Alonso Robisco and Carbó Martínez
Financial Innovation (2022) 8:70
https://doi.org/10.1186/s40854-022-00366-1

Financial Innovation

RESEARCH Open Access

Measuring the model risk-adjusted performance of machine learning algorithms in credit default prediction

Andrés Alonso Robisco and José Manuel Carbó Martínez*

*Correspondence:
jose.carbo@bde.es

Banco de España, Calle de Alcalá, 48, 28014 Madrid, Spain

Abstract

Implementing new machine learning (ML) algorithms for credit default prediction is associated with better predictive performance; however, it also generates new model risks, particularly concerning the supervisory validation process. Recent industry surveys often mention that uncertainty about how supervisors might assess these risks could be a barrier to innovation. In this study, we propose a new framework to quantify model risk-adjustments to compare the performance of several ML methods. To address this challenge, we first harness the internal ratings-based approach to identify up to 13 risk components that we classify into 3 main categories—statistics, technology, and market conduct. Second, to evaluate the importance of each risk category, we collect a series of regulatory documents related to three potential use cases—regulatory capital, credit scoring, or provisioning—and we compute the weight of each category according to the intensity of their mentions, using natural language processing and a risk terminology based on expert knowledge. Finally, we test our framework using popular ML models in credit risk, and a publicly available database, to quantify some proxies of a subset of risk factors that we deem representative. We measure the statistical risk according to the number of hyperparameters and the stability of the predictions. The technological risk is assessed through the transparency of the algorithm and the latency of the ML training method, while the market conduct risk is quantified by the time it takes to run a post hoc technique (SHapley Additive exPlanations) to interpret the output.

Keywords: Artificial intelligence, Machine learning, Credit risk, Interpretability, Bias, Internal ratings based model, IRB model, Natural language processing, NLP

Introduction

While machine learning (ML) algorithms seem to outperform traditional quantitative models in terms of predictive capabilities, especially from the supervisors' perspective, they also pose new risks. Some of them, as cited in the supervisory and regulatory literature (see European Banking Authority 2017a; 2020; Dupont et al. 2020), are associated with interpretability, biases or discrimination, prediction stability, governance or changes in the technological risk profile due to exposure to cyber risk, and dependence on external providers of technological infrastructure. Supervisors face the challenge of

allowing financial institutions and clients to maximize the opportunities stemming from technological progress and financial innovation, while observing the principles of technological neutrality, regulatory compliance, and consumer protection.

According to the International Institute of Finance (2019b), given the steep learning curve of ML as a technology, supervisors struggle to keep up with its fast-moving pace (Wall 2018). Therefore, it is appropriate to continue refining their knowledge about how financial institutions use ML to monitor new model risks as they arise and understand how they might be mitigated.

To overcome this challenge, we suggest a framework that allows the establishment of a bridge between the qualitative list of risk factors usually associated with ML and how to obtain a risk score for each model. As stated in Jung et al. (2019), regulation is not seen as an unjustified barrier, but some firms stress the need for additional guidance on how to interpret it. To the best of our knowledge, this is the first study to evaluate ML algorithms used for credit default prediction via their model risk-adjusted performance.

To build our framework, we must first identify the key components of model risk from the supervisor's perspective. For this purpose, we study the compatibility of ML techniques with the validation process of internal ratings-based (IRB) models to calculate the minimum regulatory capital requirements. Although the IRB approach is restricted to capital requirements, it has an impact beyond this use, considering that the risk components estimated using IRB models (e.g., probability of default) must be aligned with those used internally for any other purpose. We identify 13 components that we refer to as risk factors and classify them into 3 different risk categories, namely, statistics, technology, and market conduct issues. Of these risk factors, we focus on a subset of them to represent the overall risk of the model. For instance, in our exercise, in the *statistics* category, the risk score is computed only on the basis of the stability of the predictions, measured as the standard deviation of the predictions of the models when using different sample sizes, as well as the number of nonzero hyperparameters. For the *technology* category, the score depends on the transparency of the algorithm and the latency (number of seconds) of the training as a proxy for the carbon footprint (Strubell et al. 2019). For the *market conduct* category, the score depends on the latency (number of seconds) of the computation of the SHapley Additive exPlanations (SHAP) for the interpretability of the results. The final risk associated with a particular ML model will depend on the final score of each risk category weighted by the importance or intensity of the regulatory requirements of each category, subject to each use case of the model (capital, credit scoring, or provisioning). To compute these weights, we propose a novel approach based on natural language processing (NLP). First, we collect a series of regulatory texts for each use case, and we calculate the importance of each risk category according to the intensity of mentions in the documents, using our own risk terminology based on expert knowledge, representative of the universe of each risk category. For instance, we find that statistical risks are more important for regulatory capital, while technology and market conduct risks are more important for credit scoring.

We test our framework with five of the most used ML models in the credit risk literature: penalized logistic regression using least absolute shrinkage and selection operator (LASSO), decision tree (CART), random forest, Extreme gradient boosting (XGBoost), and deep learning. Using a public database available on Kaggle.com for credit default

prediction, we compute the potential model risks from validating these models and their potential predictive performance. We find that, for this particular database, XGBoost and random forest are the most efficient risk-adjusted models.

The remainder of this paper is organized as follows. in "Literature review" section, we review the literature on the use of ML for credit default prediction. In "Identifying the risks: compatibility of ML with the IRB validation process" section, we identify the potential limitations of using ML in credit default prediction by reviewing the IRB system under the Basel Accords. In "Quantifying the model risk" section, we demonstrate how our framework measures the potential model risks and benefits from the use of ML by credit institutions from the supervisor's perspective. In "An empirical example" section, we show an empirical example of our framework, and the conclusion is presented in "Conclusion" section.

Literature review

The emerging use of ML in financial systems is transforming society and industry. From hedge funds and commercial banks to contemporary financial technology service providers (Lynn et al. 2019; Kou et al. 2021b), many financial firms today are heavily investing in the acquisition of data science and ML expertise (Wall 2018; Institute of International Finance [IIF] 2019a).

Financial risk analysis is an area where ML is mainly applied by financial intermediaries (see, e.g., Jung et al. 2019 for a survey on UK financial services; Kou et al. (2014) regarding the evaluation of clustering algorithms using credit risk datasets; or Li et al. 2021 for fraud detection). However, within this field, the application with the greatest potential for this technology is credit default prediction (Königstorfer and Thalmann 2020). There is an extensive review of the literature on the predictive gains of ML on this topic. We collect a series of papers that use ML algorithms to predict the impairment of loans, mortgages, retail exposures, corporate loans, or a mixture thereof. In all the studies analyzed, the target variable to predict is the probability of default (PD). To robustly assess the results obtained from different models and samples, we focus on the classification power using the area under the curve—receiver operating characteristic (AUC-ROC) metric, out-of-sample.[1] The ROC curve shows the relationship between the true and false positive rates for all possible classification thresholds. The area below the AUC-ROC curve measures the predictive power of the classifier. Figure 1 presents all the papers included in our literature review in an orderly manner. On the horizontal axis, we divide the papers based on the ML technique used and the a priori algorithmic complexity.[2] First, we distinguish between parametric and nonparametric models. Among the nonparametric models, we consider that deep learning models are more complex than tree-based models because the number of parameters to estimate is higher and their interpretability requires post hoc techniques. Finally, we consider reinforcement

[1] In Butaru et al. (2016) predictive power is measured with the Recall, which represents the percentage of defaulted loans correctly predicted as such. In the case of Cheng and Xiang (2017), predictive power is measured by means of the Kolmogorov–Smirnov statistic, a metric similar to AUC-ROC that measures the degree of separation between the distributions of positives (default) and negatives (non-default).

[2] Similarly, in Gu et al. (2020) the authors estimate the time varying complexity within each ML model by reporting the number of selected components in each model, like for instance, the number of features selected to have nonzero coefficients for Lasso regressions; or the average tree depth for Random forest.

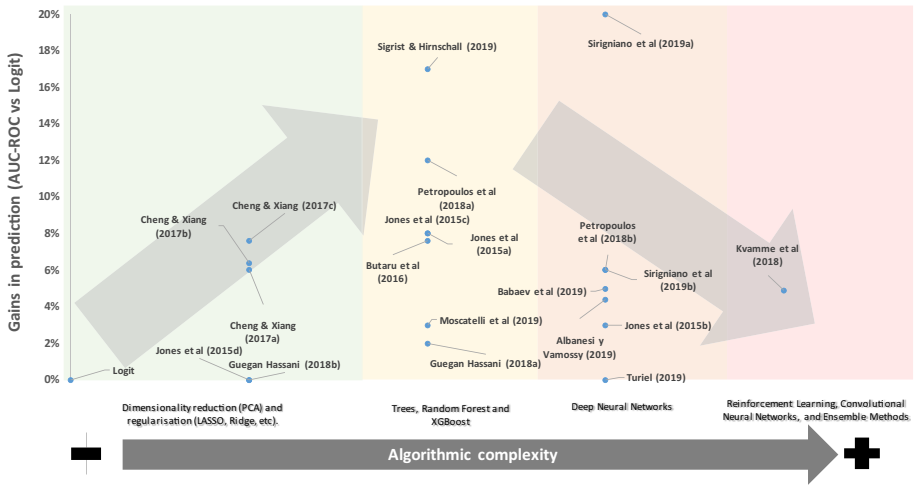

Fig. 1 The dilemma between prediction gains and algorithmic complexity

learning and convolutional nets as the most complex models because the former requires a complicated state/action/reward architecture, while the latter entails a time dimension and thus an extra layer of complexity with respect to deep neural networks.[3] On the vertical axis, we measure the gain in predictive power in terms of AUC-ROC relative to the discriminatory power obtained using a logit model on the same sample.[4] While the sample sizes and the nature of the underlying exposures and ML models differ between studies, they all find that more advanced ML techniques (e.g., random forest and deep learning) give better predictions than traditional statistical models. The predictive gains are very heterogeneous, reaching up to 20%, and do not behave monotonically as we advance toward more algorithmically complex models. We contribute to this literature by providing, in "An empirical example" section, our own estimates of the predictive gains from the use of ML in our empirical example. To the best of our knowledge, we adjust the statistical performance for the first time by measuring the model risk embedded in each algorithm. Our results are consistent with the main findings in the literature. We find gross gains of up to 5% in the AUC-ROC metric from the use of ML, while deep learning models do not necessarily outperform tree-based models, such as random forest or XGBoost, which also turn out to be the most efficient in risk-adjusted terms. The framework we propose in this study aims to measure the performance and risks (i.e., to measure risk-adjusted performance) of different ML models depending on the use case. We do not aim to indicate how to overcome or mitigate their intrinsic risk factors. We include a table in the Appendix Table 13, that summarizes all the papers based on the ML model that they use.

[3] Metrics like the VC dimension (see Vapnik–Chervonenkis 2015) could be used to account for the capacity of the algorithms, when a particular architecture is taken into account. However, for comparison reasons we solely aimed to illustrate the changes in the "structural" algorithmic complexity, in terms of ability to adapt to non-linear, highly dimensional problems. Therefore, changes to this rank could be considered depending on the set of parameters and hyper-parameters considered in each model.

[4] In Butaru et al. (2016) predictive power is measured with the Recall, which represents the percentage of defaulted loans correctly predicted as such. In the case of Cheng and Xiang (2017), predictive power is measured by means of the Kolmogorov–Smirnov statistic, a metric similar to AUC-ROC that measures the degree of separation between the distributions of positives (default) and negatives (non-default).

On the one hand, in the literature on the performance of ML in credit default prediction, there are other potential gains mentioned from the use of this technology. These include positive spillover, such as increasing the financial inclusion of underserved population segments owing to the possibility of using ML together with massive amounts of information, including alternative data, such as the digital footprint of prospective clients, thereby allowing new individuals with little to no financial history to access new credit (Bartlett et al. 2022; Barruetabeña 2020; Dobbie et al. 2021; Huang et al. 2020; Kou et al. 2021a; Fuster et al. 2022).

On the other hand, the literature on the potential model risk from the use of ML for credit scoring is more limited. Several studies have reported negative spillovers if credit scoring models are over-reliant on digital data, which could discriminate against other individuals that lack or decide not to share this sort of personal data (Bazarbash 2019; Jagtiani and Lemieux 2019). Little attention has been paid to the risks and costs that the use of ML by institutions may pose to supervisors. There are studies in the literature that try to explain which factors matter for the governance of ML algorithms on a qualitative basis, such as Dupont et al. (2020) and IIF (2020). However, these articles mention risk factors out of order and do not comprehensively discuss how the risk associated with each model should be classified or measured. Our contribution is that we endeavor to identify the factors that may constitute a component of model risk when validating or evaluating ML models, thus presenting a consistent approach to measure the resulting risk-adjusted performance.

Finally, because we measure how the interpretations of ML model differ, our work is also related to the literature on explainable ML. Notwithstanding that ML models are sometimes considered black box models, a growing and recent field that attempts to elucidate their explanations exists. One of the main approaches toward interpreting an ML model consists of applying post hoc evaluation techniques (or model-agnostic techniques) that explain which features are more relevant to the prediction of a particular model. If we are interested in how they influence a particular prediction, these techniques provide local explainability. Among local explainability techniques, the most popular is probably LIME, as propounded by Ribeiro et al. (2016). However, if we are interested in the relevance of the features for all predictions on a dataset, we use global interpretability techniques. The most important ones are permutation feature importance (Breiman 2001; Fisher et al. 2019) and SHAP (Lundberg and Lee 2017; Lundberg et al. 2020).[5] These two global techniques are based on measuring how the prediction (SHAP) or accuracy (permutation feature importance) of an ML model changes when we permute the values of the input features. The manner in which we permute the values of the features differs depending on the technique used. There is an ongoing debate on the efficacy of these techniques as ML interpretation tools. On the one hand, there are papers that argue that the level of explanation obtained by SHAP could be comparable to that of parametric models such as LASSO and logit (Ariza-Garzón et al. 2020). Furthermore, Albanesi and Vamossy (2019) demonstrates the effectiveness of SHAP for explaining the outcome of ML algorithms in a credit scoring context. Another example

[5] Shap can be used as well as local interpretability technique.

is Moscato et al. (2021), who apply a wide range of interpretability techniques (LIME, Anchors, SHAP, BEEF, and LORE) to random forest and multilayer perceptron models for loan default prediction. They find that LORE has better results, while SHAP is more stable than LIME. On the other hand, other studies affirm that the explanation of the outcome of an ML model by SHAP could be biased or seriously affected by the correlations of the features of the dataset (Mittelstadt et al. 2019; Barr et al. 2020). Arguably, SHAP is still an evolving method, and many authors are making extensions based on this theory (Frye et al. 2019; Heskes et al. 2020; Lundberg et al. 2020).

Identifying the risks: compatibility of ML with the IRB validation process

Our goal is to establish a methodology that allows supervisors to quantify the model risk-adjustment of any given ML method. To do this, we first need to identify and classify all the factors that might constitute a source of model risk from the supervisors' perspective. We do so by harnessing the validation process of IRB systems. Although the IRB approach is restricted to the calculation of minimum capital requirements, it has an impact beyond this use, as the risk components estimated using IRB models (e.g., PD) must be aligned with those used internally for any other purpose.[6] We first identify up to 13 factors that could represent risk from the supervisor's perspective. Thereafter, we classify them into three different categories, namely, statistics, technology, and market conduct.

Credit institutions are responsible for evaluating the performance of IRB systems. However, there are explicit requirements in the Basel Accords vis-à-vis how this process should be undertaken (Heitfield 2005). In this regard, the supervisor's tasks include ensuring that the models are correctly validated. When using the foundation IRB approach, as a general rule, institutions will only have to estimate the PD, while the remaining risk components will be predetermined by the regulation.[7] Once the design of the statistical model has been approved and the estimation is aligned with the supervisor's requirements, the result will be entered into an economic model for computing regulatory capital. This part of the validation is primarily quantitative. In parallel, IRB systems also involve certain issues, such as data privacy and quality, internal reporting, governance, and how to solve problems while operating in a business-as-usual mode. The last part of the validation is mostly qualitative, and it is more dependent on the supervisor's expertise, skills, and preferences. The importance of both issues depends on the purpose of the model (e.g., credit scoring, pricing, or regulatory capital calculation).[8]

In Fig. 2, we list the key risk factors usually mentioned in the regulatory literature, placing them within the IRB validation process and discovering a total of 13 factors.

[6] Article CRE36.60 of the Basel general framework requires that models under the IRB approach be used in the management of the institution's business, requiring alignment between IRB systems and the risk factors used internally in any other field, such as credit scoring, internal risk management or corporate governance.

[7] All the remaining risk factors (e.g.: loss given default or maturity adjustments and credit conversion factors) are defined in the regulation, depending on the type of underlying credit exposure.

[8] In "Quantifying the model risk" section we will quantify both the systemic or common risk factors that matter to the supervisor given any purpose of the predictive model and the idiosyncratic or local factors that refer specifically to each use involving credit default prediction.

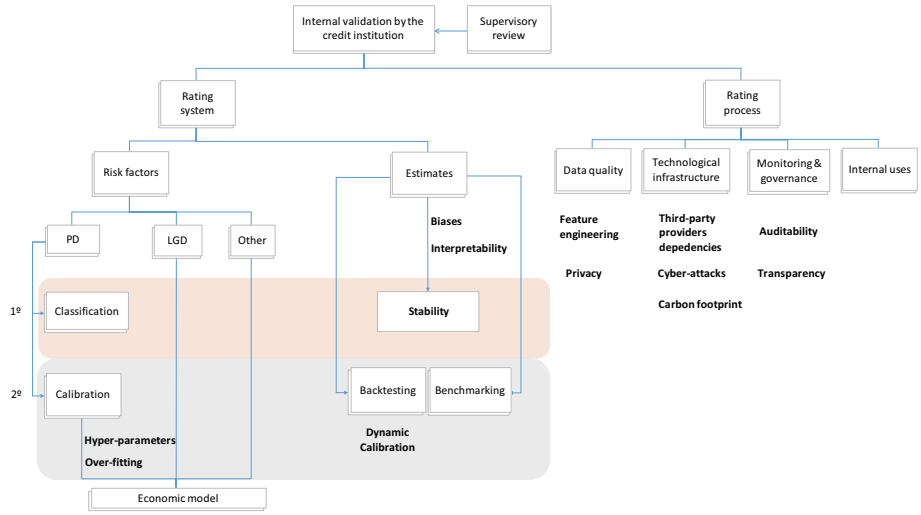

Fig. 2 Validation of IRB systems and its compatibility with ML

The risk factors: a tale of statistics, technology, and market conduct

Having identified these risk factors, we proceed to map them into the scheme of the validation process of IRB systems, as shown in Fig. 2.

Statistics

Statistical model risk may be categorized into different components. Assuming that models rely on usual econometric methods, it allows us to distinguish between estimation risk, associated with the inaccurate estimation of parameters,[9] and misspecification risk. In the banking supervision field, Kerkhof et al. (2010) propose an additional component of statistical model risk, in particular, identification risk that arises when observationally indistinguishable models have different consequences for capital reserves.

To account for this uncertainty, the European Banking Authority (2013) states that "Institutions should include the impact of valuation model risk when assessing the prudent value of its balance sheet. [..] Where possible an institution should quantify model risk by comparing the valuations produced from the full spectrum of modelling and calibration approaches." In this sense, the model risk concept that is considered in this study includes a holistic approach to all the above-mentioned components that affect the uncertainty of pointwise estimations, as is usually carried out following the IRB approach. Specifically, there are a series of factors that could affect the model's estimation and are commonly associated with the use of ML;[10] these include the presence of hyperparameters, the need to preprocess the input data (feature engineering), or the complexity of performing back testing when dynamic calibration is required (for instance, in reinforcement learning models), as it would be infeasible to "freeze" the

[9] See Tarashev (2010) for an application of parameter uncertainty on measuring portfolio credit risk correctly, or Farkas et al. (2020) who derive confidence intervals for the metric Value-at-Risk and evaluate its impact on capital requirements for market risk.

[10] See Babel et al. (2019).

model and evaluate its performance outside the sample, as proposed in Basel.[11] Similarly, the concern about overfitting is always present in the use of ML because of its high flexibility as well as the stability of the classifications (European Banking Authority 2017a) aiming to prevent the procyclicality of the estimates and following the possible migrations observed between credit ratings, such that the robustness of the model throughout an economic cycle can be demonstrated.

Technology

One of the areas associated with algorithmic complexity is the technological requirements necessary for its implementation and maintenance in production while operating normally; this is usually associated with cyber technology and cloud computing. Although cyber losses constitute a small fraction of total operational losses, they account for a significant share of the total operational value-at-risk (Aldasaro et al. 2020). Additionally, higher algorithmic complexity reduces transparency, which increases the cost of financial institutions due to the need to engage external auditors for regulatory compliance (Masciandaro et al. 2020).

One variable to measure the technological burden of ML could be the time required for the computation of the model and its consequent environmental impact, that is, the carbon footprint derived from its electricity consumption (see Alonso and Marqués 2019). Although recent industry research, such as CDP (2020), shows that indirect financed emissions (scope 3) are considered to outweigh both directly produced emissions (scope 1) and indirect consumption-related emissions (scope 2), the truth is that, currently, the vast majority of reported emissions in the financial industry remains only in scopes 1 and 2 (see Moreno and Caminero 2020 for a breakdown of climate-related disclosures by significant Spanish financial institutions). Therefore, when credit underwriting represents a significant share of banks' business models, keeping this technology under environmental evaluation may represent a key element of the current (carbon) net-zero commitments by these institutions.

Another factor that should be considered is the increasing dependence on services provided by third-party providers, such as cloud computing, or those related to fast data processing through the use of GPUs or TPUs (Financial Stability Board 2019)[12] and the potential change in the exposure to cyber risk. The integration of these services with legacy technology is one of the main challenges for institutions, and it is presented as one of the most important obstacles when putting ML models into production (see IIF 2019a). Notably, some institutions are exploring the use of cloud computing providers to avoid

[11] In the calculation of capital add-on for market risk, errors are counted on a daily basis, and depending on whether they amount to one threshold or another, they are counted as green, yellow, or red. This data is abundant and can be used to improve the models. Furthermore, in credit risk, the scarcity of defaults means that a time series usually contains only up to ten annual data points, such that the confidence in the credit risk estimates is significantly lower than in the market risk estimates. To correct this weakness, there is the possibility of counting the errors based on the rating migrations observed for the debtors, since there will be a higher frequency of observed data. In any case, if a bias is identified in the quantification of risk, it must always be adjusted, beyond the estimation's own margin of error, by establishing a margin of conservatism.

[12] The graphics processing unit (GPU) has an advantage over the central processing unit (CPU) when training complex ML models because of its distinct architecture. While the CPU is made up of a small number of complex cores that work sequentially, the GPU is made up of many simple, small cores that are designed to work on multiple tasks simultaneously. The ability to perform multiple calculations in tandem makes the GPU a very efficient tool for using ML. Likewise, the Tensor Processing Unit (TPU) is an application-specific integrated circuit, AI accelerator, developed by Google specifically for machine learning.

such challenges and utilize new data sources, which are particularly relevant to financial authorities because of their potential to impact data privacy.[13]

Market conduct

Similarly, data quality and, in particular, all privacy-related matters are additional aspects to be considered by institutions when applying ML. According to the EBA (2020), one of ML's main limitations concerns data quality. Institutions use their structured data as the main sources of information in predictive models, prioritizing compliance with privacy regulations and the availability of highly reliable data. It follows that, in the context of lending, there is no widespread use of alternative data sources (e.g., information from social networks), while advanced data analytics are used to some extent. To consider all these issues, the system of governance and monitoring of ML models acquires particular relevance, including some aspects such as transparency in the programming of algorithms as well as the auditability of models and their use by different users within the institutions, from the management team to the analysts (see Babel et al. 2019).

Finally, there are two areas, interpretability and control of biases, whose importance transcends statistical or technological evaluation, thereby influencing legal and ethical considerations with repercussions for clients and consumer protection. For instance, the proposal for a regulation of the European Parliament and the Council laying down harmonized rules on artificial intelligence (European Commission 2021) explicitly classifies credit scoring as a "high risk" activity due to its potential economic impact on people's lives. Therefore, from a supervisory perspective, these aspects mostly belong to the field of market conduct. Perhaps, these two additional factors represent ML's most important new model risks with respect to traditional statistical models. Unlike traditional statistical models, most ML models are not inherently interpretable; therefore, we require post hoc interpretation techniques to evaluate their outcomes.[14] However, these interpretation techniques can lead to misleading conclusions (Rudin 2019), and they have limitations regarding controlling for biases (Slack et al. 2020).

In summary, we use the IRB system to place the list of factors that may constitute a source of model risk within its validation process. In Table 1 we group these factors into three categories: (1) statistics, (2) technology, and (3) market conduct.

Purpose of the model

Finally, we must reiterate the circumstance that any model risk-adjustment process should depend on, which is the actual use of the algorithm. We consider three possible use cases: credit scoring and monitoring, regulatory capital, and provisioning (IIF 2019a).[15] For instance, it might be argued that credit institutions usually enjoy greater flexibility when using statistical models for provisioning rather than in other fields, such as regulatory capital, although they must still comply with the regulations and

[13] See European Banking Authority (2017c).

[14] We will introduce in "Quantifying the model risk" section some of the most popular of post hoc explanations techniques. For a complete review of post hoc explanations of machine learning models, see Molnar (2019).

[15] To this end, we disregard the potential use of ML techniques to build a master model by the supervisor to assist with the benchmarking task.

Table 1 Components of ML model risk

Model risk components
Statistics
Stability
Hyper-parameters
Over-fitting
Dynamic calibration
Feature engineering
Technology
Transparency
Carbon-footprint
Third-party providers
Cyber-risk
Market conduct
Privacy
Auditability
Interpretability
Biases

principles of prudence and fair representation.[16] In fact, provisions could be envisaged as an accounting concept governed by the International Accounting Standard Board.[17] Specifically, IFRS 9.B5.5.42 requires "the estimate of expected credit losses from credit exposures to reflect an unbiased and probability-weighted amount that is determined by evaluating a range of possible outcomes […] this may not need to be a complex analysis." Similarly, it has been established that the information used to compute provisions can only be qualitative, although the use of statistical models or rating systems will be occasionally required to incorporate quantitative information (IFRS 9. B 5.5.18). However, granting new loans or credit scoring is a field wherein the use of ML could have a greater impact because of the availability of massive amounts of data that could increase the value provided by more flexible and scalable models (see IIF 2019b). Nonetheless, precisely because of its importance, credit scoring is a field that is subject to special regulation in particular market conduct rules,[18] as is also the case with regulatory capital (see IIF 2019a).

[16] In "An empirical example" section we run an empirical exercise to assess precisely this.

[17] See Appendix 9 from Circular 04/2017, November 27th, Banco de España.

[18] The EBA (2019b) guidelines on loan origination and monitoring determines that when technological innovation is used to grant credit, institutions must, inter alia, (1) manage the risks derived from the use of this technology; (2) evaluate the potential bias that can be introduced into the decision-making process; (3) be able to explain the result ensuring their robustness, traceability, auditability and resilience; (4) document the correct use of the tool; and (5) ensure that the entire management team and analysts understand how it works. Based on the principle of proportionality, the national competent authorities will require documentation on the credit scoring models, and their level of understanding within the entity, both by managers and employees, as well as the technical capacity for their maintenance. Likewise, given the relevance of the use of this technology in the Fintech sector, the ECB (2018) incorporates the evaluation of structural aspects of the governance of the credit granting process, as well as the credit evaluation methodologies and the management of the data. In fact, the use of AI (including ML) for credit scoring is one of the practical cases recently discussed by the Single Supervisory Mechanism (SSM) with the Fintech industry in one of its latest dialogues (May 2019).

Quantifying the model risk

In this section, we propose a methodology to measure the perceived model risk from a supervisory standpoint when validating any ML-based system, which depends on the intrinsic characteristics of the ML algorithm and on the model's intended use. The methodology consists of two phases. First, we discuss the assignment of a score to a given ML model for each of the three risk categories. Second, we discuss how to assign a weight to each risk category for each analyzed use case by means of using the regulatory texts as the supervisor's benchmark, and in the absence of a specific supervisor's loss function or usefulness measure, as summarized in Sarlin (2013). Assuming that policymakers are not cost-ignorant and aim to facilitate innovation, we also acknowledge that they do not disclose their loss function, which describes the trade-off between not allowing a model to be deployed and permitting its use by financial institutions. Therefore, this second phase allows financial institutions to interpret the regulation in the absence of this information from supervisors. Notwithstanding this, the output of our framework can only be read as a ranking of models based on their risk-adjusted performance, and it lacks the knowledge on which threshold value to choose in order to determine the optimal model.

First phase: computing the risk scores

In "Identifying the risks: compatibility of ML with the IRB validation process" section, we identified up to 13 factors that could constitute a source of model risk from the perspective of the supervisor during the validation process. We divided those factors into three categories: statistics, technology, and market conduct. For a given ML model, we assign a score in every category. The score ranges from 1 to 5, where 5 indicates the highest level of risk perceived by the supervisor when deciding whether to approve the model or not.

For each category, we focus on a subset of factors. In the statistics category we have selected "Stability" and "Hyperparameters," in the technology category we have selected "Carbon Footprint" and "Transparency," and in the market conduct category we have selected "Interpretability." We deemed these factors as representative for two reasons. First, because they are highly relevant in their categories, and notably, their evaluation has implications for the other factors. Second, because they can be computed using any empirical database in the absence of prior information on specific characteristics of the financial institution under consideration, while for the remaining factors, we need additional information. In any case, we include a discussion on how we could potentially quantify the remaining factors in the "Appendix" section.

For the statistics category, we counted up to five factors: stability, overfitting, hyperparameters, dynamic calibration, and feature engineering or data pre-processing, having selected "Stability" and "Hyperparameters" as highly representative. The first one, if we count the number of mentions of "Stability" in the regulatory documents we have collected (see "Second phase: assigning weights to risk categories" section for further details), has 6.4 mentions per document, much more than hyperparameters (0.2 times per document), overfitting (0.16), dynamic calibration (0.02), and feature engineering (0.33). Therefore, we consider it as a preferred candidate for quantification. There are different methods to measure the stability of a prediction. According to Dupont et al.

(2020), the stability of the predictions can be understood as the absence of drift over time, the generalization power when confronted with new data, and the absence of instability issues during retraining. Either of these three descriptions works for our purpose. However, because not every dataset contains temporal dimensions, we will focus on the second and third definitions, measuring the standard deviation of the predictions of the models when using different sample sizes, in particular, retraining the models 100 times, each with different sample sizes, always keeping it within the range between 60 and 100% of the training set. Thus, we can test the stability of the predictions. Additionally, we have also considered it relevant to count the number of hyperparameters, as this factor is used for calculations once the ML model has been trained and validated, and it might add valuable information on the remaining statistical factors, as it could be argued that all of them would somehow be linked to the size of the models. Finally, we show in the Appendix on pages 24 to 26 a suggestion on how the "Dynamic Calibration" could be computed, along with suggestions on how to calculate "Overfitting" and "Feature Engineering."

For the technology category, we counted up to four factors: transparency, carbon footprint, third-party providers, and cyber-attacks. We have selected “Carbon Footprint” as the most representative term because as we will explain below, with its calculation we could get an approximation of the magnitude of the remaining factors in this category, as all of them rely on the required computer power to run the algorithms. We measure it through the model training latency, that is, how long in seconds it takes to train the model. For the sake of completeness, we also consider the “Transparency” of the models by first distinguishing between parametric and nonparametric models. Among the nonparametric models, we consider that deep learning models are more complex than tree-based models because the number of parameters to estimate is higher. However, while the calculation of the risk factors "third-party providers" and "Cyber risks" will depend on the specific dataset and the particular circumstances of the financial institution, they share with “Carbon Footprint” the underlying importance of the computer requirements. The longer it takes to train a model, the more likely it is that the user needs some cloud service (thus increasing the risk of the third-party provider), and the more prone processes can be to cyber-attacks. We further discuss how we could potentially calculate these two technological risk factors in the “Appendix” section.

For the market conduct category, we counted up to four factors: privacy, auditability, interpretability, and bias. We have selected “Interpretability” as the most relevant term in the category. We assign a score for “Interpretability” to a given ML model using the latency (number of seconds) of the computation of the SHAP values for the explanation of the results. SHAP (Lundberg and Lee 2017) is an interpretability technique that allows for the global and local interpretation of any model. There is an ongoing debate on the efficacy of SHAP as an ML interpretation tool (see the “Literature review” section). In any case, SHAP, along with permutation feature importance, is one of the most popular options for interpreting complex ML models. If we use an ML model to predict a target variable based on a set of features, SHAP can rank all the features depending on their importance in the final prediction. The ranking of a given feature depends on its contribution to the predicted result in a particular observation, compared with the average prediction. These contributions are known as Shapley values. Once the Shapely

values for each feature and each instance have been computed, we can obtain the overall SHAP importance by adding them together.[19]

Both SHAP and permutation feature importance have advantages and limitations (Molnar 2019), but it seems that SHAP, despite its drawbacks, is gaining popularity as the leading global interpretation technique (Hall et al. 2021). While SHAP can deliver a clear ranking among features, it is computationally expensive. Therefore, we consider that the time and resources required to execute SHAP could be a good proxy for how easy it is to interpret a model. To the best of our knowledge, this is the first attempt to use a standard explainable artificial intelligence technique, common in the academic literature, in the supervisory model validation process to provide assistance to market participants interpreting the regulation.

To calculate the remaining risk factors in the market conduct category, we need information that may not be included in all the databases. For example, to calculate "Bias," we would need information from features' labels, and these features should contain sensitive information. It is not always possible to access this type of information. In any case, we consider that our "Interpretability" score can be a reference for the rest of the risk factors in the market conduct category, as understanding how easy it is to interpret a model could be a proxy itself for how easy it would be to detect biases, the probability of breaching privacy rules, and the traceability of the model. We discuss how we could potentially calculate "Bias," "Privacy," and "Auditability" in the "Appendix" section.

Second phase: assigning weights to risk categories

How important should each category be when determining the risk in the ML model? In the second phase, we assign weights to each category depending on the purpose of the model. We consider three possible use cases: regulatory capital, credit scoring and monitoring, and provisioning. For example, it could be argued that interpretability should matter more if the purpose of the model is to grant new credit, but less if the purpose is to compute provisions on outstanding loans.

The ideal way to carry out this exercise would be to know the real preferences of supervisors when assessing these three risks. However, these preferences are unknown or ambiguous at best. Therefore, we propose a method to estimate weights for the risk categories that reflect their actual regulatory importance in each possible model use. For each risk category, we create terminology with a list of representative words (plus their lemmatization) associated with the category based on expert knowledge. The lists are provided in the Appendix Table 11. We include 54 words for the statistics category, 30 words for the technology category, and 51 words for the market conduct category. Although this list is not exhaustive, we aim to obtain a representative sample of words for each category. Thereafter, we select a sample of regulatory documents referring to each of the three possible use cases (capital, credit scoring, provisioning) and a set of documents that we assume belong to a common area, as they refer to all possible uses

[19] To compute the Shapley value of a feature of interest for a given instance, we start by considering all possible coalitions of features that exclude the feature of interest, including the empty set. For all those different coalitions, we compute the difference in the predicted outcome with and without the feature of interest. The Shapley value of the feature of interest is the weighted average of the differences in the predictions among all coalitions. When the number of features is high, the number of coalitions can be almost impossible to manage, and that is why there are several techniques for approximating the results.

of the model. A methodical search for each purpose of the models has been conducted in the repository of the European Banking Authority, Basel Committee on Banking Supervision, and European Central Bank. Further documents referring in particular to artificial intelligence have been included, as a recent proposal from the European Commission ("Artificial Intelligence Act") or working papers from central banks (e.g., Bracke et al. 2019; Dupont et al. 2020). Finally, some guidelines from auditors or international institutions, such as the World Bank, have been selected because of their reliability. The selection of texts includes binding requirements and non-binding recommendations. This list is included in Table 12 of the Appendix. The weight of the risk categories in each use of the model depends on the number of times the terms of the category are mentioned in the model's use documents.

Consequently, we have three risk categories (statistics, technology, and market conduct), and four sets of regulatory documents (capital, credit scoring, provisions, and common area). Let us call $H_{i,j,k}$ the percentage of words from risk category i over all words of document k belonging to a set of documents j:

$$H_{i,j,k} = W_{i,j,k}/N_k$$

where $W_{i,j,k}$ refers to the number of times a word from risk category i appears in document k from the set of documents j, and N_k is the total number of words in document k. We call $\overline{H}_{i,common}$ the overall frequency of risk category i in the common area set of documents:

$$\overline{H}_{i,common} = \sum_{k}^{M_{common}} H_{i,common,k}/M_{common}$$

where M_{common} refers to the number of documents collected from a common area. Finally, we compute the overall relative frequency of risk category i in a model using j as:

$$\overline{H}_{i,j} = \sum_{k}^{M_j} H_{i,j,k}/M_j + \overline{H}_{common}$$

where M_j is the total number of documents analyzed for model use j.

In Fig. 3 and Table 2, we compare the overall relative frequency of words for each risk category: capital, credit scoring, and provisions. Our results show that the statistics risk category is more important for capital requirements, whereas technology and market conduct risks are more important for credit scoring. Another key insight is that while capital and credit scoring have approximately 14% of the percentage of mentions of the three risk categories, provisions has only 12%. This may indicate that provisioning is the area with the lowest perceived supervisory model risk.

Are these differences significant? In Table 3 we check if the average intensity of the categories is significantly different across the model´s uses. We perform a t-test based on the following T *statistic*, built under the null hypothesis that two means of the populations are equal.

Table 2 Overall relative frequency of categories in model uses

	Statistics	Technology	Market conduct	Total
Capital	7.90%	2.50%	3.75%	14.15%
Credit scoring	6.41%	3.09%	5.13%	14.63%
Provisions	6.38%	2.36%	3.90%	12.64%

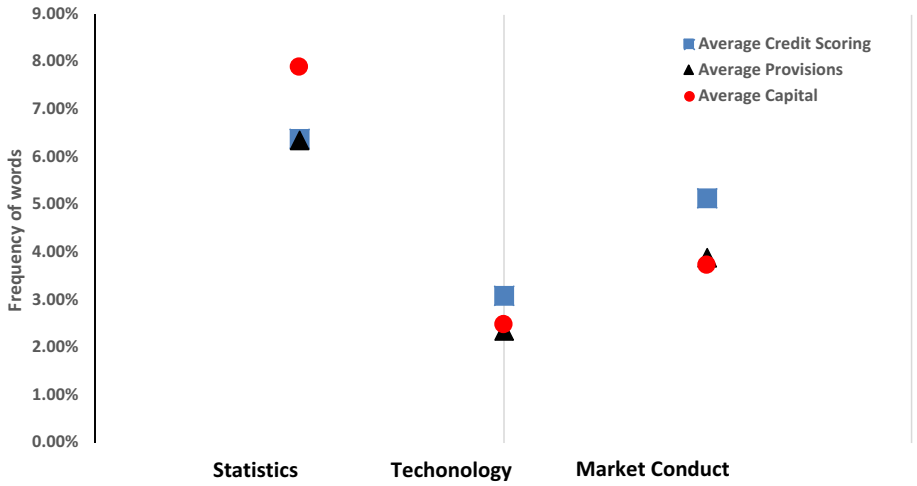

Fig. 3 Comparison between different model purposes

$$Tstatistic = \frac{Mean_1 - Mean_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

where $Mean_1$ and $Mean_2$ are the mean values of each sample, s_1 and s_2 are the standard deviations of the two samples, n_1 and n_2 are the sample sizes of the two samples, and *n-1* are the degrees of freedom. With the T statistic value and degrees of freedom, we can compute the corresponding p-values of every possible comparison of means. The p values are shown in Table 3. It can be seen that the differences between capital and credit scoring are significant for statistics and market conduct, the differences between capital and provisions are significant for statistics and technology, and the differences for credit scoring and provisions are significant for technology and market conduct. We recognize that the number of mentions of these risk factors in the selected documents may be insufficient to reflect their importance. Therefore, we perform a robustness analysis, in which, instead of counting the number of times the risk terms appear in the documents, we count the number of negative words that surround each mention of those terms. In this way, we can capture the intensity with which the terms appear (counting the times they appear in the documents) and the tone toward them (counting the negative words that surround them). The results are presented in the Appendix, "Robustness exercise: sentiment analysis" section, in this document, and they complement the results of the benchmark exercise. We leave the construction of a more complex sentiment analysis on

Table 3 Hypothesis contrast (*p* value) about difference in mean values

	Statistics	Technology	Market conduct
Capital versus credit scoring	**0.0181**	0.1897	**0.003**
Capital versus provisions	**0.0133**	**0.0180**	0.5618
Credit scoring versus provisions	0.9404	**0.1097**	**0.0128**

Bold means that differences between model uses are statistically significant at 10% of significance level

the perception of statistical, market conduct, and technological risks in different uses of the model for further investigation.

Once we have computed the score of the ML models for each risk category and the relative importance of risk categories for each possible use case, we can define the supervisory model risk of the ML algorithm m for use j as follows:

$$Supervisory\ model\ risk_{m,j} = \sum_{i}^{3} \overline{H}_{i,j} Z_{i,m}$$

where $Z_{i,m}$ is the darkness score (after the black box definition surrounding ML) of model m for risk category i, *and* $\overline{H}_{i,j}$ is the overall relative frequency of risk category i in a model use j.

For instance, the darkness Z of model m in the category i = statistics, should capture the ordinal importance for the model validator of each risk factor (i.e., stability of predictions, number of hyperparameters, overfitting, dynamic calibration, and feature engineering) between all models being evaluated. As every risk category will be calculated based on heterogeneous proxy variables with different measurement scales, we propose to leave at the discretion of the model validator (e.g., the supervisor) how to aggregate them into a single score $Z_{i,m}$. For illustrative reasons, we will assume in Table 4 a discrete choice between the range [1,5] for each $Z_{i,m}$. Assuming that we are comparing five different models, the darkest (riskiest) possible model would have a maximum value of 5 in every category. Therefore, we can compute the darkness score of any given model, normalizing the respective $Z_{i,m}$ using the maximum score, and thus obtain a relative valuation of the model riskiness.

Supervisory model riskiness for capital = (2 * 7.9 + 4 * 2.5 + 3 * 3.75)/(5 * 7.9 + 5 * 2.5 + 5 * 3.75) = 0.523.

Supervisory model riskiness for credit scoring = (2 * 6.41 + 4 * 3.09 + 3 * 5.13)/(5 * 6.41 + 5 * 3.09 + 5 * 5.13) = 0.554.

Supervisory model riskiness for provisioning = (2 * 6.38 + 4 * 2.36 + 3 * 3.90)/(5 * 6.38 + 5 * 2.36 + 5 * 3.90) = 0.536.

The construction of the supervisory model riskiness is a multidisciplinary task that aims to quantify the requirements to comply with the regulation. While expert knowledge of statistics and technology is required in the first phase to open the algorithmic black box, an in-depth understanding of financial supervision is key in the second phase to break down how the model fits into the regulatory requirements. Our scorecard offers a structured methodology for estimating this exercise from a neutral standpoint,

Table 4 Example scorecard of a model risk assessment

Risk category	Darkness Score	% Regulatory capital	% Credit scoring	% Provisions
Statistics	2	7.9%	6.41%	6.38%
Technology	4	2.5%	3.09%	2.36%
Conduct	3	3.75%	5.13%	3.9%

identifying for the first time a set of risk categories and their corresponding risk components that may be quantified using some proxy variables. Indeed, we assume no preferences from the supervisor or model validator for the weight of each risk category, which is estimated directly from the regulatory texts. This will allow supervisors to provide credit institutions with a neutral assessment of ML as a technology to be used in predictive models in a standardized format.[20] Notwithstanding this, further research is needed to investigate different alternatives to aggregate the identified proxy variables into each score $Z_{i,m}$.

An empirical example

In this Section, we propose an empirical example of the framework using a database available at Kaggle.com, called "*Give me some credit*". It contains data on 120,000 granted loans. For each loan, a binary variable indicates whether the loan has defaulted. Additionally, 11 characteristics are known for each loan: borrower age, debt ratio, number of existing loans, monthly income, number of open credit lines, number of revolving credit accounts, number of real estate loans, number of dependents, and the number of times the borrower has been 30, 60, and 90 days past due. To capture nonlinear relationships, we include the square of these characteristics as additional variables until we have a total of 22 explanatory variables. We apply the framework to five of the models that appear most frequently in the academic literature on credit default prediction: penalized logistic regression via LASSO, decision tree, random forest, XGBoost, and deep neural network. The deep learning model used in our study is an artificial neural network in which we consider the possibility of having three to six hidden layers. Therefore, we use a multilayer perceptron model, where the number of hidden layers and nodes in each layer has been chosen according to proper cross-validation to obtain the largest AUC in the validation sample. In particular, we divide our data into training (80%) and testing (20%) sets. We choose the hyperparameters for each model that maximize the AUC-ROC out-of-sample through a fivefold cross-validation. The hyperparameters of each model are as follows: the depth of trees for CART (7), the depth of trees and the number of trees for random forest (20 and 100, respectively), the depth of trees and the number of trees for XGBoost (5 and 40), and finally the optimal number of hidden layers (3) and nodes (300, 200, and 100), while activation functions would be rectified linear unit for the hidden layers and sigmoid for the output layer, and the optimization method is Adam.

[20] This challenge is not the first of its kind to occur in financial supervision. For instance, when assessing the capital requirements for market risk, supervisors have agreed on the methodology (i.e.: Value-at-Risk, or Expected Shortfall) but discretionally assume that in the back testing the surcharges due to deviations from the estimations are calculated simply by using a traffic-light test, BIS (1996).

Table 5 Scorecard phase 1: "measuring the darkness of the algorithmic black-boxes"

		Lasso	CART	RF	XGBoost	MLP
Statistics	Stability (SD)	0.001	0.005	0.003	0.001	0.002
	Number hyper-parameters	0	1	2	2	+5
	Darkness Score	1	3	3	2	5
Technology	Latency training	3.92 s	0.67 s	19.25 s	1.29 s	31.42 s
	Transparency	1	2	3	3	5
	Darkness Score	1	1	3	2	5
Market conduct	Latency SHAP	0	0	60 s	8 s	2000 s
	Darkness Score	1	1	2	2	3

As mentioned before, we will use a subset of five risk factors that we deem representative of each risk category to showcase this methodology. In "First phase: computing the risk scores" section, we provide a comprehensive explanation of why we choose these five proxies. In summary, we selected factors that could be representative of their corresponding categories, which could be estimated using a common credit database and in the absence of prior information on specific characteristics of the financial institution under consideration. In the "Appendix" section, we suggest a method for quantifying the remaining components of model risk. We leave for future research a deeper discussion on the calculation of these factors.

Our results for the scores of the five aforementioned models are shown in Table 5 and in Fig. 4 we map the assessed riskiness of each model per risk factor into a scorecard.

Model risk

We calculate the score for statistics based on the models' stability and the number of hyperparameters, as these two factors stand out as highly relevant in this category (see "First phase: computing the risk scores" section for a detailed explanation of this). For the stability of the predictions, we show the standard deviation in the AUC-ROC for 100 simulations with different train–test partitions. It can be seen that the models with the highest standard deviations are deep learning and CART, whereas the models with the lowest standard deviations are LASSO and XGBoost. Computing the number of hyperparameters is straightforward, that is, the LASSO model with the lowest need for hyperparameters and the deep learning model with the highest need. Considering the values for these factors, we assign a score of 1 to LASSO, a score of 2 to XGBoost, a score of 3 to both CART and random forest, and a score of 5 to deep learning. As mentioned in "Quantifying the model risk" section, we aggregate the estimations of each proxy variable in each risk category into a single score using expert knowledge.[21]

The technology score will depend only on the models' transparency and on the latency of the training, measured in seconds (see "First phase: computing the risk scores" section for an explanation of why we focus on those two factors). For the model´s transparency, following our explanation in "Literature review" section, LASSO falls into the category of parametric models, CART into nonparametric models, random forest and XGBoost

[21] c.f. footnote 20.

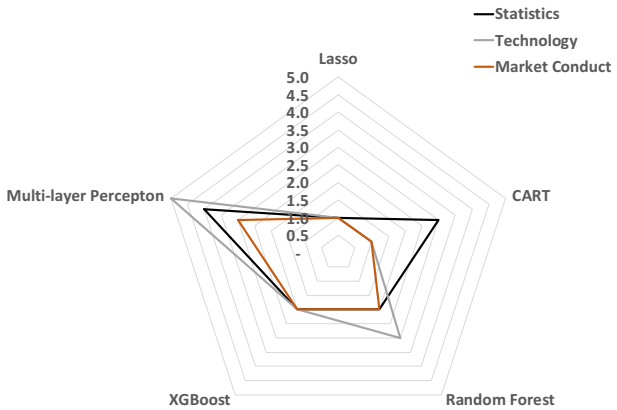

Fig. 4 Breaking down the models' riskiness

into the category of nonparametric ensemble models, and finally, deep learning falls into the most complex category. Considering the latency of training, we assign a score of 1 to LASSO and CART, 2 to XGBoost, 3 to random forest, and a score of 5 to deep learning.

The score for market conduct will be calculated based on the latency of the SHAP method measured in seconds, as we find this as a good proxy for the feasibility of interpreting a black box model, and therefore how easy it would be to spot biases and audit them (IIF 2019c), capturing the interconnectedness between all risk market conduct risk factors (see "First phase: computing the risk scores" section for a detailed explanation).[22] As mentioned in "First phase: computing the risk scores" section, for a given ML model, SHAP is a technique that allows us to rank the features according to their contribution to the predicted result for a particular instance, compared to the average prediction of the entire dataset. These contributions can be added to obtain the final importance of the features (for more details, see "First phase: computing the risk scores" section). Therefore, SHAP allows us to interpret the decisions and predictions of any ML model. However, evaluating SHAP contributions is computationally expensive. Therefore, we consider the time (latency) required to implement SHAP to be a good signal of how easy it is to interpret an ML model. Because interpretability is one of the main issues in market conduct risk, we consider that SHAP latency can be used as a good proxy for this category. While we are interested in the time it takes to compute the SHAP values, we include in Figs. 6 and 8 of the Appendix, the results from the application of SHAP to the ML models.

Because LASSO and CART are interpretable models, we assign them a score of (1). For XGBoost and random forest, we assign a score of (2) because the latency of the SHAP method is relatively low. We assign a score of (3) to deep learning because of the long time it takes to calculate its SHAP values.

Once we assign the score to each category for the five models, we use the weighting for different use cases to compute the overall model risk, as shown in Table 6. Independent of the purpose of the model, LASSO has the lowest model risk and deep learning has

[22] In any case, as mentioned before, we suggest a way to compute each remaining factor separately in the "Appendix" section.

Table 6 Scorecard phase 2: computing the supervisory riskiness for regulatory capital

	Lasso	CART	RF	XGBoost	MLP
Statistics	1	3	3	2	4
Capital: 7.9%					
Credit: 6.41%					
Provisions: 6.38%					
Technology	1	1	3	2	5
Capital: 2.5%					
Credit: 3.09%					
Provisions: 2.36%					
Market conduct	1	1	2	2	3
Capital: 3.75%					
Credit: 5.13%					
Provisions: 3.9%					
Supervisory model risk capital	14.15	29.95	38.7	28.3	55.35
Supervisory model risk credit	14.63	27.45	38.76	29.26	56.48
Supervisory model risk provisions	12.64	25.4	34.02	25.28	49,02

the highest. CART has a higher model risk than XGBoost for capital (owing to its poor stability) and a lower perceived model risk when using it for provisioning and credit. Provisions is the purpose with less model risk from a supervisory perspective, especially for deep learning, thus reflecting the flexible nature of statistical modeling regulation in this area. The net amounts of regulatory requirements for credit scoring and capital are very similar. While the statistical requirements for credit scoring are lower than those for regulatory capital (6.41% vs. 7.9% frequency of terms' occurrence), this is offset by the higher level of requirements regarding market conduct issues in this area, which has a frequency of occurrence of 5.13% in our texts, compared to 3.75%, respectively. However, the level of implementation observed in the industry nowadays indicates that credit scoring is a field in which ML is being deployed more actively (see IIF 2019a). This could imply that the statistical requirement represents a barrier to the introduction of ML in the short term, whereas the need for interpretable results of ML (associated with market conduct requirements) represents a challenge in the medium term.

Gross predictive performance

There are different methods to compute the prediction gains of an ML model. As we showed before, one of the most popular measures, and the one we will use, is to compare the statistical performance of the models using the AUC-ROC metric.[23] The results are presented in Table 7, where we show the increase in the AUC-ROC of each model with respect to the one achieved by logit.[24] XGBoost is the model with the largest gain in terms of prediction, approximately 5% in the AUC-ROC, followed by the random forest with 4%, and deep learning with 1.7%. CART and LASSO have 0.4% and 0.2% gains on average, respectively, compared with logit. This ranking based on AUC-ROC gains is in line with the results from the literature reviewed in Fig. 1, where the models with the

[23] Other measures include GINI, Kolmogorov–Smirnov or the confusion matrix.

[24] The average AUC-ROC for Logit is around 80%

Table 7 Results of the estimated AUC-ROC using "Give Me Some Credit" dataset

	Logit	Lasso	CART	Random forest	XGBoost	MLP
AUC-ROC	80%	80.2%	80.4%	84.2%	85.3%	81.7%
AUC-ROC	79.3%	79.6%	79.7%	83.7%	84.8%	81%
95% interval	80.7%	80.8%	81.2%	84.8%	85.8%	82.3%
AUC-ROC (difference with logit)	–	0.2%	0.4%	4.2%	5.3%	1.7%

highest prediction gains are XGBoost and random forest, and deep learning does not necessarily predict better than the other algorithms. Because our dataset lacks a time dimension, it is not possible to calculate the predictive performance using reinforcement learning algorithms or convolutional neural networks.[25] Although this particular ranking among ML models may change when other databases are used, our exercise provides a quantification of the predictive gains that will assist us in the challenge of measuring the risk-adjusted performance of these models.

Model risk-adjusted predictive performance

Once we know the model risk of the algorithms (Table 6) and their gross predictive performance (Table 7), we plot their risk-adjusted performances in Fig. 5. CART shows a moderate increase in predictive performance with respect to logit, but it does show a considerable increase in riskiness. However, while random forest and XGBoost have similar levels of riskiness compared to CART, they display a better predictive performance. In any case, XGBoost clearly outperforms the other models as the most risk-efficient. This is driven by the good results of this model in terms of the computational power required (approximated by the latency), comprehensive statistical nature (stability of predictions), and interpretability (quantified using computability of SHAP values), which allows it to be a well-balanced solution for the regulatory requirements for all purposes, compared to the rest of the ML models. In any case, this exercise should be complemented with more advanced calculations of the performance of the ML models as, for instance, it might be argued that the benefits of being able to classify better credit defaults are more important in credit scoring than in capital, or that the calibration of the models (which might behave differently depending on the level of PD) will be crucial for computing provisions (expected loss, i.e., higher PD) or capital (unexpected loss, i.e., lower PD). Therefore, more research should be conducted on this dimension (Alonso and Carbó 2021). Furthermore, this exercise refers only to the supply side of the ML models. To find an "optimal model" in equilibrium, we should strike a balance with an indifference curve representing supervisors' preferences.

This empirical exercise shows the potential for this methodology to discern the important factors inside the "algorithmic black box," and connect them in a realistic manner

[25] Unlike supervised algorithms, reinforcement learning algorithms receive an assessment for each given response, and learn based on the reward or punishment they receive for hit or miss, so the time dimension of loans would be an essential variable for reinforcement learning. Instead, we extrapolated in "Model risk-adjusted predictive performance" section the prediction gains for reinforcement learning as a function of the gains from Deep Learning (deep neural network).

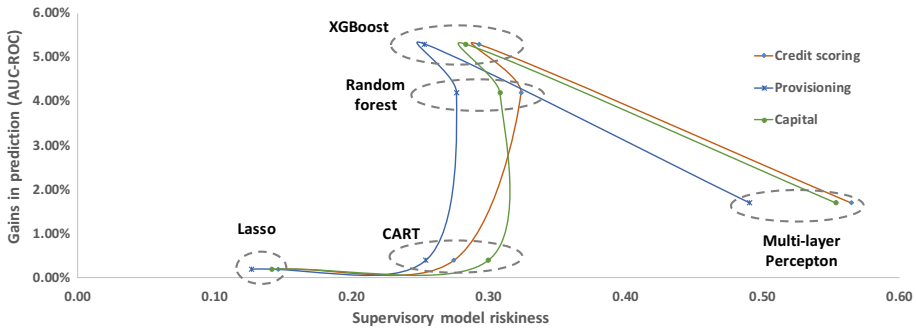

Fig. 5 Risk-reward, by model purpose

with the regulatory requirements to obtain a transparent result that is easier to communicate to the banking industry.

Conclusion

According to recent surveys, credit institutions in the field of credit risk management are at different stages of ML implementation. These range from the calculation of regulatory capital to credit scoring or estimation of provisions. In this environment, financial authorities face the challenge of allowing financial institutions and clients to maximize the opportunities derived from progress and innovation, while observing the principles of technological neutrality, regulatory compliance, and consumer protection. To duly address this challenge, we present a framework to measure the model risk-adjusted performance of ML used in the area of credit default prediction.

To calculate the model risk when evaluating the statistical performance of ML, we first identified 13 factors that could make this technology incompatible with the IRB validation system. We divided these 13 factors into three categories, statistics, technology, and market conduct, and described a procedure to assign a score to each category based on the ML model being used. The importance of these categories when calculating model risk depends on the use of the model itself (credit scoring, regulatory capital, or provisions). We collect a series of regulatory documents for each use case, and, using NLP, we compute the importance of each risk category according to the intensity of mentions. We find that statistical risks are more important for regulatory capital, while technology and market conduct risks are more important for credit scoring. We tested our framework to measure the model risk of five of the most popular ML algorithms, using a publicly available credit default database. When comparing the model risk of each of them with their respective predictive performance in terms of the AUC-ROC, we can assess which of the ML models have a better risk-adjusted performance.

Thus, the evolution of ML in the financial sector must consider the supervisory internal model valuation process. It should also be in line with the explanatory needs of the results, something that the academic literature is promoting with important developments in the field of interpretable ML. Several challenges remain for further research. First, all variables in each risk category were measured methodically. Of particular interest could be to investigate new approaches to capture the presence of risk factors in the regulatory texts relating to statistics, technology and market conduct. For instance,

using latent Dirichlet allocation (Blei et al. 2003) for topic modeling, or keyword discovery (Sarica et al. 2019) for semantic research could support the reproducibility of the results, as well as our methodology based on expert risk terminology. In addition, the benefits of employing ML models using larger datasets should be quantified. Integrating macro-prudential considerations into this assessment could be the cornerstone of any policy decision. For this purpose, further assumptions could be made regarding supervisors' preferences to assess how banks respond when choosing models with certain risks.

Appendix

On the computation of the remaining components of ML model risk

Statistics

Overfitting is a problem that arises when the predictive model has poor generalization performance. The more specialized the model is on the training data, the less it can generalize on new test data. To calculate the overfitting, we can compare the performance of the model (the loss function) on the train sample and on the test sample after each update during training or after including more data. The graphical representation of this comparison is called the "learning curve". We might consider an overfitting problem to exist if the training loss plot decreases with experience, while the validation loss plot does not decrease, or decreases to a point and starts to increase again.

Dynamic calibration refers to the need to re-train the model as new data is fed continually into the system. While a static model can be trained offline, a dynamic model adapts to changing data which requires to be trained online, incorporating new observations through continues updates. Today, thanks to new technologies, many sources of information actually change over time, so the more features the model has, the greater the need to monitor the input data for changes. In this sense, since ML is capable of handling larger amounts of data (EBA 2020), we could represent this risk component by counting the number of features that the model has in production, after data pre-processing.

Feature engineering is necessary in those ML models that require the transformation of input variables or features to work correctly or to improve their performance. On the other hand, by transforming the variables and not dealing with raw data, we could lose control over their economic sense.

A non-exhaustive list of techniques considered as feature engineering could include: (1) Data imputation (numerical and categorical), (2) Handling outliers (cap or drop the observation), (3) Binning, (4) Log transformation, (5) One-hot encoding (transform a categorical variable into a set of binary values), (6) Grouping observations (e.g.: highly correlated variables), (7) Feature split, (8) Scaling (either normalizing or standardizing), or (9) extracting a date. One way to calculate the feature engineering risk factor for different ML models could be to assess how sensitive the performance of ML model is to some of the aforementioned techniques.

Technology category

Third party-providers constitutes a risk exposure to the extent that an institution cannot control the outcome of a service within its own in-house risk framework. This would be

subject to the characteristics of the IT system and human capital of each institution. The more complexity the ML model presents, the higher the probability that an institution requires outsourcing to an external SaaS[26] provider. This could be proxied by a dummy variable if an ML model requires outsourcing or not.

Cyber-risk is a sub-class of third-party provider risk, but due to its potential impact it deserves to be evaluated separately. Our aim is to assess a potential change in the risk exposure if an institution requires too much computational power (number of operations per second) forcing it to shift from in-house deployment to cloud services.

Usually preparing an ML model for production involves four steps: (1) Pre-processing input data, (2) Training the model, (3) Storing the trained model, and (4) Deployment of the model. Clearly, training the ML model is the most computationally intensive task, especially for Deep Learning. In this scenario, we define cyber-risk as cloud migration risk (Akinrolabu et al. 2019), which could be evaluated as the marginal contribution of a single ML model to the total computational requirements of the overall models currently in production in the institution.

Market conduct category

Privacy refers to the legal mandate to protect personal data, as any piece of information that relates to an identifiable person. Both in the US[27] and EU[28] institutions must comply with strict security and privacy requirements, as regulators strive to protect discrimination in credit decisions by automated systems. 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. In this context, the notion of data minimisation (to collect as little data as possible and hold it for as short a time as possible according to the purpose for which it was collected) arises. This runs against ML as an enabler of big data analysis, and requires a qualitative assessment on the probability of each model being able to comply with current legal requirements. This would depend on variables such as the number of features, the number of transformations during data pre-processing and the frequency of updates in data feeding, as well as in-house characteristics of the financial institution regarding data storage architecture.

Auditability is required to comply with model risk governance regulation both in US[29] and EU.[30] Institutions must be able to ensure the robustness, traceability, auditability and resilience of the models. This would include dealing with issues like time or storage limitations for deployment, and production bottlenecks in delivering certain models

[26] Software as a System.

[27] See for instance Smith (2020), detailing the work of the Federal Trade Commission on privacy and data security in artificial intelligence.

[28] The EU's General Data Protection Regulation (GDPR) defines in Article 4 "personal data" as any information relating to an identified or identifiable natural person. An identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

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

[30] See for instance EBA (2019b), on Loan Origination and Monitoring, and EC (2021) with the proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (artificial intelligence act).

to the market, as well as an adequate understanding by the management. This becomes a challenge for ML, which require re-training of models, and complex statistical processes. Therefore, an idea would be to use surrogate models, as a solution that mimics the decision boundary of an original complex model, but through interpretable or "white box" models as regressions or simple classification trees. New techniques are now available that attempt to copy the behaviour of ML models, retaining the original accuracy, but including desired characteristics like interpretability, or reduced number of features (see Unceta et al. 2020). Following this rationale, if a sufficiently accurate copy of an ML model could be found, then we could conclude that there exists a low level of auditability risk.

Biases and potential discrimination in credit decisions by automated systems raised early concerns within regulators both in US[31] and EU.[32] Institutions need to ensure that any model's decision does not rely upon any protected characteristic of an individual. In this context, new ML interpretability techniques, like counterfactuals, are showing promising results, as mentioned in Wachter et al. (2017). The underlying idea would be to use adversarial perturbations by generating synthetic data points close to an existing one (e.g.: race, either white or black) such that the new instance is classified differently than the original one. For example, a counterfactual analysis could suggest for a particular classifier to change the race only to "black" people in order to alter the outcome of the model, while not suggesting that "white" people's race should be varied. If the result of the counterfactual analysis shows that there is some discriminant variable that affects the result of an ML model, then we consider that there is a risk of bias for that model.

SHAP results from empirical exercise

As we discussed in the main text, SHAP is an interpretability technique of ML models that allows us to classify characteristics according to their contribution to the prediction of the ML model. It is model agnostic, so it could be applied to the result of any ML modelling technique. Executing this method requires a considerable amount of time. For this reason, we view the time it takes to run SHAP as a signal of market conduct risk. Therefore, in our framework, what we are mostly interested in is the latency measured in seconds of SHAP execution for different ML models. In this Section, we show for illustrative purposes the results of SHAP when applied to Random Forest, XGBoost, and Deep Learning in our empirical exercise. The results are in Figs. 6, 7 and 8. Features are ranked from most important to least important. The colour red is associated with high values of the feature, and the colour blue is associated with low values of the feature. On the x-axis we can find the impact of the features on the output of the ML model. To understand these numbers, we can take a look at Fig. 6. The

[31] 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.

[32] 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).

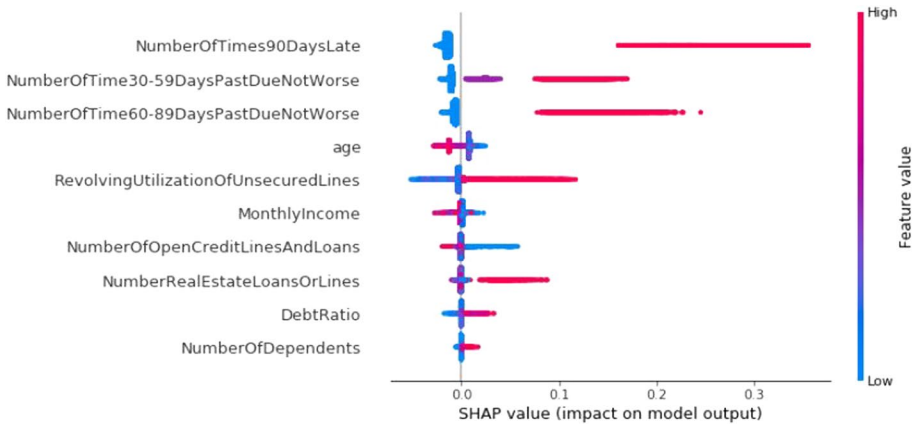

Fig. 6 SHAP interpretation of Random Forest in the empirical example

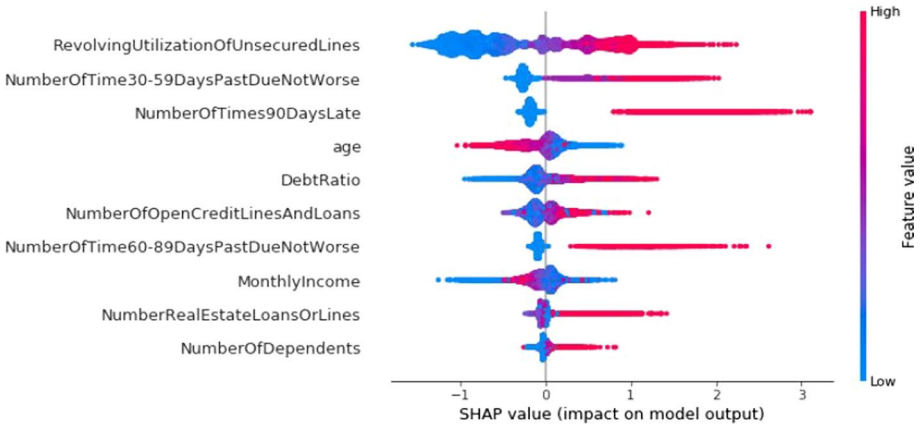

Fig. 7 SHAP interpretation of XGBoost in the empirical example

fact that "NumberOfTimpes90DaysLate" appears first means that, on average, it is the characteristic with the greatest impact on the Random Forest predictions. The higher the values of this characteristic, the greater the impact on the model. This is the opposite of what happens to "Age". The higher the "Age", the lower the impact of the model. The ranking of features slightly changes from one model to another. The features “NumberOfTimpes90DaysLate” and “NumberOfTime3-0-59DaysPastDueNotWorse” appear for the three models among the top three most important variables. But there are some discrepancies. For instance, “RevolvingUtilizationOfUnsecuredLines” appears as the most important variable for the outcome in XGBoost, but not for the output in random forest and deep learning. On the other hand, “Age” always appears the fourth most important variable, and its impact has always the same direction. We leave for further research the study of these discrepancies.

Robustness exercise: sentiment analysis

In our main exercise, in order to weight each risk category (statistics, technology, market conduct) for each possible use case (capital, credit rating, provisions), we first select a

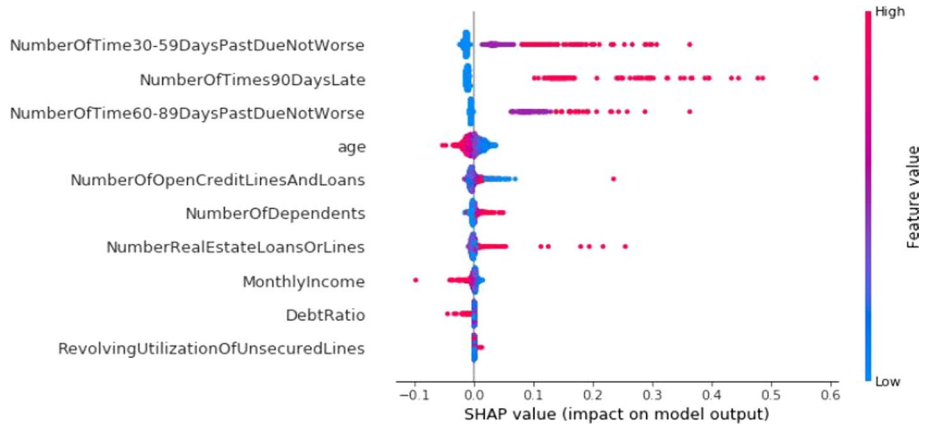

Fig. 8 SHAP interpretation of Deep Learning in the empirical example

series of regulatory documents for each use of the model, and then we count the number of mentions of terms related to statistical, technological and market conduct risks in those documents (see Table 11). We recognize that the number of times a term is mentioned in a document may not be indicative of its importance in the document, or of the sentiment in the document toward the term. For this reason, we complement our main analysis with the exercise that we report below.

Maintaining the same regulatory documents as in our main exercise, we now count the number of negative words surrounding each mention of terms related to statistical, technological, and market conduct risks in those documents. We use the dictionary by Hu and Liu (2004), which is a popular and comprehensive list of negative words in English with up to 4780 terms.[33] Our goal is to weight the number of surrounding negative words by the number of mentions of the terms in every document.

Let i refer to the risk categories (statistics, technology, and market conduct), let j refer to the types of regulatory documents (capital, credit scoring, provisions), and k to any document. We count the number of negative words, if any, that are within d words[34] of distance of each mention of terms from category i in each document of type j. We save that counting in vector $X_{k,i,j}$, a vector with as many positions as mentions of terms from risk category i in document k of type j.

We then calculate $T_{i,j}$ as the average number of negative terms in the d words surrounding the mentions of terms from category i in the type j as follows:

$$T_{i,j} = \frac{\sum_{k=1}^{K_j} \sum_{n=1}^{N_{k,i,j}} X_{k,i,j}^{n}}{\sum_{k=1}^{K_j} N_{k,i,j}}$$

where $X_{k,i,j}^{n}$ is the nth position of vector $X_{k,i,j}$ (i.e.: the number of negative words within d words of the nth mention of a term from category i in document k of type j), term K_j indicates the total number of documents on a type of regulatory documents j, and term $N_{k,i,j}$ indicates the number of mentions of terms from risk category i in document k of type j.

[33] We acknowledge, as stated by one of the creators of the dictionary, Bing Liu (2010), that the appearance of one or more negative words in a sentence that contains the term of interest does not necessarily imply a negative sentiment. Still, we believe the exercise serves us to approximate the interest.

[34] We have tried different specification for parameter d, as $d = 5$, 10, 15 and 20.

Table 8 Average number of negative terms in the 10 words surrounding the key terms

	Statistics	Technology	Market conduct
Capital	0.79	0.59	0.78
Credit scoring	0.93	1.06	0.95
Provisions	0.89	0.42	0.75

Table 9 Average number of negative terms in the 20 words surrounding the key terms

	Statistics	Technology	Market conduct
Capital	1.53	1.17	1.46
Credit scoring	1.85	2.04	1.76
Provisions	1.76	0.96	1.38

This way, instead of calculating the intensity of the appearance of the term, we measure the tone toward the term. The results are shown in Tables 8 and 9, for $d=10$ and 20 respectively (the results do not change substantially for different values of d like $d=5$ or $d=15$). In those tables we show the average negative words surrounding the risk terms by risk category and by use of the model. One of our main findings is that sentiment regarding technology and market conduct risks is more negative for credit scoring than for regulatory capital and provisions. In documents related to credit scoring, terms related to technology and conduct risk tend to be more surrounded by negative terms (1.06 and 0.96 negative words for every 10 words, and 2.04 and 1.76 for every 20 words) than in documents related to capital or provisions. This is true even though credit scoring documents have the lowest percentage of negative words out of total words, as shown in Table 10. This is in line with our main analysis. The terms related to technological risks and market conduct risks appear more frequently and with more negative sentiment in documents related to credit scoring than in regulatory capital or provisions documents.

Second, there are fewer negative words around key terms in provisions documents than in regulatory capital or credit scoring documents (except for statistical risk). This effect is even more significant if we take into account that provisions documents are those with the most negative terms overall, by a wide margin (Table 10). Again, this supports our main exercise, in which we found that the mentions of risk terms for the three categories were lower for provisions, suggesting that provisioning is the category in which ML could has the lowest perceived risk. On the other hand, sentiment towards statistical risk in regulatory capital documents is no worse than in credit rating or provisions documents. This is at odds with our main exercise, in which we consider statistical risk to be particularly relevant for regulatory capital. We leave the study of this result for future research.

Table 10 Percentage of negative words over total

	Percentage of negative words over total words
Capital	3.87%
Credit scoring	3.09%
Provisions	4.88%

Remaining figures

See Fig. 9 and Tables 11, 12 and 13.

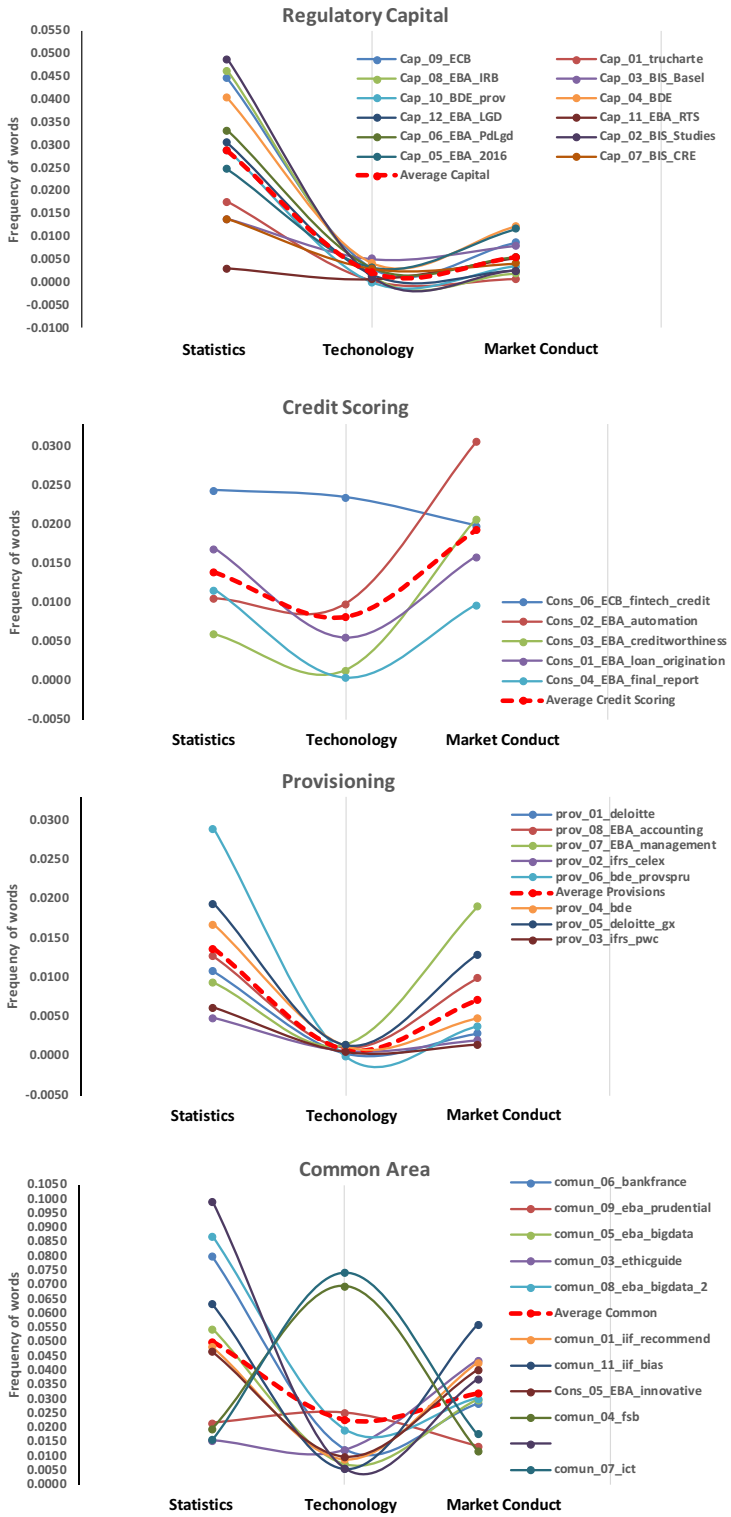

Fig. 9 Calculation of regulatory requirements using NLP, for each model purpose

Table 11 Terminology for each model risk category

Statistics	Technology	Market conduct
Stability	Transparency	Privacy
Over-fitting	Carbon footprint	Auditability
Over-fit	External providers	Interpretability
Hyper parameters	Dependencies	Biases
Dynamic calibration	Cyber-risk	Expert judgement
Feature engineering	Third-party	Expert personnel
Forecast	Cloud	Human judgement
Parameters	IT system	Human-in-the-loop
Test	Legacy	Human-on-the-loop
Calibration	Infrastructure	Governance
Features	Computing	Ethics
Explanatory variables	Computational power	Ethical
AUC	Deployment	Human
ROC	in-house	Compliance
Recall	Development	Management
Prediction	Pilot	Explainability
Logit	ICT	Internal control
Algorithm	Architecture	Knowledge
Scenario	Resilience	Consumers
Data quality	Security	Consumer protection
Back-testing	Operational risk	Discrimination
Benchmarking	Outsourcing risk	Uncertainty
Model	Reversibility	Accountability
Optimisation	DevOps	Market abuse
Dimensionality	Software	Complexity
Validation	Hosting	Decision making
Metrics	Cyber-risk	Soundness
Structured	Model stealing	Conduct
Unstructured	Poisoning attacks	Internal audit
Semi-structured	Adversarial attacks	Fairness
Classification	Open-source	Fair
Tree-based		Diversity
Neural network		Oversight
Regression		Simplicity
Clustering		GDPR
Support vector machine		Transparency
Reinforcement learning		Traceability
Parametric		Opaqueness
Non-parametric		Black-box
Performance		Black boxes
Prociclality		Surrogates
Train		Trust
Training		Trustworthiness
Volatility		Influence
Tuning		SHAP
Threshold		Shapley
Cross-validation		Independent conditional expectations
Compilation		Partial dependence plots
Out-of-sample statistical		ICE
Predictive		PDP
Challenger model		Complex
Correlation		
Confidence level		

Table 12 Set of regulatory texts analysed, classified by purpose of the model

Capital
"Credit Portfolios and Risk-Weighted Assets: Analysis of European Banks". Trucharte et al. (2015). Estabilidad Financiera No 29. Banco de España
"Studies on the Validation of Internal Rating Systems". BIS Working Paper No 14. Heitfield (2005)
" The Internal Ratings-Based Approach. Supporting document of the New Basel Accord". BIS (2001)
"Implementation and Validation of Basel II Advanced Approaches in Spain". Banco de España (2006)
"On the specification of the assessment methodology for competent authorities regarding compliance of an institution with the requirements to use the IRB Approach in accordance with Articles 144(2), 173(3) and 180(3)(b) of Regulation (EU) No 575/2013". EBA (2016a)
"Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures". EBA (2017a)
"Calculation of RWA for credit risk". BIS (2020)
"EBA Report on IRB modelling practices. Impact assessment for the GLs on PD, LGD and the treatment of defaulted exposures based on the IRB survey results". EBA (2017b)
"ECB Guide to internal models". ECB (2019)
"Provisioning models vs Prudential models". García Céspedes (2019)
"Discussion paper on draft regulatory technical standards on prudent valuation, under Article 100 of the draft Capital Requirements Regulation (CRR)". EBA (2013)
"Guidelines for the estimation of LGD appropriate for an economic downturn ('Downturn LGD estimation')". EBA (2019c)
Credit scoring
"Draft Guidelines on loan origination and monitoring". EBA (2019b)
"Report on automation in financial advice". European Supervisory Authorities (2016b)
"Guidelines on creditworthiness assessment". EBA (2015)
"On the Decision of the European Banking Authority specifying the benchmark rate under Annex II to Directive 2014/17/EU (Mortgage Credit Directive)". EBA (2016b)
"Report on innovative uses of consumer data by financial institutions". EBA (2017c)
"Guide to assessments of fintech credit institution licence applications". ECB (2018)
Provisioning
"Applying the expected credit loss model to trade receivables using a provision matrix". Deloitte (2018)
"COMMISSION REGULATION (EU) 2016/2067 of 22 November 2016 amending Regulation (EC) No 1126/2008 adopting certain international accounting standards in accordance with Regulation (EC) No 1606/2002 of the European Parliament and of the Council as regards International Financial Reporting Standard 9" EC (2016)
"In depth IRFS 9 impairment: how to include multiple forward-looking scenarios". PwC (2017)
"Financial Stability Consequences of the Expected Credit Loss Model in IFRS 9". Sánchez Serrano (2018)
"The implementation of IFRS 9 impairment requirements by banks" Deloitte (2016)
"Provisioning models vs Prudential models". García Céspedes (2019)
"Guidelines on management of non-performing and forborne exposures". EBA (2018a)
"Guidelines on credit institutions' credit risk management practices and accounting for expected credit losses". EBA (2017d)
Common area
"IIF Machine Learning Recommendations for Policymakers". IIF (2019a)
"Explainability in Predictive Modelling". IIF (2018)
"Bias and Ethical Implications in Machine Learning". IIF (2019b)
"Ethics Guidelines for Trustworthy AI". High-Level Expert Group on Artificial Intelligence set up by the European Commission (2019)
"Third-party dependencies in cloud services. Considerations on financial stability implications". Financial Stability Board (2019)
"Joint Committee Discussion Paper on the Use of Big Data by Financial Institutions". European Supervisory Authorities (2016a)
"Governance of Artificial Intelligence in Finance". Dupont et al. (2020)
"Guidelines on ICT Risk Assessment under the Supervisory Review and Evaluation process (SREP)". EBA (2017e)
"EBA Report on Big Data and Data Analytics". EBA (2020)
"EBA Report on the Prudential Risks and Opportunities Arising for Institutions from Fintech". EBA (2018b)

Table 13 Summary of papers on ML and credit default, by methodology used

Model	Papers
Lasso	Brown and Mues (2012) Jones et al. (2015) Cheng and Xiang (2017) Guegan and Hassani (2018) Moscato et al. (2021)
Tree	Khandani et al. (2010) Brown and Mues (2012) Jones et al. (2015) Guegan and Hassani (2018)
Random Forest	Brown and Mues (2012) Jones et al. (2015) Moscatelli et al. (2020) Moscato et al. (2021)
Boosting	Brown and Mues (2012) Jones et al. (2015) Petropoulos et al. (2019) Sigrist and Hirnschall (2019) Butaru et al. (2016)
Neural network	Brown and Mues (2012) Babaev et al. (2019) Petropoulos et al. (2019) Kvamme et al. (2018) Sirignano and Cont (2019) Turiel and Aste (2019) Albanesi and Vamossy (2019) Moscato et al. (2021)

Abbreviations

ML	Machine learning
IRB	Internal-ratings based
SHAP	SHapley Additive exPlanations
NLP	Natural Language Processing
PD	Probability of default
CART	Classification and regression tree
AUC-ROC	Area under the curve—receiver operating characteristic
TPR	True positive rate
FPR	False positive rate
LGD	Loss given default
GPU	Graphics Processing Unit
TPU	Tensor Processing Unit
SSM	Single supervisory mechanism
IASB	International Accounting Standard Board

Acknowledgements
The authors appreciate the comments received from Ana Fernández, Sergio Gorjón, José Manuel Marqués, Carlos Conesa, Juan Ayuso, Carolina Toloba and Arancha Gutiérrez, as well as the information shared by Supervision Department IV of Banco de España and the suggestions and feedback received from colleagues attending two internal webinars and from participants of the International Risk Management Conference 2021.

Author contributions
AA carried out the analysis of the current regulation in regulatory capital, credit scoring and provisions in order to identify the risks associated with machine learning in those fields. AA also made the compilation of regulatory texts. JMC carried out the empirical analyses, both those related to machine learning methods, as well as those corresponding to interpretability of ML and natural language processing. Both (AA and JMC) conceived the design of the study, the interpretation of the results, wrote the manuscript, and provided revisions to the manuscript. Both authors read and approved the final manuscript.

Declarations

Competing interests
We declare that we have no competing financial, professional or personal interests that might have influenced the performance or presentation of the work described in the manuscript. The opinions and analyses in this paper are the

sole responsibility of the authors and, therefore, do not necessarily coincide with those of the Banco de España or the Eurosystem.

Received: 16 June 2021 Accepted: 12 May 2022
Published online: 12 July 2022

References

Akinrolabu O, Nurse JR, Martin A, New S (2019) Cyber risk assessment in cloud provider environments: current models and future needs. Comput Secur 87:101600
Albanesi S, Vamossy DF (2019) Predicting consumer default: a deep learning approach (No. w26165). National Bureau of Economic Research
Aldasoro I, Gambacorta L, Giudici P, Leach T (2020) Operational and cyber risks in the financial sector
Alonso A, Carbó JM (2021) Understanding the performance of machine learning models to predict credit default: a novel approach for supervisory evaluation
Alonso A, Marqués JM (2019) Financial innovation for a sustainable economy. Banco de España Occasional Paper (1916)
Ariza-Garzón MJ, Arroyo J, Caparrini A, Segovia-Vargas MJ (2020) Explainability of a machine learning granting scoring model in peer-to-peer lending. IEEE Access 8:64873–64890
Babaev D, Savchenko M, Tuzhilin A, Umerenkov D (2019) Et-rnn: applying deep learning to credit loan applications. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining, pp 2183–2190
Babel B, Buehler K, Pivonka A, Richardson B, Waldron D (2019) Derisking machine learning and artificial intelligence. McKinsey Quarterly, Business Technology Office
Banco de España (2006) Implantation and validation of Basel II advanced approaches in Spain. Paper presented at Bankers Association's regulatory compliance conference, Orlando, FL, 12 June
Bank of International Settlements (1996) Supervisory framework for the use of "back testing" in conjunction with the internal models approach to market risk capital requirements
Bank of International Settlements (2001) The New Basel Capital Acord
Bank of International Settlements (2020) Calculation of RWA for credit risk
Barr B, Xu K, Silva C, Bertini E, Reilly R, Bruss CB, Wittenbach JD (2020) Towards ground truth explainability on tabular data
Barruetabeña E (2020) Impact of new technologies on financial inclusion. Banco De Espana Article 5:20
Bartlett R, Morse A, Stanton R, Wallace N (2022) Consumer-lending discrimination in the FinTech era. J Financial Econ 143(1):30–56
Basel Committee on Banking Supervision (1996) Supervisory framework for the use of "back testing" in conjunction with the internal models approach to market risk capital requirements
Basel Committee on Banking Supervision (2001) The internal ratings-based approach. Supporting document of the New Basel Accord
Bazarbash M (2019) Fintech in financial inclusion: machine learning applications in assessing credit risk. International Monetary Fund
Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022
Bracke P, Datta A, Jung C, Sen S (2019) Machine learning explainability in finance: an application to default risk analysis
Breiman L (2001) Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat Sci 16(3):199–231
Brown I, Mues C (2012) An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Syst Appl 39(3):3446–3453
Butaru F, Chen Q, Clark B, Das S, Lo AW, Siddique A (2016) Risk and risk management in the credit card industry. J Bank Finance 72:218–239
CDP (2020) The time to green finance. CDP Financial Services Disclosure Report 2020
Chen H, Xiang Y (2017) The study of credit scoring model based on group lasso. Procedia Comput Sci 122:677–684
Deloitte (2016) The implementation of IFRS 9 impairment requirements by banks
Deloitte (2018) Applying the expected credit loss model to trade receivables using a provision matrix
Dobbie W, Liberman A, Paravisini D, Pathania V (2021) Measuring bias in consumer lending. Rev Econ Stud 88(6):2799–2832
Dupont L, Fliche O, Yang S (2020) Governance of artificial intelligence in finance. Banque De France
European Banking Authority (2013) Discussion paper on draft regulatory technical standards on prudent valuation, under Article 100 of the draft Capital Requirements Regulation (CRR)
European Banking Authority (2015) Guidelines on creditworthiness assessment
European Banking Authority (2016a) RTS on the specification of the assessment methodology for competent authorities regarding compliance of an institution with the requirements to use the IRB Approach in accordance with Articles 144(2), 173(3) and 180(3)(b) of Regulation (EU) No 575/2013
European Banking Authority (2016b) On the Decision of the European Banking Authority specifying the benchmark rate under Annex II to Directive 2014/17/EU (Mortgage Credit Directive)
European Banking Authority (2017a) Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures
European Banking Authority (2017b) Report on IRB modelling practices. Impact assessment for the GLs on PD, LGD and the treatment of defaulted exposures based on the IRB survey results
European Banking Authority (2017c) Report on innovative uses of consumer data by financial institutions

European Banking Authority (2017d) Guidelines on credit institutions' credit risk management practices and accounting for expected credit losses
European Banking Authority (2017e) Guidelines on ICT Risk Assessment under the Supervisory Review and Evaluation process (SREP)
European Banking Authority (2018a) Guidelines on management of non-performing and forborne exposures
European Banking Authority (2018b) Report on the prudential risks and opportunities arising for Institutions from Fintech
European Banking Authority (2019a) Progress Report on the IRB Roadmap
European Banking Authority (2019b) Draft Guidelines on loan origination and monitoring
European Banking Authority (2019c) Guidelines for the estimation of LGD appropriate for an economic downturn ('Downturn LGD estimation')
European Banking Authority (2020) Report on Big Data and Advanced Analytics
European Central Bank (2018) Guide to assessment of fintech credit institution license applications
European Central Bank (2019) ECB guide to internal models
European Comission (2016) COMMISSION REGULATION (EU) 2016/2067 of 22 November 2016 amending Regulation (EC) No 1126/2008 adopting certain international accounting standards in accordance with Regulation (EC) No 1606/2002 of the European Parliament and of the Council as regards International Financial Reporting Standard 9
European Commission (2019) Ethics guidelines for trustworthy AI. High-level expert group on artificial intelligence
European Commission (2021) Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (artificial intelligence act). COM/ 2021/ 206
European Supervisory Authorities (2016a) Joint Committee Discussion Paper on the use of Big Data by Financial Institutions
European Supervisory Authorities (2016b) Report on automation in financial advice
Farkas W, Fringuellotti F, Tunaru R (2020) A cost-benefit analysis of capital requirements adjusted for model risk. J Corp Finan 65:101753
Financial Stability Board (2019) Third-party dependencies in cloud services. Considerations on financial stability implications
Fisher A, Rudin C, Dominici F (2019) All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. J Mach Learn Res 20(177):1–81
Frye C, Rowat C, Feige I (2019) Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. arXiv preprint arXiv:1910.06358
Fuster A, Goldsmith-Pinkham P, Ramadorai T, Walther A (2022) Predictably unequal? The effects of machine learning on credit markets. J Financ 77(1):5–47
García Céspedes JC (2019) Provisioning Models vs. Prudential Models. Revista de estabilidad financiera. Nº 36 (primavera 2019), p 125–146
Gu S, Kelly B, Xiu D (2020) Empirical asset pricing via machine learning. Rev Financ Stud 33(5):2223–2273
Guegan D, Hassani B (2018) Regulatory learning: How to supervise machine learning models? An application to credit scoring. J Finance Data Sci 4(3):157–171
Hall P, Cox B, Dickerson S, Ravi Kannan A, Kulkarni R, Schmidt N (2021) A United States fair lending perspective on machine learning. Front Artif Intell 4:78
Heitfield H (2005) Studies on the validation of internal rating systems. Working paper 14. Basel Committee on Banking Supervision
Heskes T, Sijben E, Bucur IG, Claassen T (2020) Causal shapley values: exploiting causal knowledge to explain individual predictions of complex models. arXiv preprint arXiv:2011.01625
Hu M, Liu B (2004) Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining, pp 168–177)
Huang Y, Zhang L, Li Z, Qiu H, Sun T, Wang X (2020) Fintech credit risk assessment for SMEs: evidence from China
Institute of International Finance (2018) Explainability in predictive modelling
Institute of International Finance (2019a) Machine learning in credit risk
Institute of International Finance (2019b) Machine learning: recommendations for policymakers
Institute of International Finance (2019c) Bias and ethical implications in machine learning
Institute of International Finance (2020) Machine learning governance
Jagtiani J, Lemieux C (2019) The roles of alternative data and machine learning in fintech lending: evidence from the LendingClub consumer platform. Financ Manage 48(4):1009–1029
Jones S, Johnstone D, Wilson R (2015) An empirical evaluation of the performance of binary classifiers in the prediction of credit ratings changes. J Bank Finance 56:72–85
Jung C, Mueller H, Pedemonte S, Plances S, Thew O (2019) Machine learning in UK financial services. Bank of England and Financial Conduct Authority
Kerkhof J, Melenberg B, Schumacher H (2010) Model risk and capital reserves. J Bank Finance 34(1):267–279
Khandani AE, Kim AJ, Lo AW (2010) Consumer credit-risk models via machine-learning algorithms. J Banking Finance 34(11):2767–2787
Kou G, Peng Y, Wang G (2014) Evaluation of clustering algorithms for financial risk analysis using MCDM methods. Inf Sci 275:1–12
Kou G, Xu Y, Peng Y, Shen F, Chen Y, Chang K, Kou S (2021a) Bankruptcy prediction for SMEs using transactional data and two-stage multiobjective feature selection. Decisi Support Syst 140:113429
Kou G, Olgu Akdeniz Ö, Dinçer H, Yüksel S (2021b) Fintech investments in European banks: a hybrid IT2 fuzzy multidimensional decision-making approach
Königstorfer F, Thalmann S (2020) Applications of Artificial Intelligence in commercial banks–A research agenda for behavioral finance. J Behav Experiment Finance 27:100352
Kvamme H, Sellereite N, Aas K, Sjursen S (2018) Predicting mortgage default using convolutional neural networks. Expert Syst Appl 102:207–217

Li T, Kou G, Peng Y, Philip SY (2021) An integrated cluster detection, optimization, and interpretation approach for financial data. IEEE Transactions on Cybernetics
Liu B (2010) Sentiment analysis and subjectivity. Handb Nat Lang Process 2(2010):627–666
Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: Advances in neural information processing systems, pp 4765–4774
Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Lee SI (2020) From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2(1):56–67
Lynn T, Mooney JG, Rosati P, Cummins M (2019) Disrupting finance: FinTech and strategy in the 21st century. Springer Nature, p 175
Masciandaro D, Peia O, Romelli D (2020) Banking supervision and external auditors: theory and empirics. J Financ Stab 46:100722
Mittelstadt B, Russell C, Wachter S (2019) Explaining explanations in AI. In: Proceedings of the conference on fairness, accountability, and transparency, pp 279–288
Molnar C (2019) Interpretable machine learning: a guide for making black box models explainable (published online)
Moreno ÁI, Caminero T (2020) Application of text mining to the analysis of climate-related disclosures
Moscatelli M, Parlapiano F, Narizzano S, Viggiano G (2020) Corporate default forecasting with machine learning. Expert Syst Appl 161:113567
Moscato V, Picariello A, Sperlí G (2021) A benchmark of machine learning approaches for credit score prediction. Expert Syst Appl 165:113986
Petropoulos A, Siakoulis V, Stavroulakis E, Klamargias A (2019) A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting. IFC Bulletins chapters, 49
PwC (2017) In depth IRFS 9 impairment: how to include multiple forward-looking scenarios
Ribeiro MT, Singh S, Guestrin C (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, pp 1135–1144
Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215
Sánchez Serrano A (2018) Financial stability consequences of the expected credit loss model in IFRS 9. Financial Stability Review
Sarica S, Song B, Low E, Luo J (2019) Engineering knowledge graph for keyword discovery in patent search. In: Proceedings of the design society: international conference on engineering design, vol 1, no 1, Cambridge University Press, pp 2249–2258
Sarlin P (2013) On policymakers' loss functions and the evaluation of early warning systems. Econ Lett 119(1):1–7
Sigrist F, Hirnschall C (2019) Grabit: Gradient tree-boosted Tobit models for default prediction. J Bank Finance 102:177–192
Sirignano J, Cont R (2019) Universal features of price formation in financial markets: perspectives from deep learning. Quant Finance 19(9):1449–1459
Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H (2020) Fooling lime and shap: adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM conference on AI, Ethics, and Society, pp 180–186
Smith A (2020) Using artificial intelligence and algorithms. US Federal Trade Commission, FTC Business Blog, April. https://www.ftc.gov/news-events/blogs/business-blog/2020/04/using-artificial-intelligence-algorithms
Strubell E, Ganesh A, McCallum A (2019) Energy and policy considerations for deep learning in NLP. arXiv preprint arXiv:1906.02243
Tarashev N (2010) Measuring portfolio credit risk correctly: why parameter uncertainty matters. J Bank Finance 34(9):2065–2076
Trucharte Artigas C, Pérez Montes C, Cristófoli ME, Ferrer Pérez A, Lavín San Segundo N (2015) Credit portfolios and risk weighted assets: analysis of European banks. Estabilidad financiera. No 29 (noviembre 2015), pp 63–85
Turiel JD, Aste T (2019) P2P Loan acceptance and default prediction with artificial intelligence. arXiv preprint arXiv:1907.01800
Unceta I, Nin J, Pujol O (2020) Copying machine learning classifiers. IEEE Access 8:160268–160284
Vapnik VN, Chervonenkis AY (2015) On the uniform convergence of relative frequencies of events to their probabilities. In: Measures of complexity, Springer, Cham, pp 11–30
Wachter S, Mittelstadt B, Russell C (2017) Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harv JL Tech 31:841
Wall LD (2018) Some financial regulatory implications of artificial intelligence. J Econ Bus 100:55–63

Publisher's Note