \ B_{1:d}$$

Let´s call agreement at the proportion of A and B that overlap at d:

$$Agree_{A,B,d} = \frac{Int_{A,B,d}}{d}$$

Now that we have defined how to compute the agreement up to certain depth, we call the average overlap (*AO*) at a given overlap depth *K*

$$AO_{A,B,K} = \frac{1}{K}\sum_{d=1}^{K} Agree_d$$

Now, we can use a vector of weights in such a way that we can give more importance to the first elements of the list. This vector of weights should be convergent, otherwise the series would go to infinite, while we want a vector of weights which sum could be represented with a finite number. Let´s consider the following vector of weights:

$$w = \sum_{d=1}^{\infty} \rho^{d-1} = \frac{1}{1-\rho}$$

With 0< ρ <1, where the weight of the *d* element, w_d, is ρ^{d-1}.

Finally, setting w_d = (1- ρ) ρ^{d-1} so that we have $\sum_{d=1}^{\infty} w_d$=1, we obtain our desired Rank Base Order measure (RBO)

$$RBO_{A,B,\rho} = \frac{1}{K}\sum_{d=1}^{K} Agree_d$$

RBO is a measure that is bounded between zero and one. The higher the value, the more similar would be the rankings. With the parameter ρ we can control the importance of the top items of the final ranking. On one extreme, when ρ is equal to zero, then the only item that matters is the first one of the lists. When ρ is equal to one, then the weights are arbitrarily flat. Consequently, we consider a range of ρ

that focus more on the top of the lists. Specifically, we will focus on the performance of the interpretability techniques when ρ is equal to 0.99 and 0.925[15]. The importance of the top variables for these possible values of the parameter is explained in the following **Table 4.** For example, if we choose $\rho = 0.925$, then the top 10 variables weight 75% in the final RBO score. The top 50 variables will weight 99% in the final score, but mainly because the top 10 variables weight already 75%. We will compare the performance of both Feature Importance and SHAP using this metric.

Table 4. Different weights for RBO

Parameter ρ	**% weight of top 10 variables**	**% weight of top 30 variables**	**% weight of top 50 variables**
0.99	27%	53%	67%
0.925	78%	97%	99%

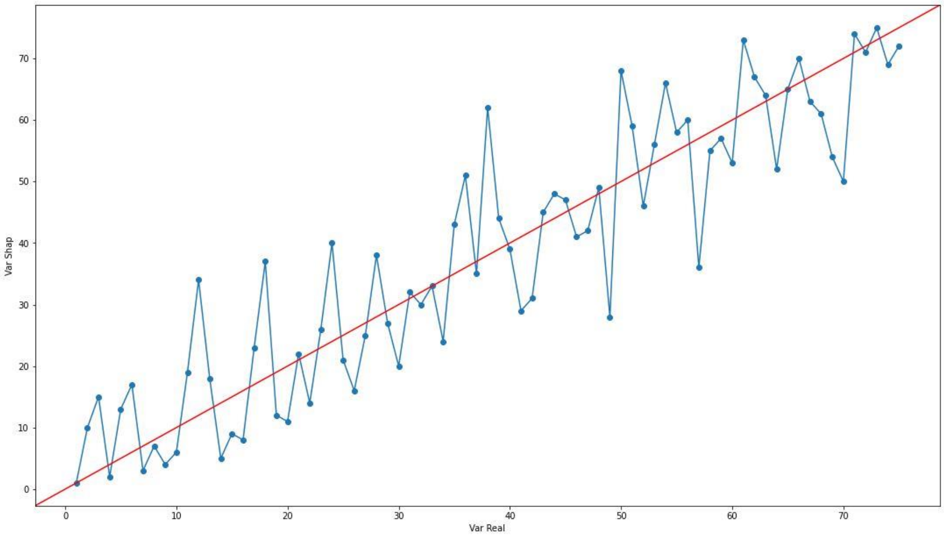
Before assessing the RBO obtained by the two techniques in all the datasets, let's visually see how the interpretability techniques perform when ordering the features in a ranking. In **Figure 3** we plot the original ranking and the ranking of SHAP after XGB was applied to the test data for one of the sets we artificially created (one with 76 features)[16]. On the x-axis we represent the real ranking, and on the vertical axis, the SHAP ranking after applying XGB. If the points on the scatterplot are on the 45-degree line that means that the feature is ranked in the same way in both the actual ranking and the SHAP ranking. It can be seen that the points do not necessarily fall on the 45-degrees' line. While the first feature of the real ranking is correctly identified as the most important feature by the SHAP ranking, the second and third features of the real ranking are identified as the 10th and 15th in the SHAP ranking. Mismatches are to be expected, as we are putting explainability techniques to a severe test. The parameters that we have chosen for Overlap, Noise, etc. make the interpretation task complicated. In any case, it can be seen that as we go through the real ranking, the SHAP ranking follows closely, and the points are near the 45-degree line. We note that despite the mismatches, the order of SHAP is in line with the order of the real ranking.

So graphically it is clear that the SHAP's ranking is similar to the real ranking, at least a priori using one out of 250 datasets as example. But how to measure this similarity? Can we conclude there is an advantage of SHAP over Feature Importance or vice-versa? To this purpose, now we quantify how accurate are the

[15] We have performed exercises manipulating the parameter from 0.995 to 0.9 and the results do not change. Using a parameter ρ of 0.995 would put an importance on the top 10 variables of 18%, while putting a parameter of 0.9 would put an importance to the top 10 variables of 85%. We consider that values of the parameter below 0.9 would put too much importance on the top variables, so we restrict our selection to 0.99 and 0.925.

[16] The characteristics of this particular set are as follows:128.000 observations, 4.9% of positives, and the AUC ROC achieved by the XGB is of 0.89.

Figure 3. Example of SHAP ranking and ground truth ranking.



actual rankings of SHAP and Feature importance using the RBO with different thresholds ρ. With that parameter we can control the importance of the top variables. We use threshold ρ=0.99 to have an idea of the entire ranking, and ρ=0.925 to focus on the top variables. In **Table 5** we summarize the results for XGB and in **Table 6** the results for Deep Learning. In both tables we show the average RBO obtained from using SHAP or Feature Importance over the 250 datasets, and we also show the minimum RBO and the maximum RBO from all those simulations.

Table 5. RBO for XGB

Parameter RBO ρ	**SHAP RBO** **Average** **(Min, Max)**	**Feature importance RBO** **Average** **(Min, Max)**
0.99 (General focus)	0.897 (0.810, 0.939)	0.861 (0.778, 0.918)
0.925 (Focus on top variables)	0.663 (0.337, 0.867)	0.564 (0.297, 0.782)

From this exercise we highlight three key results. First, the accuracy of SHAP is higher than the accuracy of Feature Importance, regardless of the parameter ρ or the ML model, XGBoost or Deep Learning. Differences are higher when we focus on top variables (when ρ is smaller), and the difference between SHAP and Feature

Table 6. RBO for Deep Learning

Parameter RBO ρ	SHAP RBO Average (Min, Max)	Feature importance RBO Average (Min, Max)
0.99 (General focus)	0.840 (0.778, 0.897)	0.836 (0.772, 0.889)
0.925 (Focus on top variables)	0.558 (0.322, 0.727)	0.555 (0.312, 0.724)

Importance is much higher for XGBoost. Actually, for Deep Learning we find that the difference between SHAP and Feature importance is not significant, as we cannot reject the null the hypothesis that the difference in average RBO between SHAP and Feature Importance is different than zero[17]. In any case, we have tried our exercise with different parameters ρ, and we find that the average RBO of SHAP is always higher than the average RBO of Feature Importance, although by small margins. The positive difference between RBO of SHAP and Feature Importance is always significant for XGBoost.

Second, we find that the RBO for both SHAP and Feature Importance under XGBoost is much higher than for Deep Learning. This might be motivated by the fact that XGBoost has a higher AUC than Deep Learning, as we will discuss later on.

And third, we like to highlight that, while RBO never reach the value of one, both the RBO for SHAP and Feature Important are remarkably high for XGBoost. That indicates that the methods, while not being perfect, are capable of a reasonable degree of interpretability of the datasets. They are able to identify the main relevant variables, even though our synthetic datasets have been built in a way that it was complicated (noisy) to capture the relations among variables. We shall remember that due to the fact that we have been completely agnostic about the data generated processes, the correlation among variables in our dataset is low, ranging between -10% to 10%. However, we expect the correlation to play an important role on the results (in the Appendix we perform a small regression analysis using the data obtained from the 250 synthetic sets in which we try to understand which factors could be influencing the performance of the interpretability techniques measured through the RBO metric). We leave for further research to compare SHAP and Feature Importance in a context with higher correlation.

[17] If we perform a traditional t-test to determine if there is a significant difference between average RBO obtained by SHAP and average RBO obtained by Feature Importance, we find that the difference is statistically different at any level of significance after performing XGBoost, but it is not statistically different after performing Deep Learning.

6.3 Accuracy of the explanations. Comparing magnitudes

In addition to comparing the order of the actual ranking with the order of the rankings obtained with SHAP and Feature Importance, we can compare the magnitude of the importance given to the features. Let's imagine that we create a dataset in which the most important variable is much more important than the second most important variable (i.e., its sum of the parameters overlap, noise, sparsity and corruption is much lower). Therefore, in a good explanation of the data, not only should the most important variable be correctly identified as first, but its assigned importance should be much greater than that assigned to the second most important variable, in order to correctly capture the relevance of both variables. In short, it should be expected that there is a negative correlation between the sum of the four parameters that define the variables in the real ranking, and the importance assigned to the features by SHAP and Feature Importance. For example, in **Figure 4,** after applying XGBoost to one of our synthetic datasets, we plot on the vertical axis the importance given by SHAP to each feature, and on the horizontal axis the sum of the parameters of each feature. It could be seen that there is a clear negative correlation, in this case of -0.81. The more negative and the more significant the correlation, the more accurate would be the importance assigned by the interpretability technique. In **Table 7** we show the average correlation between the sum of parameters of the real ranking and the SHAP values for all 250 sets after applying XGBoost, and the average correlation between the sum of parameters of the real ranking and the permutation Feature Importance values. In **Table 8** we show the same, when applying Deep Learning. It could be seen that, as with the analysis made using the RBO, the accuracy of SHAP is higher than the accuracy of permutation Feature Importance, especially for XGBoost.

Figure 4. Example of SHAP magnitudes and ground truth magnitudes

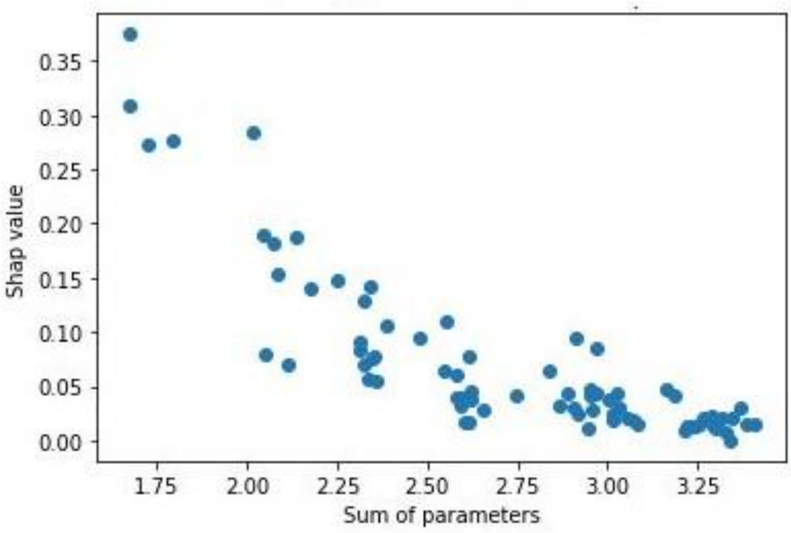


Table 7. XGBoost

Correlation SHAP magnitudes and ground truth magnitudes (Min, Max)	Correlation Feature Importance magnitudes and ground truth magnitudes (Min, Max)
-0.80 (-0.70, -0.87)	-0.62 (-0.48, -0.74)

Table 8. Deep Learning

Correlation SHAP magnitudes and ground truth magnitudes (Min, Max)	Correlation Feature Importance magnitudes and ground truth magnitudes (Min, Max)
-0.58 (-0.35, -0.77)	-0.53 (-0.32, -0.69)

6.4 Sensitivity analysis

Finally, we assess to what extent ML models that achieve higher predictive performance are associated with higher accuracy of their explanations. To this purpose, we explore which could be the effect on the performance of SHAP and Feature Importance of being applied to better or worse predictive models. For a given dataset we create a series of XGBoost models with some random hyper-parameters so that we can achieve "low AUC" models and we can then compare it with "well calibrated" XGBoost models. In order to obtain these random XGBoost, for each set we keep the original and well calibrated XGBoost and we create five additional XGBoost models, in which the number of trees is a random number with maximum value the number of trees of the original XGB model, and in which the depth of tree is a random number with maximum value the number of the trees of the original XGBoost model. Therefore, for each set we have a well calibrated XGBoost (the original one) and five additional XGBoost models with lower performance than the original one. We do this exercise for 20 datasets from our created pool of datasets, so we have 120 XGBoost models in total (20 "well calibrated" and 100 "low AUC"). In **Figure 5** we show a scatter plot where the horizontal axis represents the ratio of the AUC of the "well calibrated" models over the AUC of the "low AUC" models, and in the vertical axis we show the ratio of the RBO (with $\rho = 0.925)$ obtained by SHAP of the "well calibrated" over the RBO of the "low AUC" models. In **Figure 6** we show the same scatter plot but for Feature Importance. We can see that, for both SHAP and Feature Importance, it is obvious that the higher the difference in AUC, the higher the difference in RBO. The correlation is positive and significant. A 1% increase in AUC is associated with a 0.68% increase in overall RBO for SHAP and for a 0.64% for Feature Importance. What happens if we focus on a RBO with ρ = *0.9*, i.e., a RBO that put more emphasis in the 10 top variables? We show the results in **Table 9.** The correlation is not as strong as before. This means that the low AUC models do reasonably well for the top variables. This makes sense. The top variables are important even for "low AUC" models, and the interpretability techniques are able to pick them up in the ranking. In any case, regardless of the kind of RBO that we choose, SHAP seems to be more sensitive to the original performance of the ML model.

Figure 5. Performance of AUC vs SHAP's RBO in XGBoost

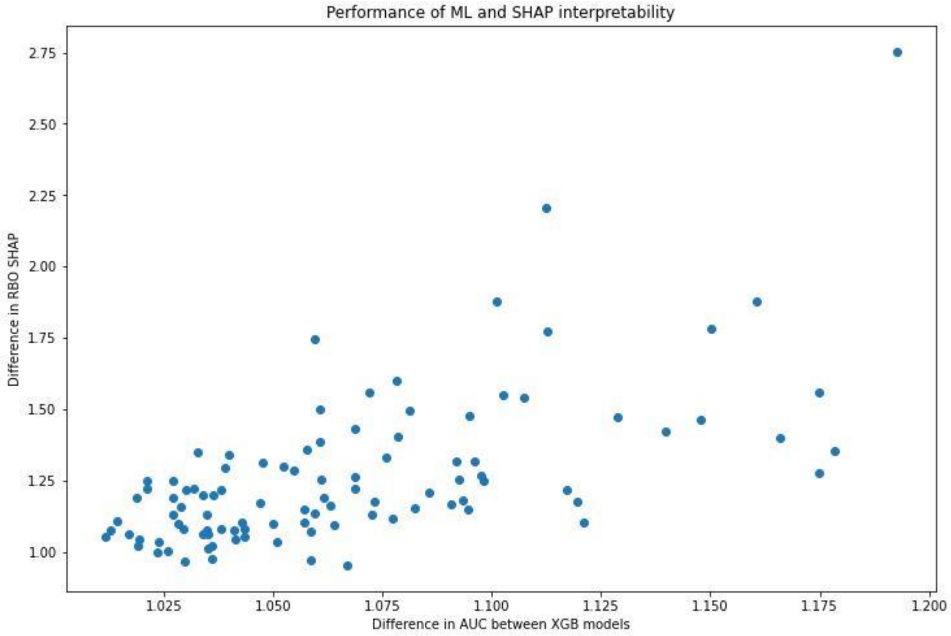


Figure 6. Performance of AUC vs FI's RBO in XGBoost

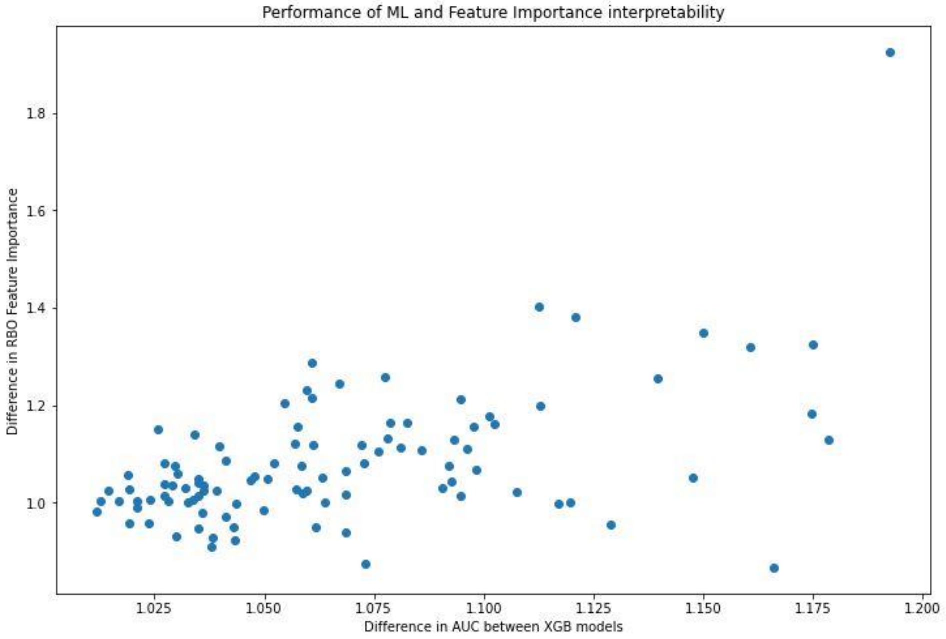


Table 9. Correlation of AUC vs SHAP values for XGBoost.

RBO measure	Correlation improvement AUC, improvement SHAP	Correlation improvement AUC, improvement Feature Importance
0.925	0,68	0,64
0.9	0,55	0,52

Moving from a pure data modeling culture to a more flexible algorithmic approach (as defined in Breiman, 2001b) in the quest for more predictive power would usually be seen as a risk of having less trustworthy explanations. However, the current access to techniques like SHAP and Feature Importance allows increases in predictions together with accuracy of explanations, though this last one at a significant slower pace. Therefore, for the task of model selection, in a world where we might have a sufficient number of good models and the explanations are focused only on the most relevant features, there seems to be room to maneuver to look for better explanatory models without jeopardizing too much predictive capacity.

7. Conclusion

The use of ML in finance is gaining momentum. Its advantage in terms of predictive capacity is undeniable in fields such as credit scoring, robo-advisors, early-warning systems or provisioning (Fernández, 2019), but the application of ML is not exempt from risks (Alonso and Carbo, 2020). Among them, one of the most important is the interpretability of the results, which is particularly relevant in the field of credit scoring. Financial regulators give special importance to achieving and ethical use of AI and ML for different reasons, such as fair lending, discrimination and model governance – see for instance, EBA (2021) for a discussion paper on ML for IRB models; BaFin (2022) for a consultation paper on this topic by the German authority; Akinwumi et al (2021) for a report on AI fair lending policy agenda for the US federal financial regulators; or Dupont et al (2022) on the ACPR Tech Sprint on the explainability of artificial intelligence.

Motivated by this need to manage new model risks, different techniques are being created to help interpret the results of ML models. But this is still an incipient area, and there is no consensus about the reliability of these *post hoc* techniques, nor about how they should be evaluated. In this article we tackle this latter concern, proposing a framework to assess the accuracy of the explanations provided by these techniques, based on the generation of synthetic datasets. The use of synthetic data allows us to define the importance of the variables, and therefore we can assess if the explanation given by the interpretability techniques is in line with the true nature of the data. We apply two ML models, XGBoost and Deep Learning to our synthetic datasets, and we use both SHAP and permutation Feature

Importance to explain the outcome of these models. From our empirical exercise, we observe that the accuracy of SHAP is better than permutation Feature Importance, particularly for XGBoost. In any case, the accuracy of both methods is reasonably high when using XGBoost, considering the difficulty of the task at hand.

We acknowledge that this methodology must be adapted to generate contexts other than those created in this article, such as situations where the correlation between variables is higher, or even generating the ground truth through other methods, like causal models. We acknowledge as well that the methodology should be tested with more post hoc interpretability techniques, capable of performing well in presence of correlation, like Aas et al (2021) or been richer in its definition of the objective functions like the one suggested by Giudici and Raffinetti (2021). Notwithstanding this, our framework is, to the best of our knowledge, the first to assess the reliability of global interpretability techniques, controlling the importance of variables in synthetically generated datasets. We conclude that using synthetic datasets seems a promising research area for financial authorities to work on, as it can be an extremely useful tool to ease the use of innovative ML models while mitigating the risks that are created.

8. References

Aas, K., M. Jullum and A. Løland (2021). "Explaining individual predictions when features are dependent: More accurate approximations to Shapley values", *Artificial Intelligence,* 298(103502).

Akinwumi, M., J. Merrill, L. Rice, K. Saleh and M. Yap (2021). *An AI fair lending policy agenda for the federal financial regulators,* Economic Studies at Brookings.

Albanesi, S., and D. F. Vamossy (2019). *Predicting consumer default: A deep learning approach,* NBER Working Paper 26165.

Alonso, A., and J. M. Carbó (2020). *Machine learning in credit risk: measuring the dilemma between prediction and supervisory cost,* Working Papers, No. 2032, Banco de España.

Alonso, A., and J. M. Carbó (2021). *Understanding the performance of machine learning models to predict credit default: a novel approach for supervisory evaluation,* Working Papers, No. 2105, Banco de España.

Álvarez-Melis, D., and T. S. Jaakkola (2018). *On the robustness of interpretability methods,* arXiv preprint arXiv:1806.08049.

Arimond, A., D. Borth, A. G. Hoepner, M. Klawunn and S. Weisheit (2020). *Neural networks and value at risk,* Michael J. Brennan Irish Finance Working Paper Series Research Paper, No. 20-7.

Ariza-Garzón, M. J., J. Arroyo., A. Caparrini and M. J. Segovia-Vargas (2020). "Explainability of a machine learning granting scoring model in peer-to-peer lending", *IEEE Access,* Vol. 8, pp. 64873-64890.

Arroyo, J., R. Espínola and C. Maté (2011). "Different approaches to forecast interval time series: a comparison in finance", *Computational Economics,* 37(2), pp. 169-191.

Avramov, D., S. Cheng and L. Metzker (2021). *Machine learning versus economic restrictions: Evidence from stock return predictability,* available at SSRN 3450322.

BaFin (2020). *Big Data Meets Artificial Intelligence. Challenges and implications for the supervision and regulation of financial services,* Tech. Rep. Germany, Federal Financial Supervisory Authority-BaFin [Google Scholar].

BaFin (2022). *Machine learning in risk models – Characteristics and supervisory priorities. Responses to the consultation paper*, Rep. Germany, Federal Financial Supervisory Authority-BaFin [Google Scholar].

Barr, B., K. Xu, C. Silva, E. Bertini, R. Reilly, C. B. Bruss and J. D. Wittenbach (2020). *Towards Ground Truth Explainability on Tabular Data,* arXiv preprint arXiv:2007.10532.

Bengio, Y., and Y. LeCun (2007). “Scaling learning algorithms towards AI”, *Large-Scale Kernel Machines,* 34(5), pp. 1-41.

Bholat, D. (2020). *The Impact of Machine Learning and AI on the UK Economy-Conference Overview,* available at SSRN 3602563.

Blattner, L., S. Nelson and J. Spiess (2021). *Unpacking the Black Box: Regulating Algorithmic Decisions,* arXiv preprint arXiv:2110.03443.

Bono, T., K. Croxson and A. Giles (2021). “Algorithmic fairness in credit scoring”, *Oxford Review of Economic Policy,* 37(3), pp. 585-617.

Breiman, L. (2001a). “Random forests”, *Machine Learning,* 45(1), pp. 5-32.

Breiman, L. (2001b). “Statistical Modeling: The Two Cultures”, *Statistical Science,* 16(3), pp. 199-215.

Bücker, M., G. Szepannek, A. Gosiewska and P. Biecek (2022). “Transparency, auditability, and explainability of machine learning models in credit scoring”, *Journal of the Operational Research Society,* 73(1), pp. 70-90.

Cascarino, G., M. Moscatelli and F. Parlapiano (2022). *Explainable Artificial Intelligence: interpreting default forecasting models based on Machine Learning,* Bank of Italy Occasional Papers, No. 674.

Dupont, L., O. Fliche and S. Yang (2020). *Governance of Artificial Intelligence in Finance,* Banque de France.

EBA (2020). *Report on Big Data and Advanced Analytics.*

EBA (2021). *Discussion paper on machine learning for IRB models,* EBA/DP/2021/04, November.

European Commission (2019). *Directorate-General for Communications Networks, Content and Technology, Ethics guidelines for trustworthy AI,* Publications Office, https://data.europa.eu/doi/10.2759/177365.

European Commission (2021). *Proposal for a regulation of the European Parliament and the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts,* EUR-Lex-52021PC0206.

Fernández, A. (2019). "Artificial intelligence in financial services", Analytical Articles, *Economic Bulletin,* 2/2019, Banco de España.

FinRegLab (2022). *Machine Learning Explainability & Fairness: Insights from Consumer Lending,* Empirical White Paper.

Fisher, A., C. Rudin and F. Dominici (2019). “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, 20(177), pp. 1-81.

Frye, C., C. Rowat and I. Feige (2019). *Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability,* arXiv preprint arXiv:1910.06358.

Fuster, A., P. Goldsmith-Pinkham, T. Ramadorai and A. Walther (2022). "Predictably unequal? The effects of machine learning on credit markets", *The Journal of Finance,* 77(1), pp. 5-47.

Ghorbani, A., A. Abid and J. Zou (2019). "Interpretation of neural networks is fragile", in *Proceedings of the AAAI Conference on Artificial Intelligence,* 33(01), July, pp. 3681-3688,

Giudici, P., and E. Raffinetti (2021). "Shapley-Lorenz eXplainable artificial intelligence", *Expert Systems with Applications,* 167(114104).

Goodell, J. W., S. Kumar, W. M. Lim, and D. Pattnaik (2021). "Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis", *Journal of Behavioral and Experimental Finance,* 32(100577).

Gosiewska, A., and P. Biecek (2019). *Uncertainty of Model Explanations for Non-additive Predictive Models,* https://arxiv. org/abs/1903.11420 v1.

Gu, S., B. Kelly and D. Xiu (2021). "Autoencoder asset pricing models", *Journal of Econometrics,* 222(1), pp. 429-450.

Hall, P., B. Cox, S. Dickerson, A. Ravi Kannan, R. Kulkarni and N. Schmidt (2021). "A United States Fair Lending Perspective on Machine Learning", *Frontiers in Artificial Intelligence,* 4(78).

Heskes, T., E. Sijben, I. G. Bucur and T. Claassen (2020). *Causal Shapley Values: Exploiting Causal Knowledge to Explain Individual Predictions of Complex Models,* arXiv preprint arXiv:2011.01625.

Hoepner, A. G., D. McMillan, A. Vivian and C. Wese Simen (2021). "Significance, relevance and explainability in the machine learning age: an econometrics and financial data science perspective", *The European Journal of Finance,* 27(1-2), pp. 1-7.

Institute of International Finance (2018). *Explainability in predictive modelling.*

Institute of International Finance (2019). *Machine Learning: recommendations for policymakers.*

Janzing, D., L. Minorics and P. Blöbaum (2020). "Feature relevance quantification in explainable AI: A causal problem", in *International Conference on Artificial Intelligence and Statistics,* PMLR, June, pp. 2907-2916.

Jullum, M., A. Redelmeier and K. Aas (2021). *Efficient and simple prediction explanations with groupShapley: A practical perspective.*

Königstorfer, F., and S. Thalmann, (2020). " Applications of Artificial Intelligence in commercial banks–A research agenda for behavioral finance*", Journal of Behavioral and Experimental Finance,* 27(100352).

Krishna, S., T. Han, A. Gu, J. Pombra, S. Jabbari, S. Wu and H. Lakkaraju (2022). *The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective,* arXiv preprint arXiv:2202.01602.

Kumar, I. E., *et al.* (2020). "Problems with Shapley-value-based explanations as feature importance measures", International Conference on Machine Learning, PMLR.

Lee, J. W., W. K. Lee and S. Y. Sohn (2021). "Graph convolutional network-based credit default prediction utilizing three types of virtual distances among borrowers", *Expert Systems with Applications,* 168(114411).

Liu, Y., and M. Schumann (2005). "Data mining feature selection for credit scoring models", *Journal of the Operational Research Society,* 56(9), pp. 1099-1108.

Lundberg, S. M., G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair and S. I. Lee (2020). "From local explanations to global understanding with explainable AI for trees", *Nature Machine Intelligence,* 2(1), pp. 56-67.

Lundberg, S. M., G. G. Erion and S. I. Lee (2018). *Consistent individualized feature attribution for tree ensembles,* arXiv preprint arXiv:1802.03888.

Lundberg, S. M., and S. I. Lee (2017). "A unified approach to interpreting model predictions", *Advances in neural information processing systems,* pp. 4765-4774.

Miller, T. (2018). *Explanation in artificial intelligence: insights from social sciences,* arXiv:1706.07269v3.

Miroshnikov, A., K. Kotsiopoulos and A. R. Kannan (2021). *Mutual information-based group explainers with coalition structure for machine learning model explanations*, arXiv preprint arXiv:2102.10878.

Misheva, B. H., J. Osterrieder, A. Hirsa, O. Kulkarni and S. F. Lin (2021). *Explainable AI in Credit Risk Management,* arXiv preprint arXiv:2103.00949.

Mittelstadt, B., C. Russell and S. Wachter (2019). "Explaining explanations in AI", in *Proceedings of the Conference on Fairness, Accountability, and Transparency,* pp. 279-288.

Nag, A. K., and A. Mitra (2002). "Forecasting daily foreign exchange rates using genetically optimized neural networks", *Journal of Forecasting,* 21(7), pp. 501-511.

Ribeiro, M. T., S. Singh and C. Guestrin (2016). "'Why should I trust you?' Explaining the predictions of any classifier", in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* August, pp. 1135-1144.

Ribeiro, M. T., S. Singh and C. Guestrin (2018). " Anchors: High-precision model-agnostic explanations", in *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April.

Rudin, C. (2019). "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead", *Nature Machine Intelligence,* 1(5), pp. 206-215.

Shen, F., X. Zhao, Z. Li, K. Li and Z. Meng (2019). "A novel ensemble classification model based on neural networks and a classifier optimisation technique for imbalanced credit risk evaluation", *Physica A: Statistical Mechanics and its Applications,* 526(121073).

Shrikumar, A., P. Greenside and A. Kundaje (2017). "Learning Important Features Through Propagating Activation Differences", in *Proceedings of the 34th International Conference on Machine Learning.*

Slack, D., S. Hilgard, E. Jia, S. Singh and H. Lakkaraju (2020). "Fooling lime and shap: Adversarial attacks on post hoc explanation methods", in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society,* February, pp. 180-186.

Unceta, I., J. Nin and O. Pujol (2020). "Copying machine learning classifiers", *IEEE Access,* No. 8, pp. 160268-160284.

Visani, G., E. Bagli, F. Chesani, A. Poluzzi and D. Capuzzo (2020). *Statistical stability indices for LIME: obtaining reliable explanations for Machine Learning models,* arXiv preprint arXiv:2001.11757.

Vreš, D., and M. R. Šikonja (2021). *Better sampling in explanation methods can prevent dieselgate-like deception,* arXiv preprint arXiv:2101.11702.

Wachter, S., B. Mittelstadt and C. Russell (2017). " Counterfactual explanations without opening the black box: Automated decisions and the GDPR", *Harv. JL & Tech.,* No. 31, p. 841.

Webber, W., A. Moffat and J. Zobel (2010). " A similarity measure for indefinite rankings", *ACM Transactions on Information Systems (TOIS),* 28(4), pp. 1-38.

Ye, T., and L. Zhang (2019). *Derivatives pricing via machine learning,* Boston University Questrom School of Business Research Paper No. 3352688.

Zhang, Y., K. Song, Y. Sun, S. Tan and M. Udell (2019). *"Why Should You Trust My Explanation?" Understanding Uncertainty in LIME Explanations,* arXiv preprint arXiv:1904.12991.

Appendix

9.1 Creation of variables from other distributions

The process to draw features from distributions other than Normal distribution is very similar to the one shown in section 5.1 for drawing features from a Normal distribution. It is also based on the selection of four parameters, overlap, noise, sparsity and corruption, that would determine the overall importance of the features on the target. Depending on the distributions, there are small differences.

Cauchy and Gamma: Drawing variables from a Cauchy or a Gamma distribution follow the exact same process as for the Normal distribution, being the only difference that instead of choosing the parameters *mean* and *standard deviation*, we choose *location* and *scale for* Cauchy*, and shape and scale for* Gamma respectively.

Beta: Drawing variables from a Beta distribution implies choosing values for parameters α and β.

Distribution if target = 1

X~Beta(α_1, β_1)

Where α_1 = Uniform(2,5), and β_1=min($4 * \alpha_1$, α_1+ $2 * \alpha_1 * (1 - ω)^{1.5}$)

Distribution if target = 0

X~Beta(α_0, β_0)

Where α_0= β_1 and β_0=α_1

Where ω is a parameter such that we obtain he desired overlap

Uniform: Drawing variables from a Uniform distribution will imply choosing values for parameters $min_0, min_1, max_0, max_1$. We assign values to these main parameters using auxiliary parameters: *a, b* and *coincidence.*

a = Uniform (minimum, maximum)

b = Uniform (minimum, maximum)

Coincidence=$\frac{|a-b|}{(2- ω)}$

Where ω is a parameter such that we obtain he desired overlap, and *minimum* and *maximum* are random real values

Distribution if target = 1

X ~ *Uniform (*min_1, max_1*)*

If parameter a is smaller than b, then:

min_1= *a*

max_1= *a + Coincidence*

If parameter a is greater or equal than b, then:

min_1= *a - Coincidence*

max_1= *a*

Distribution if target = 0

X ~ *Uniform (*min_0, max_0*)*

If parameter a is smaller than b, then:

min_0= *b - Coincidence*

max_0= *b*

If parameter a is greater or equal than b, then:

min_0= *b*

max_0= *b + Coincidence*

Categorical: The drawing from categorical variables is slightly different. Instead of drawing from two distributions (one for zeros and one for ones), we draw from three auxiliary distributions, which we call *zeros*, *ones* and *overlapping*.

Let the number of observations belonging to each of the three auxiliary distributions be:

$n_{zeros} = n_{default}$ * (1- ω)

n_{ones} = *(n-* $n_{default}$) * (1- ω)

$n_{overlapping}$= *n-*n_{ones}- n_{zeros}

where n is the total number of observations, $n_{default}$ is the total number of observations with target associated equal to 1, and ω is a parameter such that we obtain he desired overlap. Let *categories* be the number of categories (any number between 2 and 10). Then, we use binomial discrete distributions as follows:

Distribution for ones:

X ~ Binomial (n_{ones}, *categories* , p)

Distribution overlapping observations:

X ~ Binomial ($n_{overlapping}$, *categories* , 0.5)

Distribution for zeros:

X ~ Binomial ($n_{\ zeros}$, *categories* , 1- p)

Where *p* is a random number between 0.1 and 0.2

9.2 Determinants of the accuracy of the explanations

We analyze the determinants of the performance of SHAP and Feature Importance. We will focus on XGBoost, since is the best performing model. We start by considering all the following characteristics of the datasets as possible determinants of the accuracy of SHAP and Feature importance.

- Percentage of ones in the Target
- Percentage of binary and categorical variables
- Number of features
- Number of observations
- Average correlation of variables of top variables among themselves
- Performance of the original machine learning model

Of all these variables, the variable with a higher significant correlation with RBO of SHAP and Feature importance is the number of features. This makes sense, since a higher number of features would complicate the identification of the real order of variables, and thus, the RBO will be lower. Therefore, in order to control for the number of features, we decide to run the following regression: (**Equation 1**):

$$RBO_i = \beta_0 + \beta_1 N\_Feat_i + \beta_2 N\_Rows_i + \beta_3 Target_i + \beta_4 Roc_i + \beta_5 Cor\_top_i + \ \epsilon_i$$

Equation 1

Where *i* stands for each different dataset (up to 250 created datasets), RB0 is one of the measures of RBO for dataset *i* that we analyzed in **Table 4** (We do our analysis with $\rho = 0{,}925$), $NumberFeatures$ is the number of features for set *i,* N_Rows is the number of observations in set *i,* $Target$ is the percentage of positives in set *i,* Roc is AUC-ROC obtained by the XGB model used in set *i,* Cor_top is the correlation among the top 20 variables[18], and ϵ_i is the error term. We record the results in **Table 10**, where we show the estimate of the coefficients of **Equation 1** and in parenthesis the *p value* associated with the null hypothesis that the coefficient is equal to zero**.** As expected, number of features is negative and significant for the performance of SHAP and Feature Importance, but this is due to the definition of RBO. What is more relevant is that Feature Importance seems to be more affected by correlation among top variables. The coefficient of β_6 is more significant for Feature Importance than for SHAP. An increase of 1% in the

[18] We have performed this analysis considering top 10 variables and top 30 variables and the results hold.

correlation among top variables could increase 0.30% the RBO of Feature Importance, an effect that is significant at 1%, while the effect is smaller and less significant for SHAP. The other variables do not seem to be significant, except the percentage of 1's in the Target variable, which is a significant determinant for Feature Importance. Therefore, we can conclude that higher correlation among top variables and higher percentage of 1's in the target helps the interpretability techniques, but it helps particularly to Feature Importance. The rest of the variables are not significant. Since, by construction, correlation among variables in our datasets is low, and correlation among top variables seems to be an important determinant, more research is needed to understand how could correlation affect the performance of both SHAP and Feature Importance.

Table 10

Independent variable	**RBO SHAP**	**RBO Feature Importance**
N_Feat	-0,34 (0,00) ***	-0,43 (0,00) ***
N_Rows	-0,11 (0,36)	0,06 (0,64)
$Target$	0,007 (0,87)	-0,11 (0,03) ***
Roc	0,27 (0,42)	-0,04 (0,90)
$\boldsymbol{Cor\_top}$	**0.16 (0.055) ****	**0,33 (0,00) *****
N	250	250
$\boldsymbol{R^2}$	*0,096*	*0,18*

BANCO DE ESPAÑA PUBLICATIONS

WORKING PAPERS

2110 KLODIANA ISTREFI, FLORENS ODENDAHL and GIULIA SESTIERI: Fed communication on financial stability concerns and monetary policy decisions: revelations from speeches.

2111 YUNUS AKSOY, HENRIQUE S. BASSO and CAROLYN ST AUBYN: Time Variation in Lifecycle Consumption and Income.

2112 JENNIFER PEÑA and ELVIRA PRADES: Price setting in Chile: micro evidence from consumer on-line prices during the social outbreak and Covid-19.

2113 NEZIH GUNER, YULIYA A. KULIKOVA and ARNAU VALLADARES-ESTEBAN: Does the added worker effect matter?

2114 RODOLFO G. CAMPOS and JACOPO TIMINI: Unequal trade, unequal gains: the heterogeneous impact of MERCOSUR.

2115 JAVIER QUINTANA: Import competition, regional divergence, and the rise of the skilled city.

2116 PATRICK MACNAMARA, MYROSLAV PIDKUYKO and RAFFAELE ROSSI: Marginal Tax Changes with Risky Investment.

2117 RODOLFO G. CAMPOS, JACOPO TIMINI and ELENA VIDAL: Structural gravity and trade agreements: does the measurement of domestic trade matter?

2118 ALESSANDRO MELCARNE, JUAN S. MORA-SANGUINETTI and ROK SPRUK: Democracy, technocracy and economic growth: evidence from 20 century Spain.

2119 ÁNGEL ESTRADA and DANIEL SANTABÁRBARA: Recycling carbon tax revenues in Spain. Environmental and economic assessment of selected green reforms.

2120 ALEJANDRO FERNÁNDEZ-CEREZO, BEATRIZ GONZÁLEZ, MARIO IZQUIERDO and ENRIQUE MORAL-BENITO: Firm-level heterogeneity in the impact of the COVID-19 pandemic.

2121 EDUARDO GUTIÉRREZ and CÉSAR MARTÍN MACHUCA: The effect of tariffs on Spanish goods exports.

2122 JACOPO TIMINI: Revisiting the 'Cobden-Chevalier network' trade and welfare effects.

2123 ALEJANDRO BUESA, JAVIER J. PÉREZ and DANIEL SANTABÁRBARA: Awareness of pandemics and the impact of COVID-19.

2124 ALICIA AGUILAR and DIEGO TORRES: The impact of COVID-19 on analysts' sentiment about the banking sector.

2125 SILVIA ALBRIZIO, IVÁN KATARYNIUK, LUIS MOLINA and JAN SCHÄFER: ECB euro liquidity lines.

2126 ANTHONY BALD, ERIC CHYN, JUSTINE HASTINGS and MARGARITA MACHELETT: The causal impact of removing children from abusive and neglectful homes.

2127 IRMA ALONSO, PEDRO SERRANO and ANTONI VAELLO-SEBASTIÀ: The impact of heterogeneous unconventional monetary policies on the expectations of market crashes.

2128 MARÍA T. GONZÁLEZ-PÉREZ: Lessons from estimating the average option-implied volatility term structure for the Spanish banking sector.

2129 SIMÓN A. RELLA, YULIYA A. KULIKOVA, EMMANOUIL T. DERMITZAKIS and FYODOR A. KONDRASHOV: Rates of SARS-COV-2 transmission and vaccination impact the fate of vaccine-resistant strains.

2130 MATÍAS LAMAS and DAVID MARTÍNEZ-MIERA: Sectorial holdings and stock prices: the household-bank nexus.

2131 ALBERT BANAL-ESTAÑOL, ENRIQUE BENITO, DMITRY KHAMETSHIN and JIANXING WEI: Asset encumbrance and bank risk: theory and first evidence from public disclosures in Europe.

2132 ISABEL ARGIMÓN and MARÍA RODRÍGUEZ-MORENO: Business complexity and geographic expansion in banking.

2133 LUIS GUIROLA: Does political polarization affect economic expectations?: Evidence from three decades of cabinet shifts in Europe.

2134 CHRISTIANE BAUMEISTER, DANILO LEIVA-LEÓN and ERIC SIMS: Tracking weekly state-level economic conditions.

2135 SERGI BASCO, DAVID LÓPEZ-RODRÍGUEZ and ENRIQUE MORAL-BENITO: House prices and misallocation: The impact of the collateral channel on productivity.

2136 MANUEL ARELLANO, STÉPHANE BONHOMME, MICOLE DE VERA, LAURA HOSPIDO and SIQI WEI: Income risk inequality: Evidence from Spanish administrative records.

2137 ANGELA ABBATE and DOMINIK THALER: Optimal monetary policy with the risk-taking channel.

2138 MARTA BAŃBURA, DANILO LEIVA-LEÓN and JAN-OLIVER MENZ: Do inflation expectations improve model-based inflation forecasts?

2139 MÁXIMO CAMACHO, MARÍA DOLORES GADEA and ANA GÓMEZ LOSCOS: An automatic algorithm to date the reference cycle of the Spanish economy.

2140 EDUARDO GUTIÉRREZ, AITOR LACUESTA and CÉSAR MARTÍN MACHUCA: Brexit: Trade diversion due to trade policy uncertainty.

2141 JULIO A. CREGO and JULIO GÁLVEZ: Cyclical dependence in market neutral hedge funds.

2142 HERVÉ LE BIHAN, MAGALI MARX and JULIEN MATHERON: Inflation tolerance ranges in the new keynesian model.

2143 DIEGO COMIN, JAVIER QUINTANA, TOM SCHMITZ and ANTONELLA TRIGARI: Measuring TFP: the role of profits, adjustment costs, and capacity utilization.

2144 ROBERTO PASCUAL: Do analysts forecast differently in periods of uncertainty? An empirical analysis of target prices for Spanish banks.

2145 BEATRIZ GONZÁLEZ, GALO NUÑO, DOMINIK THALER and SILVIA ABRIZIO: Firm heterogeneity, capital misallocation and optimal monetary policy.

2201 RYAN BANERJEE and JOSÉ-MARÍA SERENA: Dampening the financial accelerator? Direct lenders and monetary policy.

2202 JUAN S. MORA-SANGUINETTI and ISABEL SOLER: La regulación sectorial en España. Resultados cuantitativos.

2203 JORGE E. GALÁN, MATÍAS LAMAS and RAQUEL VEGAS: Roots and recourse mortgages: handing back the keys.

2204 MONICA MARTINEZ-BRAVO and CARLOS SANZ: Inequality and psychological well-being in times of COVID-19: evidence from Spain.

2205 CORINNA GHIRELLI, DANILO LEIVA-LEÓN and ALBERTO URTASUN: Housing prices in Spain: convergence or decoupling?

2206 MARÍA BRU MUÑOZ: Financial exclusion and sovereign default: The role of official lenders.

2207 RICARDO GIMENO and CLARA I. GONZÁLEZ: The role of a green factor in stock prices. When Fama & French go green.

2208 CARLOS MONTES-GALDÓN and EVA ORTEGA: Skewed SVARs: tracking the structural sources of macroeconomic tail risks.

2209 RODOLFO G. CAMPOS, ILIANA REGGIO and JACOPO TIMINI: Thick borders in Franco's Spain: the costs of a closed economy.

2210 MARCO CELENTANI, MIGUEL GARCÍA-POSADA and FERNANDO GÓMEZ POMAR: Fresh start policies and small business activity: evidence from a natural experiment.

2211 JOSE GARCIA-LOUZAO, LAURA HOSPIDO and ALESSANDRO RUGGIERI: Dual returns to experience.

2212 ADRIÁN CARRO and PATRICIA STUPARIU: Uncertainty, non-linear contagion and the credit quality channel: an application to the Spanish interbank market.

2113 MARIO IZQUIERDO, ENRIQUE MORAL-BENITO, ELVIRA PRADES and JAVIER QUINTANA: The propagation of worldwide sector-specific shocks.

2214 HENRIQUE S. BASSO: Asset holdings, information aggregation in secondary markets and credit cycles.

2215 JOSÉ MANUEL CARBÓ and SERGIO GORJÓN: Application of machine learning models and interpretability techniques to identify the determinants of the price of bitcoin.

2216 LUIS GUIROLA and MARÍA SÁNCHEZ-DOMÍNGUEZ: Childcare constraints on immigrant integration.

2217 ADRIÁN CARRO, MARC HINTERSCHWEIGER, ARZU ULUC and J. DOYNE FARMER: Heterogeneous effects and spillovers of macroprudential policy in an agent-based model of the UK housing market.

2218 STÉPHANE DUPRAZ, HERVÉ LE BIHAN and JULIEN MATHERON: Make-up strategies with finite planning horizons but forward-looking asset prices.

2219 LAURA ÁLVAREZ, MIGUEL GARCÍA-POSADA and SERGIO MAYORDOMO: Distressed firms, zombie firms and zombie lending: a taxonomy.

2220 BLANCA JIMÉNEZ-GARCÍA and JULIO RODRÍGUEZ: A quantification of the evolution of bilateral trade flows once bilateral RTAs are implemented.

2221 SALOMÓN GARCÍA: Mortgage securitization and information frictions in general equilibrium.

2222 ANDRÉS ALONSO and JOSÉ MANUEL CARBÓ: Accuracy of explanations of machine learning models for credit decisions.