South African Journal of Economic and Management Sciences
ISSN: (Online) 2222-3436, (Print) 1015-8812

Integrating traditional and non-traditional model risk frameworks in credit scoring

Authors:
Hendrik A. du Toit[1]
Willem D. Schutte[1,2]
Helgard Raubenheimer[1,2]

Affiliations:
[1]Centre for Business Mathematics and Informatics, Faculty of Natural and Agricultural Sciences, North-West University, Potchefstroom, South Africa

[2]National Institute for Theoretical and Computational Sciences (NITheCS), Pretoria, South Africa

Corresponding author:
Hendrik du Toit,
drikus0329dutoit@gmail.com

Dates:
Received: 10 June 2024
Accepted: 20 Aug. 2024
Published: 08 Oct. 2024

How to cite this article:
Du Toit, H.A., Schutte, W.D. & Raubenheimer, H., 2024, 'Integrating traditional and non-traditional model risk frameworks in credit scoring', *South African Journal of Economic and Management Sciences* 27(1), a5786. https://doi.org/10.4102/sajems.v27i1.5786

Background: An improved understanding of the reasoning behind model decisions can enhance the use of machine learning (ML) models in credit scoring. Although ML models are widely regarded as highly accurate, the use of these models in settings that require explanation of model decisions has been limited because of the lack of transparency. Especially in the banking sector, model risk frameworks frequently require a significant level of model interpretability.

Aim: The aim of the article is to evaluate traditional model risk frameworks to determine their appropriateness when validating ML models in credit scoring and enhance the use of ML models in regulated environments by introducing a ML interpretability technique in model validation frameworks.

Setting: The research considers model risk frameworks and regulatory guidelines from various international institutions.

Method: The research is qualitative in nature and shows how through integrating traditional and non-traditional model risk frameworks, the practitioner can leverage trusted techniques and extend traditional frameworks to address key principles such as transparency.

Results: The article proposes a model risk framework that utilises Shapley values to improve the explainability of ML models in credit scoring. Practical validation tests are proposed to enable transparency of model input variables in the validation process of ML models.

Conclusion: Our results show that one can formulate a comprehensive validation process by integrating traditional and non-traditional frameworks.

Contribution: This study contributes to existing model risk literature by proposing a new model validation framework that utilises Shapley values to explain ML model predictions in credit scoring.

Keywords: machine learning models; credit scoring; model risk frameworks; model interpretability; model validation; Shapley values; model transparency.

Introduction

The field of machine learning (ML) has gained a lot of popularity in recent years. Therefore, the importance of interpretable ML models in regulated environments such as the banking sector has increased significantly over the last decade. Machine learning[1] is a name for a group of models that are typically classified under the umbrella of artificial intelligence (AI). Although some of the techniques are not new, they have been supported by the advancements in computational power (Bertsimas, King & Mazumder 2016). For this study, we do not aim to define ML. Machine learning models in this study refer to supervised classification algorithms such as tree-based models. Many banks and other financial institutions have seen the benefit of these so-called ML models and are implementing the necessary infrastructure to productionalise these models. Banks have benefited from traditional model methodologies such as logistic regression for over 30 years and developed a body of knowledge and rules that are used to judge the appropriateness of these models when used in practical applications.

The South African Reserve Bank (SARB) adopted the definition of model risk as defined in the Basel II market risk framework (SARB 2015). In this framework, two forms of model risk are identified. The first form of model risk has to do with an incorrect valuation methodology, and the second is unobservable (and possibly incorrect) calibration parameters in the valuation model.

[1] Note that some literature refers to AI, other refer to ML, but for this article, these terms are used interchangeably.

The model risk as identified by SARB (2015), is managed by banks using a function called model risk management. De Jongh et al. (2017) state that this function usually comprises robust and sensible model development, sound implementation procedures, appropriate use of models and consistent model validation. These measures exist to ensure that model risk can be measured and mitigated appropriately. Traditional model risk frameworks refer to model risk frameworks that typically already exist in model development teams and are being used to evaluate traditional models such as logistic regression. Non-traditional model risk frameworks are model risk frameworks that have been proposed in recent literature to evaluate non-traditional models such as ML algorithms.

Model validation is the set of processes and activities intended to verify that models perform as expected, in line with their design objectives and business uses (OCC 2011). The Office of the Comptroller of the Currency (OCC) (2011) report further describes that effective model validation techniques should be able to test model soundness, identify potential limitations and test certain model assumptions. The model validation process includes testing the model's accuracy, testing if the model data is stable over time, determining if the model can differentiate between events and non-events in the data and evaluating if the variables used in the model are intuitive. When referring to the validation of ML models, it relates to best practice in the application of ML models in real-world production environments.

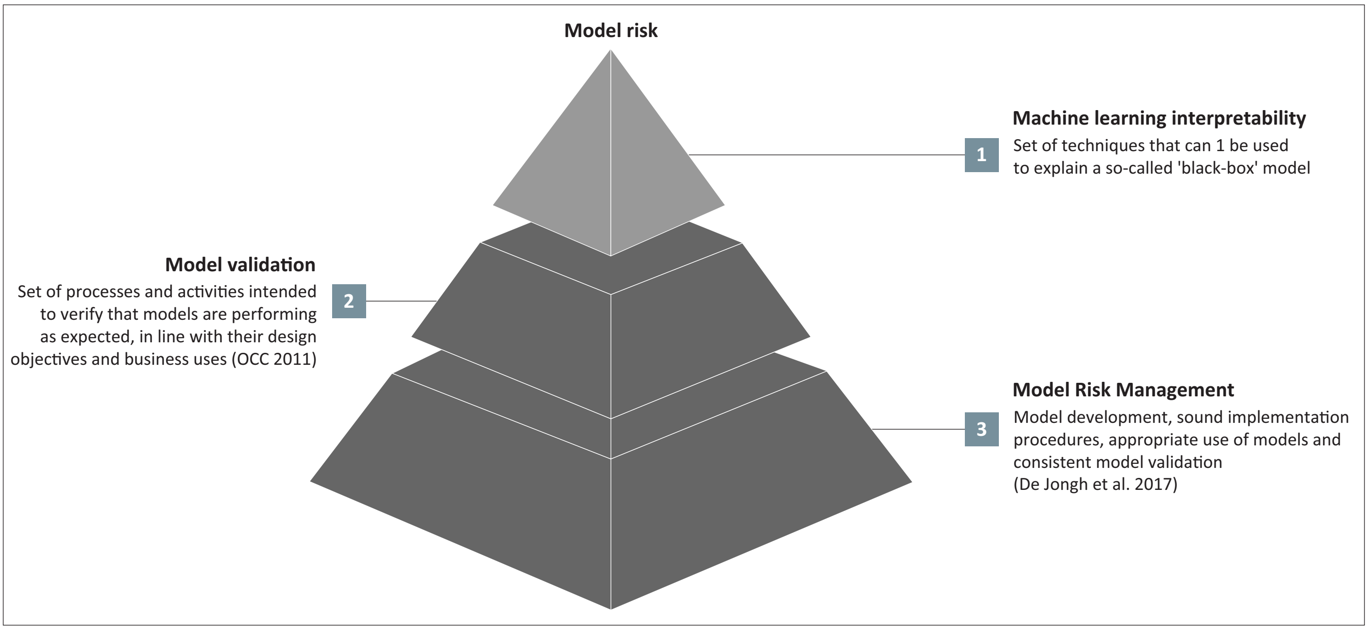
Van der Burgt (2019) notes that the financial sector is particularly important from an ML development perspective and needs an adequate regulatory and supervisory response. The reasons given are:

- The financial sector is commonly held to a higher social standard than many other industries and AI-related incidents can have serious reputation effects.
- Incidents could seriously impact financial stability, given that the financial system is interconnected in many ways (i.e., systemic risk).
- The progress of AI and the increase in the importance of these models in the financial sector directs us to rethink traditional supervisory frameworks.

Figure 1 shows the layers pertaining to model risk as discussed so far. It highlights the role of validation as an important role in the model risk framework and proposes ML interpretability techniques as an additional layer to model risk management.

The use of ML in credit scoring has been proven to be very efficient and financial institutions are exploring different ways of leveraging the accuracy of ML models (Lessmann et al. 2015). With this new domain of models entering specialist areas such as credit scoring, certain questions come to mind about how these models can be validated and proven sound. These questions, among others, have led to a new field of research called ML interpretability techniques. Machine learning interpretability can be described as the set of techniques that can be used to explain a so-called 'black-box' model. There are many different techniques, and each of them aims to describe a piece of the model, or in some cases, it attempts to describe the model entirely.

The European Banking Authority (2021) notes that the main challenges regarding ML models come from the complexity of the model, which leads to difficulties in interpreting the results, ensuring management functions adequately understand them, and, lastly, justifying their results to supervisors. The *EU Artificial Intelligence Act (EU AI Act)* was recently published (European Commission 2024). This act


Model risk
Machine learning interpretability
1
Set of techniques that can 1 be used to explain a so-called 'black-box' model
Model validation
2
Set of processes and activities intended to verify that models are performing as expected, in line with their design objectives and business uses (OCC 2011)
Model Risk Management
3
Model development, sound implementation procedures, appropriate use of models and consistent model validation (De Jongh et al. 2017)

Note: Please see the full reference list of the article, Du Toit, H.A., Schutte, W.D. & Raubenheimer, H., 2024, 'Integrating traditional and non-traditional model risk frameworks in credit scoring', *South African Journal of Economic and Management Sciences* 27(1), a5786. https://doi.org/10.4102/sajems.v27i1.578, for more information.

FIGURE 1: The layers of model risk.

includes a classification of AI applications in terms of risk. According to the act, there are four categories of AI applications:

- prohibited applications
- high-risk applications
- applications with special requirements
- low risk applications.

Credit scoring is classified as a high risk application and must fulfil comprehensive requirements in areas such as transparency. The Institute of International Finance (IIF) and Ernst & Young (EY) (2022) survey report on ML uses in credit risk and anti-money laundering applications notes that some of the key challenges in the adoption of ML included data quality, explainability and IT-infrastructure. The survey further found that existing model risk management frameworks often govern ML applications. Because of the challenge of approval, frameworks need to be created to enable a smoother road towards the implementation of ML models. As such, traditional frameworks are first inspected in the 'Overview of traditional model validation process' section. The section titled Machine Learning Model Risk Management consults recent literature on how model risk management is conducted for ML models. Because of a lack of practical examples, this section further describes the principles for using ML models in the financial sector as proposed by various authors. The section concludes by gathering recent literature on how existing model risk frameworks can be expanded to incorporate ML models. Shapley values are introduced and explained as an ML interpretability technique in the 'Methods' section. This technique is proposed as a model validation technique for ML models. The section titled 'Case study' explains the integration of Shapley values into model validation frameworks to validate ML models. Practical tests are proposed within the 'Results' section to illustrate how Shapley values can be used to validate ML models. The study concludes in the last section and proposes areas for further research.

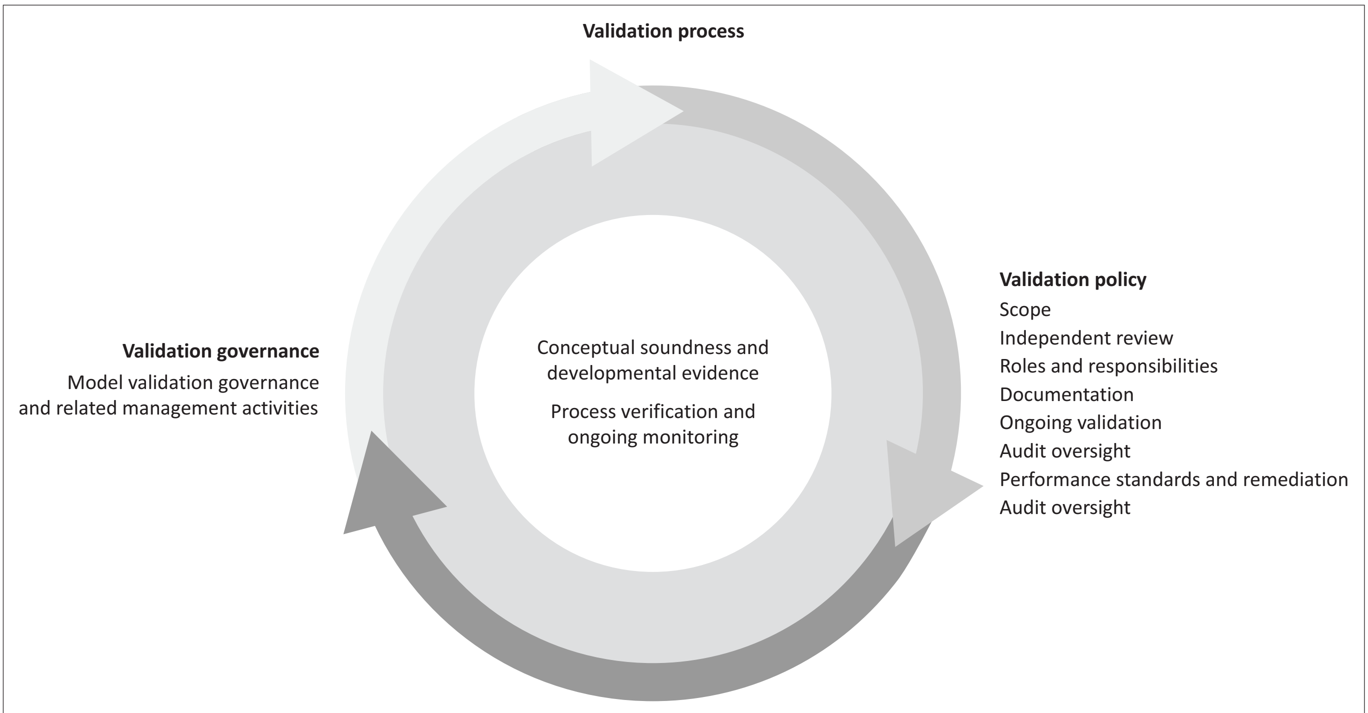
Overview of traditional model validation process

The model validation process is considered a critical step in model risk management. Therefore, regulatory authorities such as the Basel Accord have attempted to set the standard for the model validation process. However, according to our knowledge, research has not been able to establish a definite set of global standards for this process, and not a lot of focus has been placed on providing examples of how these standards could be achieved. In this section, we will focus on a few key resources that encapsulate the core aspects of a traditional model validation framework.

Quell et al. (2021) list the following typical aspects of a model that should be validated and how this should be achieved:

- Model data:
  - data representativeness
  - data traceability and data quality
  - feature engineering
  - other exploratory data analysis techniques.
- Conceptual soundness:
  - model design and algorithm selection
  - model assumptions and limitations
  - dynamic learning
  - explainability and interpretability of the model
  - overfitting and bias.

- Model implementation and ongoing validation.



Source: Adapted from De Jongh, P.J., Larney, J., Mare, E., Van Vuuren, G.W. & Verster, T., 2017, 'A proposed best practice model validation framework for banks', *South African Journal of Economic and Management Sciences* 20(1), a1490. https://doi.org/10.4102/sajems.v20i1.1490

FIGURE 2: Traditional model validation process as proposed by De Jongh et al. (2017).

TABLE 1: Model validation areas and components.

Validation area	Validation components
Inputs	• Input assumptions • Input data • Lending policies and practices
Process	• Model development • Model selection • Model implementation
Output	• Model results interpretation • Holdout sample testing • Performance monitoring and testing

Source: Abrahams, C.R. & Zhang, M., 2008, *Fair lending compliance: Intelligence and implications for credit risk management*, Wiley, Hoboken, NJ

- Model documentation and use.

De Jongh et al. (2017) propose a similar model validation framework (see Figure 2). This framework consists of validation governance, validation process and validation policy. Specifically, the model validation process consists of:

- conceptual soundness and developmental evidence
- process verification and ongoing monitoring
- outcomes analysis.

Both the model validation frameworks of Quell et al. (2021) and De Jongh et al. (2017) highlight conceptual soundness as a critical aspect of the model validation framework. Within conceptual soundness, explainability and interpretability of the model prediction are considered important validation checks. Other authors, such as Abrahams and Zhang (2008), describe model validation by identifying areas and components of importance (see Table 1).

Baesens, Roesch and Scheule (2016) explain that one can quantitatively validate a model by comparing realised numbers to predicted numbers. These numbers will rarely be identical, and therefore, appropriate performance metrics and test statistics should be specified to conduct the comparison. The authors describe the process of back testing models to check data stability. This process determines if the population on which the model has been developed is representative of the population being observed. Back testing is also used to determine if the ranking of the model predictions is comparable to the event that the model is predicting, in other words, comparing predicted events to actual events. The second type of quantitative validation method described by the authors is benchmarking. The method involves comparing the output and performance of the model that is being validated with a reference model, otherwise known as a benchmark. Examples of benchmark models are credit bureaus, rating agencies, existing models and even expert models. Where a relevant benchmark is not available, other regulatory authorities, such as Hong Kong Monetary Authority (HKMA 2006), have suggested the development of an internal benchmark or an expert based benchmark.

Authors such as Abrahams and Zhang (2008) have proposed typical statistical measures and their applications in the model validation process. These include measures such as the Kolmogorov-Smirnov (K-S) test, Receiver Operating Characteristic (ROC) curve, Gini coefficient, cumulative gains chart and the Chi-square statistic. These measures are used to measure model performance and population shift, and to analyse model input.

Siddiqi (2017) lists resources that are used as guidelines for model validation under the umbrella of model risk management. These include resources such as supervisory guidance on model risk management (OCC 2011), GL-44 guidelines on internal governance issues (EBA 2011) and the Basel Committee on Banking Supervision Working Paper 14 (BCBS 2005), to name a few. Although there are many suggested methods to validate models, both Siddiqi (2017) and De Jongh et al. (2017) specifically note that there is no single set of global standards to validate models.

Machine learning model risk management

To define the need for a revised model risk management framework, it is important to understand the dangers of ML models and what risks need to be mitigated. To this end, Quell et al. (2021) listed a few common dangers of ML models. These dangers include explainability, where they note that models such as 'black-box' models are difficult to interpret. This is significant because interpretation is often required for financial models, especially if they must adhere to external regulatory requirements. Many researchers have attempted to provide a comprehensive definition of interpretability. Miller (2017) explains that interpretability can be understood as the degree to which a human can understand the reason for a certain decision. Furthermore, Kim, Khanna and Koyejo (2016) define it as the degree to which a human can consistently predict a model's result.

Although this research will focus on interpretability as a key area of concern, the interested reader can consult Quell et al. (2021) for a complete list of the typical dangers of ML models. Other dangers of the application of ML models include:

- overfitting
- robustness and population drift
- bias, adversarial attacks, and brittleness
- development bias
- *p*-value arbitrage.

Recent literature on model risk management for ML models proposes principles that model risk frameworks need to adhere to, with less focus on proposing practical techniques that can be considered. This article aims to bridge this gap and provide practical techniques that a typical credit scoring team can implement within a credit scoring model validation framework. As such, this section aims to investigate these principles and gain an understanding of how they can be used as validation criteria within a model risk framework. To narrow the scope of this research, a clear focus is placed on a ML governance principle, namely transparency. The below-stated review solidifies what the concept of transparency means and how non-traditional model risk frameworks intend to ensure that the principle of transparency is met.

General principles for using AI in the financial sector have been proposed in recent literature. The principles provide a broad overview of what model risk management teams should keep in mind for the use of ML models in finance. Härle et al. (2015) identified six structural trends that will transform bank risk management in the future. One of these trends is the continuous expansion of the breadth and depth of regulation. Furthermore, the same authors note that compliance with existing rules will likely not be sufficient in the future and that banks will need to comply with broad principles to protect themselves against potential future rules and interpretations of existing rules.

Van der Burgt (2019) introduces general principles for a responsible application of AI in the financial sector. The author notes the following six key principles: soundness, accountability, fairness, ethics, skills and transparency. These are known as the 'SAFEST' principles. The principle of transparency is particularly interesting as it states that firms should be able to explain how and why they use AI in their business processes and how these applications function.

The Monetary Authority of Singapore (MAS 2018) outlines similar principles to aid the use of AI and data analytics in Singapore's financial sector. The MAS (2018) aims to provide the financial sector with a set of foundational principles to consider when using AI and data analytics in decision making. It also aims to assist companies in contextualising and operationalising governance of the use of AI and data analytics. Lastly, it aims to promote public confidence and trust in the use of AI and data analytics. The article introduces the principles of Fairness, Ethics, Accountability and Transparency (FEAT). The MAS (2018) clearly states that these principles are not intended to replace existing relevant internal governance frameworks and that companies should continue to comply with all applicable laws and requirements. Although these principles are not intended to be prescriptive, the authors believe that through industry engagement, there might be areas where more specific or technical guidance would benefit the industry, with the FEAT principles serving as a foundational framework. In summarising the transparency principle, the article explicitly recommends that data subjects are provided, upon request, with clear explanations on what data is used to make an AI and data analytics decision about the data subject and how the data affects the decision.

PricewaterhouseCoopers (PWC) released a report describing the practical implications of the FEAT principles on industries such as banking and insurance (PWC 2018). It advises banks to prepare processes and tools to deal with client requests for explanations. More specifically, the report requests banks to provide information on the input factors and the potential output scenarios of an ML model using non-technical language without revealing the underlying intellectual property. The report further encourages banks to evaluate explanatory techniques, especially for deep learning models.

Shifting the focus slightly to models relevant to the regulatory capital space. The European Banking Authority (EBA) (2021) released a discussion article that aims to understand the challenges and opportunities of applying ML models in the context of internal ratings-based (IRB) models to calculate regulatory capital for credit risk. The article proposes a set of principle-based recommendations based on trust elements:

- ethics
- explainability and interpretability
- traceability and auditability
- fairness and bias prevention
- data protection and quality
- consumer protection aspects and security.

The EBA (2021) lists interpretability as one of the concerns of using ML models. Together with the concerns, the article also lists some of the techniques frequently used to obtain insight into the internal logic of an ML model. More specifically, for interpretability techniques, the following are listed by the EBA (2021):

- Graphical tools such as partial dependence plots (PDP), and individual conditional expectation plots (ICE). These plots are designed to show the effect of an explanatory variable on the model.
- Feature importance measures the relevance of variables in the overall model.
- Shapley values quantify the impact of a variable on the final prediction.
- Local explanations such as Local Interpretable Model-Agnostic Explanations (LIME) and anchors give a simplified explanation of the model from a local point of view.
- Counterfactual explanations show how changing the input variables can influence a model's prediction.

Based on the overview of the general principles for using AI in the financial sector, the next section will focus on seven key pillars that outline the changes required to integrate AI and ML models into existing model risk management frameworks. These changes are a first step towards identifying the practical modifications needed to integrate AI/ML models in existing model risk management frameworks.

Integrating machine learning models in existing model risk management frameworks

KPMG (2022) suggests that model risk management for AI/ML models can be integrated into existing (traditional) model risk management frameworks. In doing this, industries can benefit from synergies that arise from using proven processes and methods. Integrating the new types of models into existing frameworks addresses many regulatory requirements as listed in the *EU AI Act*, but minor changes need to be made to address AI/ML models specifically. Considering this, the white article suggests seven key pillars that outline the changes needed to integrate AI/ML models into existing frameworks. These seven pillars are listed in Table 2.

TABLE 2: Key model risk management pillars as proposed by KPMG (2022).

Key pillar	Description
1. Establish a definition of AI/ML models.	• Banks establish an enterprise-wide definition of what AI/ML models comprise. Expand model inventory to include these models.
2. Updating the model tiering definition.	• Model tiering parameters, specifically around materiality, criticality and uncertainty, need to be revised to correctly incorporate the risk of AI and ML models.
3. Establish an appropriate risk appetite.	• Banks need to leverage peer networks to establish a first draft of a risk appetite statement and appropriate thresholds. This is necessary because traditional risk appetite statements are not designed for ML models and regulatory guidance on this matter is still being developed.
4. Identify accountability.	• Establish clear definitions of roles and accountabilities within all the risk management functions.
5. Invest in skill enhancements.	• Develop a skill set inhouse or involve external parties. External subject matter experts (SMEs) can help benchmark banks with the latest risk management, controls and techniques for model validation using AI/ML models.
6. Enhance the compensatory control framework.	• Designing additional compensatory controls around areas such as benchmarking, feature selection, bias elimination, among others, to account for the lack of transparency. • Enhancing existing data management framework. • Building control frameworks around compliance and operational risk. • Conducting enterprise-wide training programmes.
7. Develop additional tests and procedures for AI/ML models.	• Interpretability • Bias elimination • Dynamic calibration of models • Implementation • Ongoing monitoring.

AI, artificial intelligence; ML, machine learning; SME, subject matter expert.

The general principles of using ML in finance establish a clear directory of what regulatory bodies consider important and relevant when using ML models. Finally, this section introduces seven pillars that need to be adopted to incorporate ML models into existing model risk management frameworks. The following section builds on these pillars by proposing an additional test that can be linked to pillar 7 (see Table 2). This test specifically focusses on interpretability as a requirement in the model validation process. Furthermore, because Shapley values will form an integral part of the tests proposed, the next section will uncover some details about Shapley values.

Methods

A brief overview of the Shapley value

This section introduces the Shapley value as an ML interpretability technique. It provides an overview of the technique and explains the interpretation of the Shapley value with a simple example. The section concludes with a case study that illustrates how a model development team can use the technique to validate the interpretability of ML models.

Shapley value

Considering the different ML interpretability techniques listed in the section titled: Machine learning model risk management, the best method of selecting ML interpretability techniques is considering the characteristics of each available technique. Arrieta et al. (2020) provide an extensive overview of concepts, taxonomies, opportunities and challenges with respect to the proper use of AI. In light of this, the article proposes the use of model-agnostic post-hoc explainability techniques. Model-agnostic techniques can be applied to any ML model, while post-hoc explainability techniques are used to explain the inner workings of an already developed model which is not intrinsically interpretable. Model-agnostic post-hoc interpretability techniques contain feature relevance explanation techniques. These techniques measure the influence, relevance or importance of variables. Shapley additive explanations form part of this group of interpretability techniques and have been proposed by authors such as Lundberg and Lee (2017) and Chen, Lundberg and Lee (2021).

Considering the more recent integration of ML models into traditional model validation frameworks, Shapley values have been mentioned as an interpretability technique to be considered. See, for example, EBA (2021) and Scheda and Diciotti (2022). Although the technique is not without its weaknesses (as is evidenced by the shortcomings listed in Molnar et al. 2020 and Woznica et al. 2021), many researchers, such as Du Toit et al. (2023) consider this technique to be especially applicable for interpreting ML models in a model validation process.

Shapley values originate from game theory (Shapley 1953). Shapley values show the impact of a specific predictor variable on the model outcome. The interested reader can consult sources such as Du Toit et al. (2023) and Kumar et al. (2020) for a detailed explanation of how the Shapley value is calculated. SHapley Additive exPlanation (SHAP) is also proposed by Lundberg and Lee (2017). Considering how expensive the calculation of the Shapley value is, two estimation approaches to calculate SHAP values were introduced by Lundberg and Lee (2017). The first is KernelSHAP, a kernel-based estimation method inspired by local surrogate models. The second is TreeSHAP, an estimation method for tree-based models. The interested reader can consult Molnar (2020) and Lundberg, Erion and Lee (2018) for more details on the estimation techniques.

Shapley value explanation

The following simplified example, as used by Du Toit et al. (2023) illustrates how the Shapley value is applied. In Figure 3, the Shapley value ranges from –0.2 to 0.25 and for this explanation, the value is based on a model that predicts the probability of default. The Shapley value is the marginal increase or decrease in the probability of default contributed by a certain variable entering the model for a specific application. This contribution is added or subtracted from the average model prediction.

To illustrate, assume the average probability of default for the model is 25%. For the variable depicted in Figure 3, if the specific applicant had a variable value in bin 1, then the

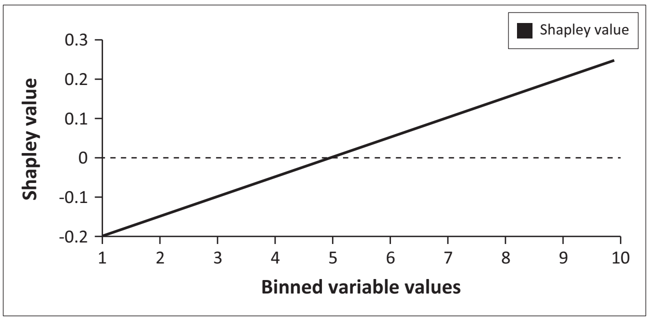


FIGURE 3: Shapley value explanation.

prediction, for instance, would be calculated as 25% + (-20%) = 5%. Conversely, if the variable value was in bin 5, we would not expect the variable to influence the average prediction (i.e., 25%) because the Shapley value in bin 5 is equal to 0. From this example, we can infer that the higher the variable bin, the higher the Shapley value for the specific instance and thus, the probability of default increases.

The research conducted by Du Toit et al. (2023) evaluates Shapley values as an ML interpretability technique for credit scoring models. The authors test the technique on simulated data generated from various underlying distributions representing real world credit variables.

The Shapley values are generated after fitting both traditional models and ML models. The authors compared the resulting Shapley value to a well-known measure called the Weights of Evidence (WOE) value to evaluate the technique. The interested reader can consult Siddiqi (2006) for more information on this measure.

These two metrics are compared using Spearman's correlation and the mean squared error between the standardised values of the two metrics. The results show that the Shapley value can explain ML models similarly to the WOE value. The study shows that the Shapley values represent the relationships and interactions simulated in the data. The study encourages model validation teams to use the technique to explore acceptable thresholds for the Shapley value explanation.

Integration of Shapley values into model validation frameworks for machine learning models

To illustrate that a new technique can be integrated into an existing model validation framework, we refer to the model validation process as proposed by De Jongh et al. (2017). Recall that the three distinct elements in the model validation process were:

- conceptual soundness and developmental evidence
- process verification and ongoing monitoring
- outcomes analysis.

The authors propose a model validation process scorecard to determine if the best practice model validation framework has been adequately assembled and implemented. The idea is to rate the overall scorecard a point out of 4, where 1 represents no evidence and 4 indicates full evidence. The authors note that the validation process scorecard consists of seven elements, as Table 1-A1 indicates. This generic scorecard will require changes depending on the product, the institution and the phase at which the process is being performed, that is development, implementation or monitoring. To this end, the following case study shows how a traditional model validation framework can be retained and enhanced to cater for ML models. Note that the case study is aimed to provide model validation teams with a methodology of validating ML models in credit scoring. The data used in this section are hypothetical and assume a model has been selected and trained on a set of data. The tests proposed are based on hypothetical data to demonstrate how this methodology can be followed. It is proposed that a practitioner uses their own model development data or model training data and validation/testing data set to conduct the proposed tests.

Case study

The model validation team of Bank ABC is tasked to validate a credit scoring model that will be used to determine the probability of default of prospective clients. The model selection and development process has been completed, and the model development team decided to implement a Random Forest[2] model based on an extensive list of selection criteria. The team identified transparency as one of the key principles that need to be adhered to when implementing ML models in the financial sector. For the purpose of this case study, the focus of the validation will be on formulating tests that will prove the transparency of the ML model. All other typical validation steps, such as testing population stability and testing accuracy, are considered out of scope for the particular task at hand. The team can use the existing model validation framework where applicable.

Proposed steps to integrate transparency tests into existing model validation framework

A possible solution to the following case study could be to utilise an existing model validation framework and integrate new research on the best practices for model validation processes in ML models. To illustrate this concept, the traditional model validation process proposed by De Jongh et al. (2017) will be used as a starting point.

The following steps outline the proposed process of incorporating new model validation steps into existing model validation frameworks in a practical manner:

- Identify the general principle(s) for the use of AI in the financial sector, which needs to be incorporated into the model validation process: This could be one or more principles depending on the state of the existing model validation process. In some cases, there might not be an

[2] Any ML model could have been used as the framework, and is not limited to random forests only.

existing model validation process available. For this specific use case, transparency is selected as the principle to incorporate.

- Identify and describe the existing model validation process:
  - Identify all the elements of the current model validation process and the typical criteria that relate to the principle as identified in step 1.
  - In some cases, there might not be an existing model validation process available. In these cases, a high level process should be established that specifies the most notable validation elements that need to be tested.
  - Typical validation elements within a model validation process include (De Jongh et al. 2017):
    - understanding and evaluating the model paradigm
    - ensuring model methods/theory is based on sound assumptions
    - determining if the model design is appropriate
    - testing data/variables used in the model
    - evaluating algorithms and codes used to develop the model
    - understanding output generated by the model
    - assessing how the model will be monitored.
- Determine which criteria within the existing model validation process can be used to evaluate the chosen ML model risk principle. Expand on this criteria where necessary:
  - As highlighted in the 'Overview of traditional model validation process', Miller (2017) explains that interpretability can be understood as the degree to which a human can understand the reason for a certain decision. Furthermore, Kim et al. (2016) define it as the degree to which a human can consistently predict a model's result. To showcase Shapley values, the focus will be on developing criteria for understanding model output, one of the main elements of a typical model validation process.
  - The following criteria, as proposed by De Jongh et al. (2017), can be used to evaluate the transparency of a model within the model output element:
    - was model output benchmarked against best practice models (e.g., against a vendor model using the same input data set)?
    - was the reasonableness and validity of model outputs assessed?
    - has a comparison of model outputs against actual realisations been performed? (Commonly referred to as 'back testing'.)
    - has a range of outputs been examined versus a range of inputs – are solutions continuous or jagged? What is the behaviour of hedging quantities and/or derived quantities over the same range?
    - are all results repeatable? (e.g., Monte Carlo simulations)
  - Additional criteria that can be added to this list are:
    - can the model prediction be explained to stakeholders on a global level?
    - can the model prediction be explained on a local level for a specific instance?
    - what is the certainty to which the model prediction can be interpreted?
    - how much variability is present in the model output over time?
    - does the model output make logical sense when compared to assumed outcomes gained from business expertise and experience developing similar models?

In this section, certain key criteria used to test model transparency are obtained from an existing model validation process. The focus is placed on model output as a key validation element to evaluate the transparency of a model. The next section proposes practical tests using Shapley values that model practitioners can perform to evaluate the transparency of ML models in credit scoring.

Ethical considerations

Ethical approval to conduct this study was obtained from the Faculty of Natural and Agricultural Sciences Ethics Committee (FNASREC), North-West University (NWU–01251–23–A9).

Results

Shapley value tests

The previous section assessed the existing model validation process and identified validation elements that need to be considered when evaluating transparency as a key ML model risk principle. The next step of the process suggests tests that will provide evidence to meet the highlighted criteria of transparency. Before we propose the various tests, it is important to note that these tests are performed in the model validation process. Although Shapley values can be used at various stages in the model development process, these tests aim to provide transparency to model predictions by explaining the model output. These tests assume that the variable selection, model selection and training, model parameter tuning and basic performance checks have been completed. The trained model is used to generate the Shapley values, which are used to perform the various tests proposed in the next section.

Proposed structure of Shapley value test

The following tests are designed in the format of a report to ensure that they can be included in the model documentation. The tests illustrate through certain calculations and graphs how Shapley values can be used to prove that the model output is logical, accurate, stable and, therefore, transparent and trustworthy. The first three tests (Tests 1–3) are variable-specific and are designed to analyse variable characteristics. This is similar to the typical scorecard development steps proposed by Siddiqi (2006), which is called variable characteristics analysis. The next two tests (Tests 4–5) are model prediction analyses and focus on providing validation on a model level, thus focussing on the global explanation of the final model prediction.

To perform the tests, certain prerequisite steps have to be completed. These steps are listed below:

- Fit the final[3] model with the development train sample and include all variables selected through variable selection techniques. Note that the development train sample is a subset of the development sample and is used to train the model. Similarly, the development test sample is also a subset of the development sample used to test the model's performance. Additionally, the out-of-time sample is an independent sample that either precedes or succeeds the development sample. This sample serves as another sample that the model developer can use to test accuracy and stability over a different period of time.
- Generate the Shapley values for the model based on the development train sample. This value is based on all observations in the development sample and will give a Shapley value per observation/row for the specified variable.
- In the case of continuous variables, the variable can be binned and the average Shapley value can be calculated to create a summarised view of the Shapley value for a range (bin) of values. Binning is a technique that is commonly used in scorecard development, see Siddiqi (2006). Although the binned results are not used as input in the model, it creates a convenient way of summarising model output for visual reports.

Variable characteristics analysis

Test 1: Shapley value versus default rate (per variable bin)

The first test visualises the Shapley value and default rate per variable bin. The Shapley value is calculated from the development sample as mentioned in the prerequisite steps. These values are grouped per variable bin (in the case of continuous variables), and the average of the binned group is reported on. The same steps are followed to obtain the default rate per bin. This test is shown with an example in Figure 4a.

Test 1 shows how the average Shapley value generated with the development sample compares to the default rate of the development train, development test and validation samples. It is expected that the average Shapley value follows a similar trend to the default rate, considering the Shapley value explanation provided in the section titled 'Shapley value explanation'. The test explains the variable's expected impact on the final prediction depending on the bin from where the observation originates. Furthermore, the tests enable us to inspect if a logical trend is present for the variable under consideration. This is a very important step, as noted by Siddiqi (2006).

An additional test that accompanies Test 1 is a correlation report as shown in Table 3.

Table 3 shows the Spearman and Pearson[4] correlation between the average Shapley value and the development train, development test and validation default rates. This illustrates how closely aligned the average Shapley value trend is to the three[5] default rate trends.

Test 2: Shapley value stability over time (per variable bin)

The second test measures the Shapley Value Stability Over Time (per variable bin). The Shapley value is calculated from the development train sample. These values are grouped per variable bin (in the case of continuous variables), and the average of the binned group is reported. The data is grouped by month for the development train sample, and the results can be depicted in a stacked bar graph, as seen in Figure 4b.

Test 2 highlights an important step in assessing the predictions' continued transparency, namely the predictions' stability over time. The test shows how stable the Shapley values are and determines how much the value fluctuates throughout the year per variable bin. This can point out certain seasonal trends that could be present in the data, which the model development team might want to cater for.

Test 3: Shapley value versus percentage of population (per variable bin)

The third test shown in Figure 4c compares the Shapley Value to the population distribution, expressed as the proportion of the population per variable bin. The Shapley value is calculated from the development train sample and grouped per variable bin (in the case of continuous variables), and the average of the binned group is reported. The average Shapley value is plotted against the volume of observations in each variable bin for the development train, development test and validation samples.

Test 3 shows the distribution of population within the 10 variable bins and overlays the average Shapley value per bin. This test shows the expected impact of a certain Shapley value prediction on the overall population, considering the size/proportion of the group it relates to. This test could point out illogical trends in the Shapley value and its impact on the overall population. This test is aimed at providing transparency by assigning the size of the impact on the Shapley value contribution for a certain subgroup in the population.

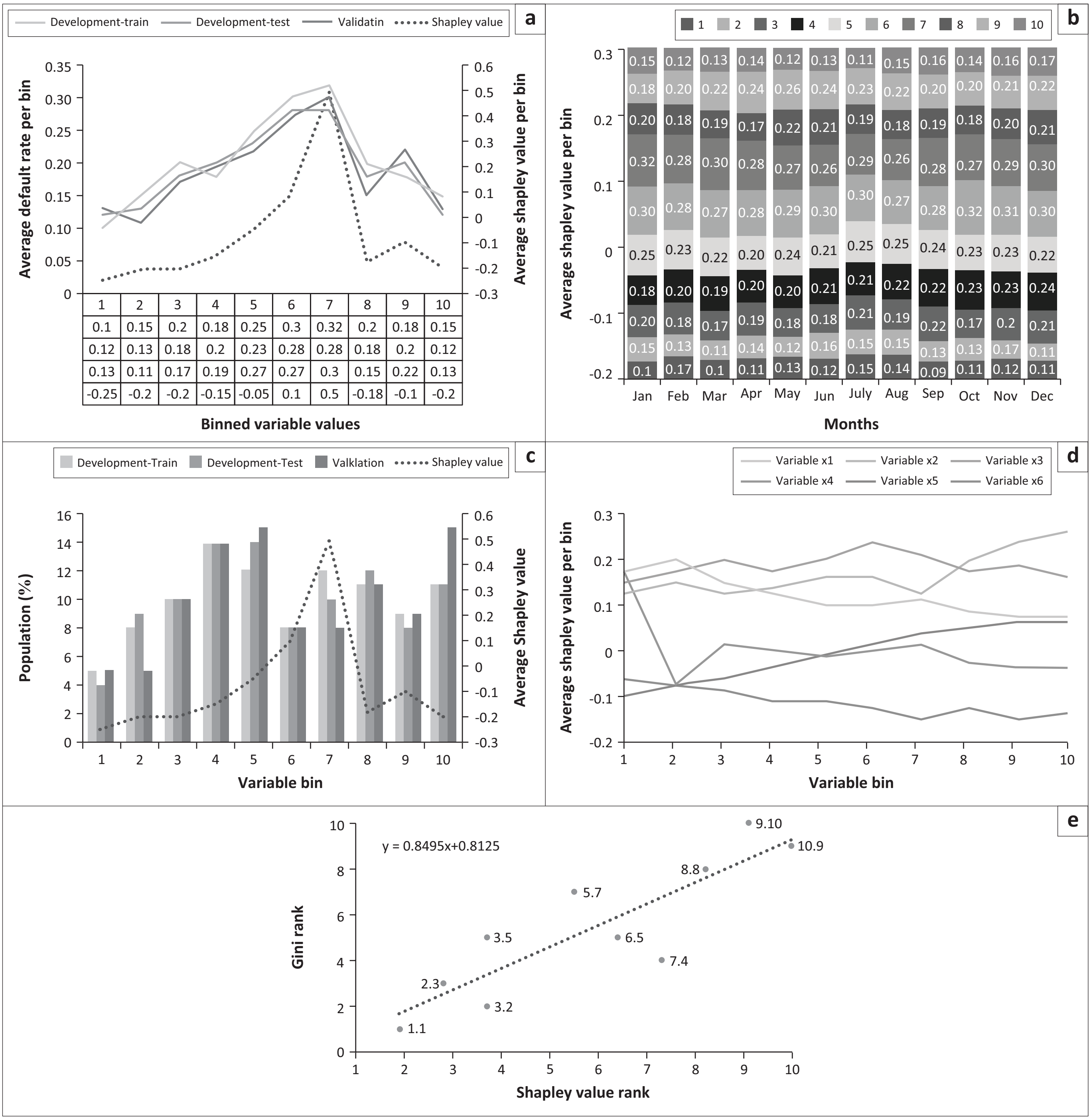
Model prediction analysis

The variable characteristics analysis focussed on tests that could be performed per variable. This section introduces a test that can be used on a model prediction level, that is understanding what influences the final model prediction.

[3] As per the case study above, the final model is a Random Forest model and it is assumed that the software used to fit the model can be used to generate Shapley values.

[4] The normality assumption should be tested before the correlation is calculated. If both the Shapley values and the default rates are normally distributed, then the Pearson correlation is sufficient. Otherwise, Spearman rank correlation is recommended.

[5] Note that the default rates are based on actual default and thus give a good indication of how accurate the Shapley value trend is tracking the actual default rate.



RHS, right-hand side.

FIGURE 4: Shapley value tests 1–5: (a) Test 1: Shapley value (RHS) versus default rate (per variable bin) (b) Test 2: Shapley value stability over time (per variable bin) (c) Test 3: Shapley value (RHS) versus percentage of population (d) Test 4: Top x positive and negative contributing variables (e) Test 5: Gini rank versus absolute Shapley value rank.

TABLE 3: Test 1: Correlation report.

Results	Development-train	Development-test	Validation
Pearson Correlation	0.86	0.82	0.87
Spearman Correlation	0.86	0.94	0.89

Test 4: Top x positive and negative contributors

The fourth test shown in Figure 4d shows the top *x* variables that positively (decrease default prediction) and negatively impact (increase default prediction) the final model prediction. The Shapley value is calculated from the development train sample. These values are grouped as before. The average Shapley value is plotted per variable bin for the top x positive and negative contributors.

Test 4 gives a high-level explanation of the final model prediction by visualising the top *x* contributing variables that positively and negatively impact the model prediction. This test can add significant value when a challenger model is being developed. Using this plot, the model developer can compare the contributions of the same subset of variables for the challenger model and investigate differences between the variable contributions of the models. This will highlight key differences between the champion and challenger models, and guide the development team in choosing the best model for the specific purpose. While not an extensive explanation, these visuals can be used to explain core drivers in the final prediction to business stakeholders.

Test 5: Gini rank versus absolute Shapley value rank

Test 5 shown in Figure 4e is used to determine if the Shapley value contribution per variable ranks similarly to the Gini value calculated for each variable. To calculate this test, the Gini value is calculated per variable. To calculate the Gini value, the model prediction and actual default rate are required. Subsequently, the Shapley value contribution for each variable is calculated. Note that both these calculations are performed on the development train dataset.

Finally, the rank of the Gini and the absolute Shapley value is determined with respect to the subset of variables selected for the model. The absolute Shapley value is used because the value can be both positive and negative. For this test, we are not interested in the direction (positive or negative) of the impact but rather the magnitude of the impact. The rank of the Gini versus the rank of the absolute Shapley value is compared and illustrated visually.

This test evaluates the assumption that high contributing Shapley values are good in discriminating between default and non-default events. If a strong correlation between the rank of the Gini and the rank of the absolute Shapley does not exist, further investigation is warranted to understand why the assumption is not evident in the data.

A common misperception when explaining ML model predictions is that the model outcome can be explained through one metric. To our knowledge, such a golden standard does not exist, at least not one that is model-agnostic. Although the tests proposed above offer practical examples for the model development team, it's important to acknowledge their limitations. For instance, the process can be time-consuming as each variable needs to be evaluated individually. Additionally, certain variable assumptions that are not met may be difficult to explain and may take some time to investigate. The test can be modified and improved to meet the requirements of the relevant stakeholders. The final test will be different for specific use cases and different audiences because the relevant stakeholders will be responsible for approving the final model.

Conclusion

The consensus has been that introducing effective methods for interpreting ML models is widely regarded as a crucial step required for the validation process in credit scoring. This article aims to contribute to this research by comparing traditional model validation processes to more recent proposed frameworks that include ML models. Furthermore, the article aims to compare these two methodologies and identifies transparency as one of the key elements to instil trust in ML models. The comparative study suggests that many of the same criteria still apply to ML models, the only difference being the methods and techniques to evaluate the criteria that need to be adjusted and/or extended for ML models.

This study motivates, through prior research such as Du Toit et al. (2023), that Shapley values hold immense potential in explaining ML models on the level of detail comparable to that provided by well-known scorecard metrics, such as the WOE metric. A list of validation elements related to ML transparency is identified, and our research guides practitioners on how Shapley values can be used to evaluate the criteria within the model validation elements practically.

This article illustrates how the typical model validation practitioner can integrate existing validation frameworks with the principle based guidelines proposed by recent research. It showcases the usability of techniques such as Shapley values and illustrates the importance of maintaining model validation processes that have proven very successful. The novelty in this research is that an interpretability technique is proposed specifically for credit scoring model validation in the banking sector.

Acknowledgements

Competing interests

The authors declare that they have no financial or personal relationships that may have inappropriately influenced them in writing this article.

Authors' contributions

H.A.d.T., W.D.S. and H.R. contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript.

Funding information

The authors received no financial support for the research, authorship, and/or publication of this article.

Data availability

Data sharing is not applicable to this article, as no new data were created or analysed in this study. The data supporting the findings of this study are available within the article.

Disclaimer

The views and opinions expressed in this article are those of the authors and are the product of professional research. The article does not necessarily reflect the official policy or position of any affiliated institution, funder, agency or that of the publisher. The authors are responsible for this article's results, findings and content.

References

Abrahams, C.R. & Zhang, M., 2008, *Fair lending compliance: Intelligence and implications for credit risk management*, Wiley, Hoboken, NJ.

Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A. et al., 2020, 'Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI', *Information Fusion* 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Baesens, B., Roesch, D. & Scheule, H., 2016, *Credit risk analytics: Measurement techniques, applications, and examples in sas*, Wiley, Hoboken, NJ.

BCBS, 2005, *Studies on the validation of internal rating systems (revised)*, Working paper 14, Bank for International Settlements, viewed 13 November 2023, from https://www.bis.org/publ/bcbs_wp14.htm.

Bertsimas, D., King, A. & Mazumder, R., 2016, 'Best subset selection via a modern optimization lens', *The Annals of Statistics* 44(2), 813–852. https://doi.org/10.1214/15-AOS1388

Chen, H., Lundberg, S., and Lee, S.-I. (2021). 'Explaining models by propagating Shapley values of local components', in A. Shaban-Nejad, M. Michalowksi, & D.L. Buckeridge (eds.), *Explainable AI in Healthcare and Medicine: Building a Culture of Transparency and Accountability*, pp. 261–270. Studies in Computational Intelligence, Volume 914. Springer.

De Jongh, P.J., Larney, J., Mare, E., Van Vuuren, G.W. & Verster, T., 2017, 'A proposed best practice model validation framework for banks', *South African Journal of Economic and Management Sciences* 20(1), a1490. https://doi.org/10.4102/sajems.v20i1.1490

Du Toit, H.A., Schutte, W.D., Raubenheimer, H., 2023, 'Shapley values as an interpretability technique in credit scoring', PhD thesis, Centre for Business Mathematics and Informatics, North-West University, Potchefstroom, South Africa.

European Banking Authority (EBA), 2011, *Guidelines on Internal Governance (GL 44)*, viewed 13 November 2023, from https://www.eba.europa.eu/regulation-and-policy/internal-governance/guidelines-on-internal-governance.

European Banking Authority (EBA), 2021, *Discussion paper on machine learning for IRB models – European Banking Authority*, viewed 01 May 2023, from https://www.eba.europa.eu/regulation-and-policy/model-validation/discussion-paper-machine-learning-irb-models.

European Commission, 2024, 'Artificial Intelligence Act', *Official Journal of the European Union*, viewed n.d., from https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

Härle, P., Havas, A., Kremer, A., Rona, D. & Samandari, H., 2015, *The future of bank risk management*, viewed 01 May 2023, from https://www.mckinsey.com/~/media/mckinsey/dotcom/client_service/risk/pdfs/the_future_of_bank_risk_management.pdf.

Hong Kong Monetary Authority (HKMA), 2006, *Validating risk rating systems under the IRB approaches*, viewed 13 November 2023, from https://www.hkma.gov.hk/media/eng/doc/key-functions/banking-stability/supervisory-policy-manual/CA-G-4.pdf.

IIF & EY, 2022, *IIF and EY survey report on machine learning white paper*, viewed 13 November 2023, from https://www.iif.com/portals/0/Files/content/32370132_iif_and_ey_survey_report_on_machine_learning_-_uses_in_credit_risk_and_aml_applications_-_public_summary.pdf.

Kim, B., Khanna, R. & Koyejo, O., 2016, 'Examples are not enough, learn to criticize! Criticism for interpretability', in *Advances in neural informatio'n processing systems 29 (NIPS 2016)*, viewed 15 January 2024, from https://papers.nips.cc/paper_files/paper/2016/hash/5680522b8e2bb01943234bce7bf84534-Abstract.html.

KPMG, 2022, *Modern risk management for AI models – Re-imagining the model risk management function for artificial intelligence*, KPMG white paper, viewed 19 March 2023, from https://assets.kpmg.com/content/dam/kpmg/xx/pdf/2022/07/modern-risk-management-for-ai-models.pdf.

Kumar, E., Venkatasubramanian, S., Scheidegger, C. & Friedler, S., 2020, 'Problems with shapley-value-based explanations as feature importance measures', in H. Daumé III & A. Singh (eds.), *37th International Conference on Machine Learning, online*, July 13–18, 2020, pp. 5491–5500.

Lessmann, S., Baesens, B., Seow, H. & Thomas, L., 2015, 'Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research', *European Journal of Operational Research* 247(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Lundberg, S., Erion, G. & Lee, S., 2018, *Consistent individualized feature attribution for tree ensembles*, Working paper, viewed 13 November 2023, from https://arxiv.org/pdf/1802.03888.pdf.

Lundberg, S. & Lee, S., 2017, 'A unified approach to interpreting model predictions', in I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, et al. (eds.), *31st International Conference on Neural Information Processing Systems (NIPS'17) proceedings*, Curran Associates Inc., 4 December 2017, pp. 4768–4777.

Miller, T., 2017, *Explanation in artificial intelligence: Insights from the social sciences*, Working paper, viewed 13 November 2023, from https://arxiv.org/pdf/1706.07269.pdf.

Molnar, C., 2020, *Interpretable machine learning. A guide for making Black Box models explainable*, viewed 12 November 2023, from https://christophm.github.io/interpretable-ml-book/.

Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C.A. et al., 2020, *General pitfalls of model-agnostic interpretation methods for machine learning models*, Working paper, viewed 13 November 2023, from https://arxiv.org/pdf/2007.04131.pdf.

Monetary Authority of Singapore (MAS), 2018, *Principles to promote Fairness, Ethics, Accountability and Transparency (FEAT) in the use of artificial intelligence and data analytics in Singapore's financial sector*, MAS white paper, viewed 13 November 2023, from https://www.mas.gov.sg/publications/monographs-or-information-paper/2018/FEAT .

Office of the Comptroller of the Currency (OCC), 2011, *Supervisory guidance on model risk management*, Supervisory document OCC 2011–12, Board of Governors of the Federal Reserve System, Washington, DC, pp. 1–21.

PWC, 2018, *Building trust in AI and data analytics*, viewed 20 April 2021, from https://www.pwc.com/sg/en/publications/assets/building-trust-ai-data-analytics-122018.pdf.

Quell, P., Bellotti, AG., Breeden, J.L. & Martin, J.C., 2021, *Machine learning and model risk management*, Model Risk Managers' international association (MRMIA) white paper, viewed 18 March 2023, from https://mrmia.org/wp-content/uploads/2021/03/Machine-Learning-and-Model-Risk-Management.pdf.

SARB, 2015, *Directive 4/2015: Amendments to the regulations relating to banks, and matters related thereto*, viewed 10 August 2016, from https://www.resbank.co.za/Lists/News%20and%20Publications/Attachments/6664/D4%20of%202015.pdf.

Scheda, R. & Diciotti, S., 2022, 'Explanations of machine learning models in repeated nested cross-validation: An application in age prediction using brain complexity features', *Applied Sciences* 12(13), 6681. https://doi.org/10.3390/app12136681

Shapley, L., 1953, 'A value for n-person games', in K. Harold William (ed.), *Contributions to the theory of games*, vol. 2, pp. 307–317, Princeton University Press, Princeton.

Siddiqi, N., 2006, *Credit risk scorecards: Developing and implementing intelligent credit scoring*, John Wiley & Sons, Hoboken, NJ.

Siddiqi, N., 2017, *Intelligent credit scoring : Building and implementing better credit risk scorecards*, John Wiley & Sons, Hoboken, NJ.

Van der Burgt, J., 2019, *General principles for the use of AI in the financial sector*, DeNederlandscheBank white paper, viewed 13 November 2023, from https://www.dnb.nl/media/voffsric/general-principles-for-the-use-of-artificial-intelligence-in-the-financial-sector.pdf.

Woźnica, K., Pękala, K., Baniecki, H., Kretowicz, W., Sienkiewicz, E. & Biecek, P., 2021, *Do not explain without context: Addressing the blind spot of model explanations*, Working paper, viewed 13 November 2023, from https://arxiv.org/pdf/2105.13787.pdf.

Appendix starts on the next page →

Appendix 1

Table 1-A1 describes the seven elements contained in the model validation process scorecard as proposed by De Jongh et al. (2017) in more detail.

TABLE 1-A1: Model validation process scorecard as proposed by De Jongh et al. (2017).

Validation process	Score			
	1: No evidence	**2: Due consideration lacking**	**3: Some consideration**	**4: Fully evident**
Paradigm				
To what extent was the conceptual soundness of paradigm checked?				
To what extent was the review performed by suitably skilled experts?				
Methods or theory				
To what extent is the underlying model theory consistent with published research and sound industry practice?				
To what extent were research publications considered of appropriate quality/standing?				
To what extent was the methodology benchmarked against appropriate industry practice?				
To what extent are approximations made within agreed tolerance levels?				
Design				
To what extent was it ascertained that assumptions are clearly formulated?				
To what extent was the appropriateness and the completeness of assumptions checked?				
To what extent was it checked that all variables employed have been clearly defined and listed?				
To what extent have the causal relationships between variables been noted?				
To what extent have input data been assessed in terms of reasonableness, validity and understanding?				
To what extent has it been ascertained that outputs are clearly defined?				
To what extent has the design been evaluated in terms of over-complexity/over-simplification?				
To what extent has the model builder benchmarked the design against existing best practice models?				
To what extent was the design independently benchmarked against existing best practice models?				
To what extent have special cases been dealt with appropriately? (e.g. terminal conditions or products with path-dependent pay-off)				
Data or variables				
To what extent have input data been checked to gauge reliability/suitability/validity/completeness?				
To what extent has it been checked that data involving subjective assessment of expert opinion been appropriately incorporated?				
To what extent was the procedure for the collation of expert opinion scrutinised?				
To what extent has expert opinion been validated in terms of logical considerations?				
To what extent has the expert selection process been assessed as sound?				
To what extent was it verified that data are representative of relevant (general and stressed) market conditions?				
To what extent was it verified that data are representative of the company's portfolio?				
To what extent have inadequate or missing data been re-assessed and reviewed for model feasibility?				
Algorithms or code				
To what extent was the algorithms/code checked against the model formulation and underlying theory?				
To what extent were key assumptions and variables analysed with respect to their impact on model outputs?				
To what extent was an independent construction of an identical model undertaken?				
To what extent was the code rigorously tested against a benchmark model?				
To what extent was technical proofreading of the code performed?				
Outputs				
To what extent was model output benchmarked against best practice models (e.g. against a vendor model using the same input data set)?				
To what extent was the reasonableness and validity of model outputs assessed?				
To what extent has a comparison of model outputs against actual realisations been performed? (backtesting)				
To what extent has a range of outputs been examined vs. a range of inputs (e.g. are solutions continuous or jagged? What is the behaviour of hedging quantities and/or derived quantities over the same range?)				
To what extent are all results repeatable? (e.g. Monte Carlo simulations)				
Monitoring				
To what extent has the model been monitored for appropriate implementation and use?				