RISK MANAGEMENT MAGAZINE

Vol. 20, Issue 1
January – April 2025

EXCERPT

https://www.aifirm.it/rivista/progetto-editoriale/

Audit Program on Artificial Intelligence (AI)-driven Credit Risk

Valeria Anna De Palma, Alessandro Di Maria, Daniele Foschini, Vincenzo Frasca, Dario Girardi

Audit Program on Artificial Intelligence (AI)-driven Credit Risk Models

Valeria Anna De Palma (UniCredit SpA), Alessandro Di Maria (UniCredit SpA), Daniele Foschini (UniCredit SpA), Vincenzo Frasca (UniCredit SpA), Dario Girardi (UniCredit SpA)[1, 2]

Corresponding Author: Dario Girardi (dario.girardi@unicredit.eu)

Article submitted to double-blind peer review, received on 17th Decembre 2024 and accepted on 8th March 2025

Abstract

From an Internal Audit perspective, the integration of Artificial Intelligence (AI) into credit risk modelling through Machine Learning (ML) algorithms presents significant challenges due to the complexity and multidimensional nature of these models. While AI enhances predictive performance and accuracy, its inherent lack of transparency and explainability increases the risk of control deficiencies, potentially leading to financial losses, misrepresentation of information, unfair discrimination against debtors, and non-compliance with EU regulations. This paper introduces a comprehensive audit framework designed to establish robust internal controls over AI-driven credit risk models. Aligned with the Model Risk Management (MRM) lifecycle, we propose a structured set of audit tests and controls, organized by thematic area, to assess key aspects such as model design and performance, governance, reliability, and regulatory compliance. Additionally, we provide practical examples in emerging areas to illustrate their application. These audit procedures aim to identify critical vulnerabilities while ensuring adherence to regulatory standards, including EBA/REP/2023/28 and the evolving requirements of the EU AI Act.

Keywords: Internal audit, Credit Risk, Artificial Intelligence, EU regulation, Model Risk

JEL Classification: M42, G32, C45, C49

1. Introduction

In recent years, the banking system has recognized the need to improve creditworthiness evaluation in terms of precision, accuracy, and responsiveness to detect credit distress, ensuring more efficient and controlled credit processes. A key advancement in addressing this challenge has been the introduction of Machine Learning (ML) algorithms. ML has become an essential tool in credit risk management, allowing banks to assess obligors more effectively throughout the entire credit lifecycle. Unlike traditional statistical methods, ML enhances predictive accuracy, processes large volumes of data, identifies complex patterns, and quickly adapts to new information. Notably, ML models leveraging high-frequency transactional data have demonstrated superior precision and discriminatory power, even in cases with limited information (Moscatelli *et al*, 2019). Despite these advantages, integrating ML into credit risk assessment introduces new challenges, especially when it comes to internal control framework and particular 3rd level/audit controls. The dynamic nature of ML models, which continuously learn from and adapt to new data, raises concerns about consistency and reliability. Additionally, the complexity of these algorithms can lead to outcomes that are not always easily understandable, earning them the label of "black-boxes" – models where the inputs and outputs are known, but the internal workings remain largely opaque (Giudici and Gramegna, 2021). This poses a trade-off between increasing accuracy on the one hand, and interpretability on the other, of predictions made by the model, which must be balanced according to the model purpose and the context in which it is used. Finally, the regulatory landscape for credit risk management and the usage of AI techniques is stringent and constantly evolving, with regulatory bodies like the EU Parliament and the European Banking Authority setting comprehensive standards. Compliance with these regulations demands transparency, accountability, and rigorous validation of credit risk models, including evaluating the impact of ML on the credit assessment process.

Consequently, the inherent complexity of ML models presents unique challenges in meeting the aforementioned requirements. Effective audit controls must address not only the technical aspects of model assessment but also the governance framework and its role within credit processes, particularly in relation to the parallel adoption of IRB/regulatory models. So far, banks' Internal Validation functions have primarily focused on assessing regulatory models due to the constraints imposed by regulatory validation and maintenance requirements. Within this context, and given the currently limited oversight provided by the second-level control function for ML credit risk models, the role of Internal Audit becomes crucial in mitigating residual risks. Therefore, it is essential to establish a comprehensive audit framework capable of assessing the complexity of these models from multiple perspectives (e.g., quantitative, qualitative, etc.).

As reported by Clark (2018), various approaches exist for conducting a Machine Learning audit. These range from comprehensive code reviews – involving an examination of underlying mathematical assumptions and relevant human interventions, constituting a highly technical methodological evaluation of the ML algorithm – to approaches focused exclusively on assessing the ethical implications of ML. For instance, one assessment framework evaluates the different stages defined in the *CRISP-DM (Cross-Industry Standard Process for Data Mining)* model, a robust and systematic framework for data mining projects. This method enables evaluation across six fundamental stages: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. The depth of analysis at each stage may vary depending on the auditors' focus and expertise ([3]).

Another framework, known as *SMACTR (Scoping, Mapping, Artefact Collection, Testing, and Reflection)*, is designed to assess the compliance of AI algorithms with an organization's ethical values. This framework aims to ensure that AI systems are developed

[1] The views and opinions expressed in this paper are solely those of the author(s) and do not necessarily represent the official stance of UniCredit.

[2] We would like to extend our sincere gratitude to *Deloitte Financial Advisory s.r.l. S.B* for their thorough review and invaluable support in the preparation of this paper. Your expertise and insights have been instrumental in enhancing the quality of our work.

[3] For further details on *CRISP-DM*, see Clark (2018).

and implemented in alignment with ethical principles such as transparency, justice, fairness, and non-discrimination (Sandu *et al*, 2022; Inioluwa *et al*, 2020).

Despite their strengths, such approaches have significant limitations, as they primarily focus on specific aspects of Machine Learning models. For instance, frameworks like CRISP-DM do not adequately address the risks associated with using these models in the credit assessment process. Moreover, they often operate at too high level, concentrating solely on regulatory compliance without offering detailed steps for auditing the algorithms or the processes they underpin. These shortcomings make such frameworks insufficient for providing a comprehensive audit assessment that ensures compliance with the stringent requirements of the EU AI Act.

For this reason, we propose, as outlined by Sandu *et al* (2022), an audit framework grounded in the phases of the credit model lifecycle as defined in Model Risk Management. This approach facilitates a comprehensive evaluation of model risks, addressing both errors inherent to the model—such as incorrect data or flawed design—and risks arising from the misuse or misapplication of the model. It also includes an assessment of the key stakeholders within the Model Risk Management (MRM) framework, focusing on their roles and responsibilities. Furthermore, as elaborated in the following paragraphs, this framework can systematically incorporate, on top of already existing risks, the consideration of ethical risks—such as the potential for unfair discrimination—at every stage of the model's lifecycle ([4]).

It is important to note that the audit framework presented in this paper can be considered as a "baseline" to address the unique characteristics of AI-driven credit risk models used for both regulatory and managerial purposes. Nevertheless, it is essential to recognize that when AI and ML techniques are applied to regulatory models—such as Internal Ratings-Based (IRB) models used for capital requirements calculation—the internal audit framework must include additional tests beyond those discussed in this paper. These supplementary evaluations are necessary to ensure the compliance of such models with relevant regulatory requirements. Moreover, this paper tackles the distinct challenges posed by Machine Learning, including model explainability, fairness, performance, and regulatory compliance. It proposes a comprehensive framework for establishing robust audit controls. This framework is specifically designed for banks and financial intermediaries, such as fintech firms, that utilize machine learning model in the credit assessment. Given the growing reliance on ML algorithms in credit risk evaluation, defining a focused approach to auditing these models is crucial. By addressing these challenges, financial institutions can maximize the potential of machine learning while safeguarding the integrity and reliability of their credit risk assessment processes.

2. Regulatory requirements for Machine Learning in credit risk

In light of the widespread diffusion of Artificial Intelligence in recent years, regulators and policymakers across the globe have tried, at different paces and with different levels of restrictiveness, to develop new regulatory frameworks with the aim of fostering the diffusion of AI-based solutions while ensuring a correct use by companies and guaranteeing that customers and society as a whole can benefit from fair and transparent AI technology.

At a global level, the European Union has acted as a pioneer in the regulation of AI, launching several years ago a regulatory process aimed at promoting the safe and responsible development and diffusion of AI-based applications. This process culminated in July 2024 with the publication of the final version of the so-called AI Act ([5]) in the official Journal of the EU. The AI Act defines a risk-based approach to the regulation of AI, modulating its requirements on the basis of the level of risk posed by the specific AI system and tailored to the role of the specific actor in the development/deployment/usage of the system. The regulation focuses on the potential harmful impact that AI solutions might have on the fundamental rights of individuals, introducing the need for a thorough risk assessment and ethical considerations in the development, deployment and usage of AI-based solutions. Following the risk-based approach, the regulation identifies three different categories of AI systems:

1. Prohibited AI systems: these are AI systems characterized by an unacceptable risk, and which are therefore prohibited. Such systems are banned as they contradict the values of the European Union, violating fundamental rights. This category includes, among others, AI systems that can manipulate and distort a person's behavior through subliminal techniques or that exploit a person's vulnerabilities related to age, disability or specific social or economic situation, and AI systems for the evaluation or classification of natural persons and the related attribution of a social score.

2. High-risk AI systems: these are AI systems that pose a high-risk to health and safety or to the fundamental rights of natural persons and are therefore subject to a specific set of requirements, though not being prohibited. The EU AI Act identifies as high-risk systems, among others, the systems listed in Annex III of the Regulation, which include AI systems used in the areas of biometrics, critical infrastructure, education and vocational training, employment and workers management, access to essential private and public services (including healthcare services, credit, life and health insurance), law enforcement, migration, administration of justice and democratic processes. According to this classification, credit risk models which are used to assess the creditworthiness of potential borrowers for the purpose of granting loans fall within the definition of high-risk AI systems and are therefore subject to the related requirements. The AI Act requirements on high-risk AI systems will be summarized later in this paragraph.

3. Low-risk AI systems: these are AI systems that do not fall within the two above categories, and for which compliance with minimum transparency standards is required, for example, concerning the need to inform users that they are interacting with an AI system and, in the case of content generated by AI tools that may be misconstrued as authentic, to disclose that the content has been manipulated or generated using AI tools.

[4] A key component of the MRM framework is the structuring of roles, responsibilities, and accountabilities for decision making, risk control, and governance. There are several ways in which an organization can setup roles. However, it is important that reporting lines and incentives are clear. In a typical MRM framework, the "three lines of defense" model is widely adopted (Satish *et al*, 2016).

[5] *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence* [...].

In addition, it should be noted that the application of the above requirements depends also on the specific role that the organization (i.e., the bank in our case) plays with respect to the AI model. Indeed, the EU AI Act defines several key roles in the AI ecosystem:

- Provider: a natural or legal person, public authority, agency, or other body that is or has developed an AI system to place on the market, or to put into service under its own name or trademark.
- Deployer: a natural or legal person, public authority, agency, or other body using an AI system under its authority.
- Importer: any natural or legal person within the EU that places on the market or puts into service an AI system that bears the name or trademark of a natural or legal person established outside the EU.
- Distributor: any natural or legal person in the supply chain, not being the provider or importer, who makes an AI system available in the EU market.
- Product manufacturer: a manufacturer of an AI system that is put on the market or a manufacturer that puts into service an AI system together with its product and under its own name or trademark.

It is our understanding that the specific requirements applicable depend on the intersection between risk level of the AI system and the role of the organization, as better described below.

Focusing on the *high-risk systems* the EU AI Act defines the following requirements:

1) obligations for *providers:*
    a. General obligations:
        i. "Effective data governance" (art. 10).
        ii. "Maintaining appropriate technical documentation and record-keeping" (articles 11 and 12).
        iii. "Transparency and provision of information to users" (art. 13).
        iv. "Enabling and conducting human oversight" (art. 14).
        v. "Compliance with standards for accuracy, robustness, and cybersecurity for the intended purpose" (art. 15).
        vi. "Establishing and maintaining appropriate AI risk and quality management systems" (art. 17).
        vii. "Registering high-risk AI systems on the EU database before placing them on the market" (art. 49); systems used for law enforcement, migration, asylum and border control, and critical infrastructure will be registered in a non-public section of the database.

    In addition, a conformity assessment (check of compliance) should examine whether the requirements laid out above have been met. In most cases, providers can self-assess. A third-party conformity assessment by an accredited body is required if any of the following criteria apply:

    - The AI system is part of a safety component subject to third-party assessment under sectoral regulations.
    - The AI system is part of a biometric identification system.
    - Harmonized standards are not used.

    b. On-going performance and conformity checks:
        i. "Maintaining logs generated by high-risk systems for a period of at least six months" (articles 9 and 19);
        ii. "Immediately taking the necessary corrective actions for non-conforming systems already on the market and informing other operators in the value chain of the non-conforming systems" (art. 20).
        iii. "Cooperating with the national competent authorities or the AI Office" (art. 21).
        iv. "Monitoring performance and safety of AI systems throughout their lifetime" (art. 72).
        v. "Reporting to the appropriate authorities serious incidents and malfunctions that lead to breaches of fundamental rights" (art. 73).
        vi. "Undergoing new conformity assessments for substantial modifications (e.g., changes to a system's intended purpose or changes that affect how it meets regulations)" (art. 43). For AI systems that are considered to have limited or minimal risk, it will be important to check whether the original risk classification still applies after any changes.

2) obligations for *deployers* (art. 26):
    a. "Completing a fundamental rights impact assessments (FRIA) before putting the AI system in use (relevant for public bodies and private entities providing services of general interest including banks, insurers, hospitals, schools, which are deploying high-risk systems)".
    b. "Implementing human oversight by people with the appropriate training and competence.
    c. "Ensuring that input data is relevant to the use of the system".
    d. "Suspending the use of the system if it poses a risk at a national level".

e. "Informing the AI system provider of any serious incidents".

f. "Retaining the automatically generated system logs".

g. "Complying with GDPR obligations to perform a data protection impact assessment".

3) obligations for *importers* and *distributors* (art. 23):

a. "Verifying the AI system is compliant with the AI Act and that all relevant documentation is evidenced".

b. "Informing people that they might be subject to the use of high-risk AI".

In the case of a bank developing AI models for credit risk, it would primarily fall under the role of a "provider" of a high-risk system, as it develops and deploys the AI systems internally for its own use. As a provider, the bank would be subject to the regulatory requirements outlined above in point 1) for the high-risk systems, including ensuring transparency, accountability, and conformity with high-risk AI system assessments.

In addition to the general provisions provided by the AI Act for high-risk AI systems, in the context of ML-based credit risk models relevant attention must be also paid to the expectations and requirements of banking Supervision Authorities. As a matter of fact, in 2021 the European Banking Authority (EBA) initiated a discussion about potential requirements for AI-based credit risk models used for regulatory purposes (IRB models) ([6]). In 2023, the discussion was followed by the publication of the *Follow-up Report on the use of Machine Learning for IRB Models (EBA/REP/2023/28)* (European Banking Authority, 2023), which defines some principle-based recommendations to be followed by banks if they plan to adopt ML techniques for the purpose of IRB modelling. It should be noted that, while the EBA recommendations specifically refer to IRB models used to calculate regulatory capital requirements, the EU AI Act requirements in the context of credit risk models apply in principle only to models which are used for assessing the creditworthiness of individuals for the purpose of granting loans. However, the "use test" requirements established by the *Capital Requirements Regulation (CRR)* ([7]) and the *ECB Guide to Internal Models* ([8]) (European Central Bank, 2024) impose on banks the use of estimates resulting from their IRB models within their risk management, decision-making and credit approval processes, thus generating a potential overlap between IRB models and AI Act requirements that banks should carefully take into consideration.

Here follows a brief summary of the EBA recommendations related to the use of ML for IRB modelling:

1) All the relevant stakeholders, including the model Development Unit, the Credit Risk Control Unit and the Internal Validation function, should have an appropriate level of knowledge of the model's functioning. In addition, the management body and the senior management should be in a position to have a good understanding of the model and the underlying key drivers.

2) Banks should avoid unnecessary complexity and find an appropriate balance between model performance and explainability of the results.

3) Banks should ensure that the model is correctly interpreted and understood, by analyzing the statistical relationship of risk drivers with the output variable and ensuring that potential biases in the model are promptly detected.

4) Banks should ensure a proper level of understanding of the model by the relevant staff, especially when human judgement and overrides are applied.

5) Banks should ensure a reliable in-depth validation of ML models, addressing overfitting issues, challenging the model design, ensuring a sufficient level of representativeness and data quality and guaranteeing the stability of estimates.

As can be noted from the points above, both the EU AI Act and the EBA recommendations draw specific attention to the risks produced by the AI system, related not merely to its "algorithmic" component, but also to its deployment, usage, and the related governance and personnel involved in all stages of the model life cycle. Finally, additional guidelines on the adoption of AI/ML techniques in credit risk models are expected to be included in the next update of the ECB Guideline on Internal Models, scheduled for the first quarter of 2025.

The regulation of AI systems shares significant touchpoints with personal data protection regulation, such as the European Union's General Data Protection Regulation (GDPR); such interdependencies and potential overlaps should be carefully taken into consideration in the definition of an appropriate audit framework on AI-driven credit risk models.

The GDPR governs the processing of personal data of individuals in the EU, focusing on how such data is collected, stored, and used. Consequently, any AI system that processes personal data must comply with GDPR provisions, particularly those related to transparency regarding the use of such data.

For example, Articles 15 and 22 of the GDPR require that individuals subject to automated decision-making be provided with meaningful information about the logic involved in these processes. This implies that, in the context of AI-based creditworthiness assessment in loan application, a person whose request is denied can ask for information about the explanation of the decision (see also section 3.3.3 for more details on this topic).

[6] This discussion was launched with the publication of the *EBA Discussion Paper on Machine Learning for IRB Models (EBA/DP/2021/04)* (European Banking Authority, 2021).

[7] *Regulation (EU) No 575/2013 of the European Parliament and of the Council of 26 June 2013 on prudential requirements for credit institutions and investment firms.* Art. 144(1)(b) of such regulation states that, among the minimum standards required for IRB approval, competent authorities should verify that "*internal ratings and default and loss estimates used in the calculation of own funds requirements and associated systems and processes play an essential role in the risk management and decision-making process, and in the credit approval, internal capital allocation and corporate governance functions of the institution.*"

[8] Section 6.2 - Use test requirement.

In consideration of what described in the previous parts of this chapter, it is deemed that an MRM-based framework for auditing ML credit risk models is the best and most logical solution to effectively address the new challenges posed by these models while ensuring compliance with relevant regulatory requirements.

3. Credit Risk Audit Framework

3.1 Model Risk Management perspective

According to Sandu *et al* (2022), existing frameworks for auditing AI algorithms primarily focus on assessing specific dimensions, such as ethical risks, including potential discrimination based on protected or sensitive characteristics. However, as noted in the previous paragraph, the new EU AI Act and EBA reports shift the focus toward risks of incorrect outcomes that could harm the rights of individuals, especially for AI-based credit risk models. While these regulations continue to emphasize addressing ethical risks, they also highlight the importance of managing risks from errors in AI algorithms, particularly those stemming from the complexity of AI-based credit risk models.

In this context, it seems evident that in order to assess the risk of incorrect outcomes from credit risk models involving ML technology, the concepts of model risk and Model Risk Management (MRM) process can be applied. Indeed, as defined by the Federal Reserve System (2011), model risk is "*the potential for adverse consequences from decisions based on incorrect or misused model outputs or reports*". This concept is further reinforced by the European Central Bank (ECB) in its Guide to Internal Models (European Central Bank, 2024), which states: "*Effective model risk management allows institutions to reduce the risk of potential losses and underestimation of own funds requirements due to flaws in model development, implementation, or use. To mitigate these risks, institutions should have a model risk management framework in place that allows them to identify, understand, and manage their model risk for internal models across the group*."

By aligning regulatory requirements for AI-based credit risk models with MRM principles, we can leverage the concept of the model lifecycle, as defined in MRM, to systematically identify, measure, and address risks at every phase of the model's lifecycle ([9]). This alignment supports the development of a robust audit framework for AI credit risk models, incorporating tests and controls aimed at effectively mitigating all potential risks associated with these models.

These controls address not only technical risks, such as incorrect or misused model outputs—e.g., misclassifying obligors or misapplying concepts critical to creditworthiness assessments—but also ethical risks specific to AI-based models, such as unfair discrimination against certain groups of debtors.

In our opinion, by leveraging the Model Risk Management (MRM) approach as a starting point, and by integrating the ethical risks dimension on top of risks already assessed by the MRM, we can:

- Effectively address all risk dimensions associated with the model lifecycle, from its initial development through to its final application.
- Evaluate the governance of models and their role in the credit risk process.
- Account for controls implemented at different stages of the model lifecycle, wherever possible relying on analyses conducted by other lines of defence, and focusing audit efforts on areas that provide added value and enhance the organization's operations.

By adopting this approach, we can strengthen the oversight and accountability of AI-based credit risk models, ensuring their reliability, compliance, and ethical integrity across all operational stages.

3.2 Framework for Risk Assessment Across the Model Lifecycle

Our proposed framework divides the credit risk model lifecycle into four categories capable to cover all its phases ([10]), including governance arrangements, development, implementation, validation, usage, review, and monitoring. Each of these areas aim to cover specific risks embedded in the life cycle of the model:

1) Governance and Organization: this area covers the governance structures, policies, procedures, controls, and actors in place to ensure that the model development, implementation, validation, and usage processes function as intended. Its main risk is related to unclear and improper formalization of roles and responsibilities of the different stakeholders involved.

2) Model Design and Application: this area gathers all information and evidence related to the model, including the business need it addresses, the data used, the modelling techniques applied, model formalization, performance expectations, and documentation (covering the model's purpose, methodology, assumptions, and limitations). The main risk is the inaccurate definition of the model design, leading to the risk of inappropriate performance of the model and, finally, to the production of erroneous results, but also covering the assessment of ethical risks posed by the model.

3) Model Validation and Controls: this area includes all the information needed to verify that the model functions as intended. It involves critically challenging the model through "*analysis by objective, informed parties who can identify model limitations and assumptions and recommend appropriate changes*" (Federal Reserve System, 2011). The main risk is that of an inappropriate validation and controls setup for the model under investigation.

[9] This approach is also proposed in Sandu *et al* (2022).
[10] The MRM process can be divided into various areas of activity. In our framework, we identified four distinct areas; however, other authors may define a different number of areas. For more details, refer to the International Professional Practices Framework (2018).

4) Model Use, Monitoring and Review: this area gathers all information and evidence about how the model is used in the credit processes. The main risk is that of an inappropriate use of model results, as well as that of an improper monitoring framework that should ensure that the model works properly over time.

Each of the areas included in our framework and details of the related checks are summarized in the tables below. Again, as already specified within the introduction, it should be noted that the audit framework presented below is intended to cover only the specific peculiarities of AI-driven credit risk models, without encompassing additional areas of analysis that should be considered when dealing with regulatory models used for capital requirements calculation.

In that case, the Internal Audit framework presented below should be integrated with additional tests aimed at assessing the compliance of these models with relevant regulatory requirements, as already done when dealing with traditional statistical models.

Governance and Organization	
Test Objective	Test Description
Governance	Verify: 1) whether a clear definition and segregation of roles and responsibilities exists across all functions involved and throughout every stage of the model lifecycle. 2) the compliance with internal rules and policies related to the development, maintenance, and documentation of Machine Learning Models. 3) the existence of a clear strategy regarding ML credit-risk models and coherence with such a strategy.
Compliance with EU AI Act Requirements	Verify the alignment with EU AI Act requirements in terms of Model Governance, including Information Obligation, Risk Assessment and Model Registry requirements.
Compliance with GDPR Requirements *(only for Individuals models)*	Verify the alignment with EU GDPR requirements in terms of collection, processing, and storage of personal data.

Table 1 - Audit tests in the area of Governance and Organization (source: elaboration of the Authors)

For more practical details on the execution of the checks on “Governance”, see section 3.3.1.

Model Design and Application	
Test Objective	Test Description
Model Design	Verify: 1) the alignment of the model design with its intended purposes and objectives, also considering internal and external factors that may influence its development and performance. 2) the completeness of the model documentation, including description of model parameters (e.g., hyperparameters), assumptions, and limitations (e.g., use cases for the algorithms), and its coherence with the model development steps, ensuring it complies with both internal policies and EU AI Act requirements. 3) whether expert-based choices are adequately justified.
Data Assessment	Verify: 1) the adequacy of the estimation data perimeter and its consistency with the model's objectives. 2) the quality of the data sources used, and the adequacy of the data quality checks performed. 3) the alignment with the EU AI Act regarding the use of non-discriminatory and representative input data.

Data Processing & Features Creation	Verify the adequacy of the data processing (including sampling procedures, data treatment and feature creation) with respect to the methodology used and the supporting rationales.
Methodology	Verify: 1) the alignment of modelling methodologies (e.g., input mapping, data treatment, ML algorithms used) with industry best practices and relevant literature. 2) the consistency between the code workflows in model development and corresponding documentation. 3) the methodology used for the feature selection and final model selection (including module integration). 4) the presence and adequacy of human oversight in accordance with the EU AI Act (such as the involvement of experts with independent feedback on the model development steps and results). 5) the correctness and reproducibility of the model's results.
Model Performance, Interpretability & Fairness	Verify: 1) whether the model maintains adequate performance when applied to alternative datasets (out-of-sample / out-of-time); 2) whether the model's specification remains stable under different distributional assumptions. 3) whether the complexity of the trained model is justified by a greater performance compared to simpler alternatives. 4) whether the model's results are clearly interpretable and aligned with business sense, leveraging on interpretability techniques such as SHAP and LIME. 5) whether the results of the trained model do not discriminate against individuals based on sensitive attributes.
Model Implementation	Verify: 1) whether the UAT activities performed ensure the alignment with business requirements, if the IT documentation is complete to support correct implementation, and the consistency between the code workflows of the model in production and corresponding documentation. 2) whether the model's results in production are consistent with those obtained in the development environment. 3) whether there is an appropriate level of human oversight (e.g., periodic consistency checks evaluated by human beings on the outputs of the model in production) and transparency during the model's production phase.

Table 2 - Audit tests in the area of Model Design and Application (source: elaboration of the Authors)

For more practical details on the execution of the checks on “Model Performance, Interpretability & Fairness”, see section 3.3.2 (for tests 1), 2) and 3)), section 3.3.3 (for test 4) and section 3.3.4 (for test 5), while for details on the execution of the checks on “Model Implementation” see section 3.3.5.

Model Validation and Controls	
Test Objective	Test Description
Validation & Controls	Verify: 1) the adequacy of accuracy controls on data, IT processes and outputs related to the model. 2) whether relevant internal regulation on ML credit risk model validation exists and if the validation has been conducted in compliance with such regulations. 3) whether the validation results have been interpreted correctly and documented comprehensively. 4) whether the findings align with identified gaps also in terms of severity, if a timeline is set to address these gaps, and if there is a system for tracking the resolution of findings.

Table 3 - Audit tests in the area of Model Validation and Controls (source: elaboration of the Authors)

Model Use, Monitoring and Review	
Test Objective	Test Description
Model Use	Verify whether adequate documentation supports the model's use, and whether the people involved are properly trained to enable them to use the model effectively and implement corrective actions in case of anomalies.
Model Monitoring	Verify: 1) whether there is an adequate framework for model monitoring and maintenance, focusing on performance, stability, and interpretability, and ensuring compliance with EU AI Act requirements (e.g., model fairness). 2) whether an appropriate records-keeping system is in place in accordance with the EU AI Act requirements (e.g., maintaining logs for a period of at least six months).
Model Review	Verify whether there is an adequate framework for model review, update and discharging.

Table 4 - Audit tests in the area of Model Use, Monitoring and Review (source: elaboration of the Authors)

For more practical details on the execution of such checks, see section 3.3.6.

As previously stated, a key component of the Model Risk Management (MRM) framework is the structured definition of roles, responsibilities, and accountabilities in decision-making, risk control, and governance for models. While organizations may adopt different role structures, a typical MRM framework in financial institutions classifies stakeholders into three lines of defence, each with distinct roles and responsibilities ([11]). The first line of defence, primarily composed of model developers and business units, is responsible for designing, developing, and using models, while actively managing associated risks. The second line of defence consists of independent risk management functions, such as internal validation and compliance. This function ensures oversight by establishing policies, validating models, and advising on model risk mitigation strategies, thereby maintaining risk control and regulatory compliance. The third line of defence, Internal Audit, performs independent assessments to evaluate the effectiveness and soundness of governance, risk management, and control processes related to model risk ([12]).

Given these distinctions, our goal is to prevent overlapping lines of defence while ensuring that Internal Audit adds value through independent assessments of governance, risk management, and control processes ([13]). This also includes exploring innovative areas, such as non-regulatory models, where Machine Learning techniques are increasingly adopted. In these areas, Internal Validation's role may be less extensive, as its focus remains primarily on regulatory models. Therefore, we propose defining control areas taking into consideration the assessments already conducted by the first two lines, particularly the Internal Validation function. A structured approach may consider:

- If Internal Validation has not assessed the Machine Learning model, Internal Audit could perform a deep-dive assessment on all the areas represented before.
- If Internal Validation has instead performed an assessment of the Machine Learning model, the activity of Internal Audit could be steered towards a challenging of the assessment performed by the Validation function on the most relevant areas (as also described in the following chapter) or could be limited towards making reliance on Internal Validation outcomes. In addition, it could also cover areas not assessed by the Internal Validation, such as governance topics.

It is noted that the activities of Internal Audit in this regard should not be considered a substitute for the proper performance of the duties of the Internal Validation function, which are in any case aimed at ensuring a second level control presidium on internal credit models to guarantee their soundness and robustness.

3.3 Main Building Blocks

In this paragraph we provide a detailed exploration of the potential assessments that Internal Audit could conduct, following the previously summarized audit framework. The focus is to suggest a possible approach to conduct audit tests in more innovative areas, with careful attention to fostering effective collaboration and avoiding overlap with Internal Validation activities, as previously outlined.

[11] For further details please see "*Best Practices for Effective Model Risk Management*", Satish *et al*, 2016.
[12] For further details please see "*Time to audit your AI algorithms*", Sandu *et al*, 2022.
[13] For further details please see the Institute of Internal Audit https://www.theiia.org/en/standards/what-are-the-standards/definition-of-internal-audit/.

We believe that, compared to standard models, the most innovative areas of analysis that could benefit from an introduction of standardized best practices are the following:

- Governance and Organization, focusing on aspects related to EU AI Act requirements.
- Model Design and Performance, focusing on particular dimensions of ML models, such as interpretability, fairness, performance, overfitting, complexity and causality.
- Model Implementation, focusing on transparency and human oversight aspects.
- Model Use, Monitoring and Review, focusing on aspects related to human oversight and the appropriate training of model users, as well as on an appropriate model monitoring and records-keeping.

While each aspect of these areas is analyzed in detail in the following sections, we here emphasize that machine learning (ML) algorithms are not inherently resource intensive. Nevertheless, their true potential is unlocked when applied to large volumes of heterogeneous data, including unstructured data (e.g. web data). Consequently, financial institutions seeking to enhance their credit processes with AI must invest in advanced analytics platforms capable of handling vast data volumes and executing complex computations efficiently. In this evolving landscape, Internal Audit, as an independent control function, must align with the AI infrastructure used for model development. Operating within the same IT environment as the model development team ensures consistency in software versions, computational resources, and data infrastructure. This alignment enables Internal Audit to accurately replicate model outcomes, conduct unbiased testing, and provide objective feedback without being hindered by IT discrepancies. To effectively oversee AI-driven credit processes, Internal Audit must also align its AI expertise through specialized training and hands-on experience. By strengthening its technical capabilities, auditors can better monitor AI models, adapt to emerging technologies, and maintain independence, transparency, and accountability. This, in turn, reinforces trust and validation in AI-based decision-making for credit risk assessment. Therefore, a structured roadmap for adopting and scaling the proposed audit tests must be aligned with the institution's current AI infrastructure and its planned evolution, considering both IT resources and skill development.

3.3.1 Governance and Organization

Regarding the governance aspects of credit risk models, the audit framework should assess the governance structure supporting the model's development, implementation, usage, and monitoring. This includes evaluating compliance with internal policies and organizational rules, as well as ensuring the clear and accurate formalization and effectiveness of roles and responsibilities and their segregation throughout the model lifecycle.

Focusing on the innovative aspects of Machine Learning models, a new item of assessment emerges, that is compliance with the EU AI Act requirements related to model governance. Indeed, as explained in section 2 regarding the general obligations for providers, the EU AI Act establishes some specific requirements in this area, namely:

- Ensuring an appropriate risk assessment and risk-based classification of the AI system, including identifying relevant requirements if it falls within the high-risk category.
- Providing an appropriate level of transparency and necessary information to deployers and users.
- Registering the system in the EU database for high-risk AI systems.
- Conducting a conformity assessment to verify compliance with general obligations for providers of models used in creditworthiness evaluations.

It is therefore important to ensure compliance with such requirements, where necessary establishing, as in our proposed framework, appropriate audit checks on the compliance with EU AI Act model governance requirements for ML models falling into the high-risk category.

It is important to note that the EU AI Act requirements for high-risk systems will take effect 24 months after the Regulation's publication (i.e., from August 2026). As such, compliance is not yet mandatory. However, during this transition period, we propose evaluating whether effective initiatives are in place to ensure the future compliance of credit models with these requirements once they become applicable.

3.3.2 Performance, Overfitting and Complexity

Performance

As already discussed in the introductory parts of this paper, one of the main drivers of the diffusion of ML techniques in several fields, including risk management and credit risk modelling, is linked to their capacity to boost the performance of predictive models, both in terms of increased accuracy and enhanced reactivity to credit deterioration signals compared to traditional models. Against this assumption, it would be therefore useful to back-test such hypothetical increasing predictive capacity with respect to alternative models. How to do so?

In this case, one possible check that we propose is the so-called "*time-to-default*" assessment. The idea underlying this assessment is to verify, for a sample of counterparties that entered into default status, what is the "time-to-default", meaning the time elapsed between the first signal of creditworthiness deterioration detected by the model and the actual default of the counterparty, comparing the Machine Learning model with an alternative benchmark model (possibly non-ML based). The objective of this test is to answer to the following question: is the Machine Learning technology producing a tangible increase in model performance? The expectation is therefore that the ML model is more responsive compared to the non-ML one, meaning that it would anticipate signs of deterioration in creditworthiness in a more gradual manner, and as such it is expected to observe a greater "time-to-default" for the ML model.

Before entering into details, looking at the following graph represented in Figure 1 helps to better explain the main idea underlying the proposed test.

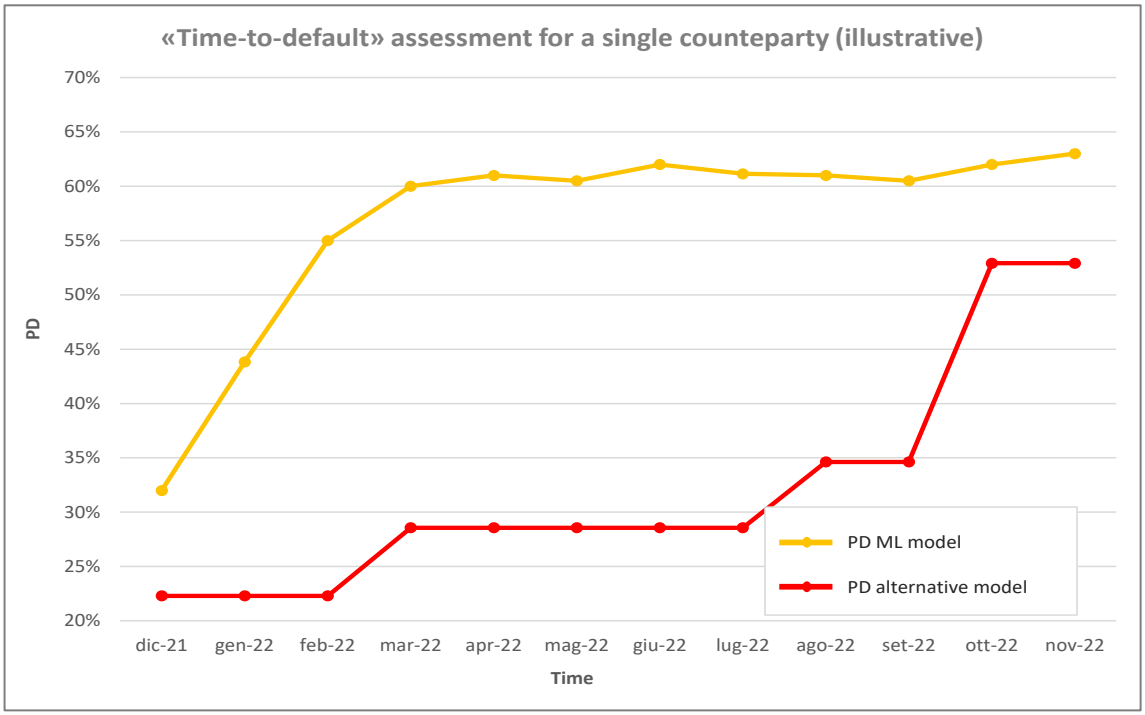

Figure 1 - Example of "time-to-default" assessment for a single counterparty (source: elaboration of the Authors)

The above chart shows, for a given counterparty which entered default status on 31/12/2022, the PD produced by the ML model (yellow line), and the PD produced by the alternative, non-ML model (red line).

Looking at the lines, one notices that the yellow line anticipates the deterioration signals of the counterparty, showing an increase of the PD since almost one year before the default (January 2022), while the red one exhibits a sudden increase only 2 months before the default (October 2022).

In such a case, one can conclude that the ML model is indeed more reactive in detecting a deterioration of the creditworthiness of the counterparties. It should be of course noticed that the proposed approach is intended to be a simple and intuitive method to test the ML model, and more complex and sophisticated approaches can be implemented, also depending on the available data and observations.

Here below we summarize the main steps to conduct such an assessment:

1. Identify the sample population of analysis, which should be a set of counterparties that have experienced a default, and which are scored both by the ML model under investigation and by an alternative scoring model.
2. For each of these counterparties, calculate the monthly (or with different frequency, according to the time granularity of data available) change in the score / PD calculated by each model m:

$$\Delta\, PD_t^m = \frac{PD_t^m - PD_{t-1}^m}{PD_{t-1}^m}$$

3. Identify the widest change in the ΔPD calculated above and assume that such maximum change corresponds to the first signal of deterioration identified by the model.
4. Count the months elapsed between the moment when the signal of deterioration emerged, and the moment of actual default of the counterparty. Such time elapsed is the “time-to-default".
5. Cluster the magnitude of PD / score changes, calculated as per point 2., into different risk buckets (e.g., deciles) to calculate the average “time-to-default” for each bucket. It should be noticed that this last step allows the production of an aggregate result taking into consideration the different risk profile of the counterparties analyzed; also in this case, more advanced approaches can be conceived to effectively produce an aggregate outcome for the test.

An example of application of the above testing steps and related results is reported in Figure 2 below, where the sample population has been clustered in 10 risk buckets by dividing it in deciles according to the value of monthly PD change.

The yellow line represents the average time-to-default per bucket based on the estimated PD of a Machine Learning model (in this specific case the ML technique employed is an *Extreme Gradient Boosting, XGBoost* for short, but the proposed test is “model-agnostic” and it can be applied independently of the modelling approach adopted), while the red line shows the average time-to-default per bucket stemming by the application of a traditional (logistic) PD model.

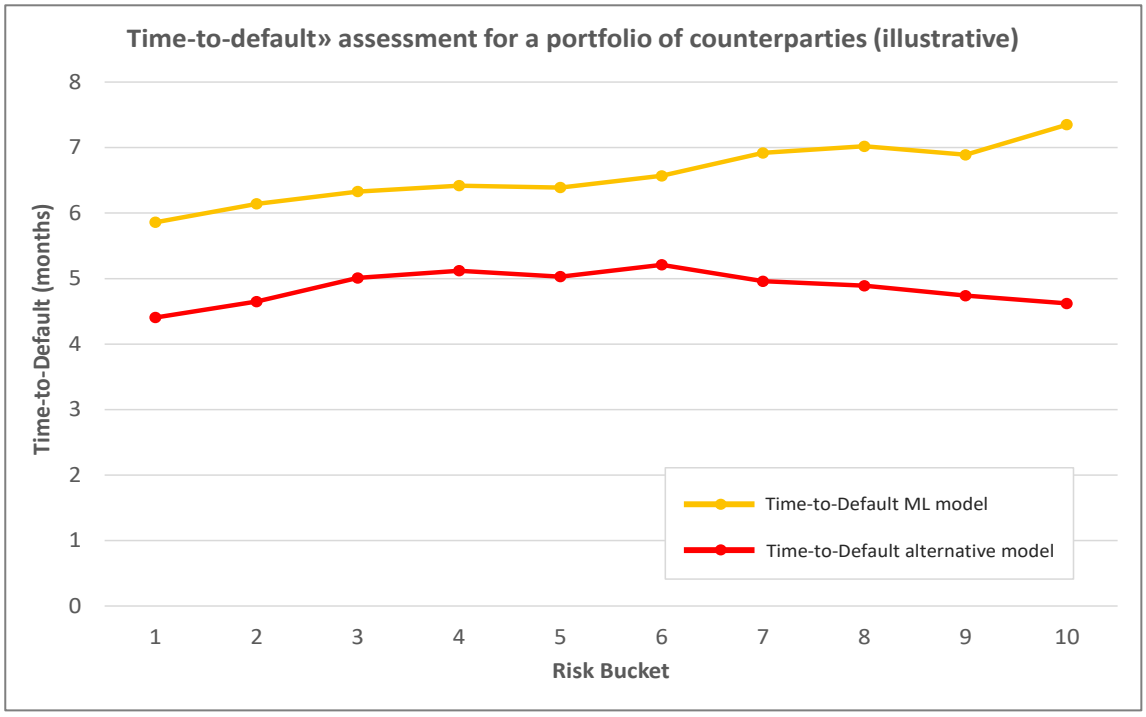

Figure 2 - Example of "time-to-default" assessment for a portfolio of counterparties (source: elaboration of the Authors)

As it can be noticed, the yellow line is consistently above the red one across all buckets, providing supporting evidence for the greater reactivity of the ML model in detecting signals of early distress for counterparties migrating to default. Additionally, it is also interesting noting that the stronger is the PD variation (identified ascendingly by risk buckets) the earlier the Machine Learning PD anticipates the result of the alternative model. This can be seen in the chart where the distance between yellow and red lines increases as the deterioration signal gets more evident in terms of PD.

It is important to remark that the aim of the above represented example is only to illustrate in a practical way how the proposed test could be employed to assess the predictive performance and reactivity of a Machine Learning model. The considerations and conclusions drawn above should not be interpreted as a general statement of stronger ability of ML models in predicting the default of a counterparty, compared to traditional models. As a matter of fact, the outcomes of the presented analysis might change from one case-study to another, depending on several elements, such as the modelling techniques used, the characteristics of the portfolio considered, etc.

Overfitting

Machine Learning models are prone to overfitting: this can happen when the model trains for too long and becomes too complex, for example a decision tree with an excessive number of splits, which adapts too closely to the training data, and then underperforms when applied to alternative datasets.

In contrast to the previous sub-paragraph, here we refer to performance as the accuracy of predictions provided by the model, measured for instance by the Accuracy Ratio (AR) or Area Under the curve (AUC) in case of binary target variable (e.g., for a PD model). This situation is known as overfitting and is surely an undesirable property for a predictive model.

Several techniques have been developed to mitigate the risk of overfitting. Indeed, in addition to the standard out-of-sample / out-of-time testing, widely used also in the development and validation of traditional models, more advanced techniques, such as *k-fold cross validation* are generally used when training ML models.

This involves splitting the dataset into (k) equally sized folds, and the model is trained (k) times, each time on (k-1) folds and tested on the remaining fold. The final performance of the model is then calculated as the average of the performances obtained from all (k) iterations. This method helps ensure that the model's performance is robust and not dependent on a particular train-test split, allowing a greater generalization power of the model when applied to different data.

A robust audit framework should include controls aimed at assessing the risk of overfitting, first by verifying whether an overfitting assessment has been performed during the development phase or by the Internal Validation function.

In case such checks have not been performed, or to challenge them also when they have been executed, we propose two possible controls:

- A "standard overfitting" test, based on the evaluation of the performance of the model when applied to an alternative testing set, which could be a classical out-of-sample or, better, an out-of-time sample, with more recent observations, allowing to make an assessment on the robustness of the model's performance over time. If the out-of-sample or out-of-time performance does not show a significant decrease with respect to the in-sample performance, then the model does not show overfitting issues.

We report below a practical example of such assessment: the depicted table shows the performance (Area Under the Curve, AUC) of a credit scoring model, trained for illustrative purposes ([14]), measured both on the training sample and on the test sample (i.e., a portion equal to 20% of the original sample kept out of the training phase for testing purposes).

Sample	AUC
In-sample	71.954%
Out-of-sample	69.295%
Delta	**-2.659%**
Delta %	**-3.696%**

Table 5 - Standard overfitting test: comparison between "in-sample" and "out-of-sample" performance (source: elaboration of the Authors)

As it can be noticed, the performance measured "in-sample" stands at around 71.95%, while "out-of-sample" it is equal to around 69.30%, suggesting a slight decrease, even though not extremely material, of the model's discriminatory power. As a best practice, thresholds for this test are usually set at -5%/-10%, meaning that relative drops of performance beyond these levels should be carefully assessed and the underlying drivers investigated. Generally, real-life credit risk models deployed in production might be supported by a "monitoring dashboard", which periodically provides performance metrics (such as AUC, stability indexes, etc.) updated to most recent data, allowing to continuously monitor and assess the proper functioning of the model over time and promptly detect potential deterioration of its performance. In this context, the first step of analysis for the Internal Audit function could be to analyze and assess the results of this periodic monitoring, where necessary complementing it with specific deep-dive assessments (e.g., analyzing the performance of specific model features, or investigating the performance of the model on specific sub-buckets of the population, etc.).

- A "robustness of performance" test, which could be carried out by assessing the performance of the model on a meaningful set of alternative samples. This test has the aim to investigate the behavior of the model's performance under a wide set of possible scenarios. Considering that in practice it might be difficult to identify and define multiple test samples, the idea is to leverage on the bootstrapping procedure to "build" a (potentially large) set of alternative sub-samples starting from the testing sample at hand (possibly also allowing for different distributional properties, i.e. through simple random sampling, without stratification).Then, test the model performance on each of these sub-samples, and verify if the performance of the model measured on the training sample falls within a reasonable boundary (e.g., 5-95 or 10-90 percentile) of the "bootstrapped performances". Such verification allows to get a sense on how reliable the prediction model is and how it is expected to perform over time, subject to distributional changes in the application dataset.
  As done for the previous test, we report below a practical example of such assessment performed on the same illustrative credit scoring model described before. The test sample kept out of the training phase consists of 200 observations (20% of the original sample): to conduct this assessment, we used simple random sampling (without replacement) to build 100 testing sub-samples (each of them with 100 observations) starting from the original testing set. Then the model is applied, and its performance evaluated (in terms of AUC) on each of the sub-samples, constructing a distribution of "bootstrapped performances". Finally, it is evaluated how the model's performance on the training sample behaves when compared to such distribution, to assess whether it falls within a reasonable range of this distribution or if it rather positions at its the extremes (i.e., beyond a certain percentile).
  Here below a histogram depicting the distribution of the AUC measured on the set of 100 alternative testing sub-samples:

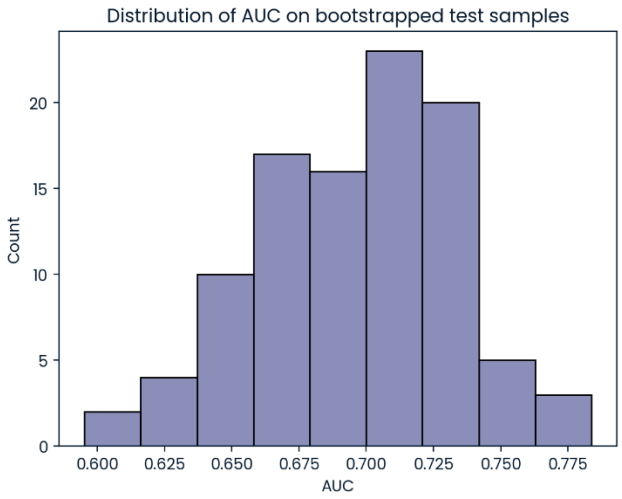

Figure 3 - Distribution of model's performance on bootstrapped testing samples (source: elaboration of the Authors)

[14] For this paper, we developed a credit risk model using the publicly available "German Credit Data" dataset prepared by Prof. Hofmann. This dataset contains 1,000 records of individuals who received credit from a bank, each categorized as a "good" or "bad" credit risk based on various features.
After an initial data exploration phase, we conducted a features engineering step. This involved handling missing values and applying label encoding to categorical features, such as "Housing", "Saving accounts" and "Checking account" to prepare the data for modelling.
Then, we selected LightGBM as the model due to its strong performance with structured data, as it can directly utilize categorical variables, allowing us to avoid One-Hot Encoding and instead rely on label encoding. To optimize model accuracy, we fine-tuned hyperparameters using *RandomizedSearchCV*. Model performance was evaluated primarily through accuracy to provide a reliable measure of predictive quality. For interpretability, we employed SHAP values and LIME analysis to better understand the model's predictions. For further details or the dataset please contact the authors.

As it can be noticed, the distribution is centered around 70% (the median is 70.02%), and the 5th and 95th percentiles equal to 63.54% and 75.29%, respectively. Recall also that the model's performance measured on the training sample is equal to 71.95%, thus quite close to the center of the distribution. Also consider that, out of 100 testing samples, only 12 of them exhibits an AUC below the relative threshold of -10% of the training AUC (64.76%), while 36 of them fall below the -5% relative threshold (68.36%). Based on these results, there is no strong evidence of the model overfitting when applied to a bunch of alternative testing samples. However, if the number of scenarios with performance under the defined thresholds resulted to be higher, this could be a warning signal of potential overfitting issues which should be investigated, for instance evaluating the model's performance on more recent data to assess whether they could be an early warning of performance deterioration over time.

Complexity

Another point of attention when dealing with Machine Learning models is complexity: it is without doubt that a ML model is more complex, in terms of training procedures, debugging, features maintenance, interpretation and communication of results, than a simpler, traditional, statistical or regression model.

Such increased complexity is justified, on the other hand, by greater predictive performance. But what if this is not the case? What if a level of performance similar to that of our ML model can be achieved by a simpler model, which leverages on a less complex modelling approach or less advanced and more interpretable variables?

To answer to these questions, we propose to introduce a test within the audit framework aimed at challenging the complexity-performance trade-off of the model under investigation, by verifying if it is possible to achieve similar performances by means of a simpler model. Also in this case, such checks could be possibly carried out also during the development stage of the model or by the Internal Validation function during its assessment of the model. On the other side, whenever such controls are deemed to be not exhaustive, Internal Audit could challenge them through a dedicated complexity assessment. To do so, we must first define what we mean by "complexity" when speaking of a model. Here below we summarize some possible dimensions of model complexity and propose specific checks to assess each of them:

Complexity dimension	Description	Audit check on this dimension
Complexity of the algorithm	An algorithm that is too complex ("black box") could lead to results that are not easily interpretable and produce additional costs in terms of development and retraining	If made possible by the structure of the training dataset and its features, train a simpler model, such as a simple decision tree or a logistic regression, and compare its performance with that of the audited model
Complexity of the model features	A model that relies on complex features (e.g., advanced variables based on complex mathematical transformations) could lead to results that are difficult to interpret and explain to relevant stakeholders	Train an alternative model based only on the simplest and most interpretable variables and compare its performance with that of the audited model
Complexity of the model in terms of number of features	A model that relies on a huge number of features could determine additional costs in terms of model maintenance (e.g., in terms of data collection and storage) and complicate the relationship between independent variables and the target, making it more difficult to interpret	Metrics such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) can be used to assess this dimension, as they penalize models with an increasing number of features. In addition, an alternative model can be trained with a lower number of variables, and its performance compared with that of the audited model

Table 6 - Complexity dimensions and related possible audit checks (source: elaboration of the Authors)

For the sake of awareness, it is important to highlight that model complexity, particularly in terms of the number of features, can be evaluated using metrics such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). These criteria are widely used in regression models, such as logistic regression, and probabilistic machine learning models, including Bayesian Networks and Hidden Markov Models (HMMs). They help balance model fit and complexity by penalizing complex model, ensuring a trade-off between overfitting and underfitting. However, in credit risk modelling, where machine learning techniques rely heavily on ensemble methods based on decision trees, such as XGBoost, LightGBM, and CatBoost, AIC and BIC are not directly applicable. These ensemble methods do not rely on probabilistic estimation, making traditional complexity measures unsuitable. Instead, alternative approaches must be used to manage the number of features and control model complexity. Among the possible approaches for limiting the number of features may use:

- Recursive Feature Elimination ("RFE") which iteratively removes the least important features and retrains the model to identify an optimal feature subset ([15]).
- "Early-Stopping" methods, which means adding up features during the training until their marginal contributions can be considered negligible in terms of discriminatory power according to some expert-based thresholds ([16]).
- Additional regularization techniques can be applied through hyperparameter tuning to effectively control the number of features used in tree-based ensemble methods. Specifically, it is possible to limit tree depth using parameters such as *max_depth* and *num_leaves*, which constrain tree growth and reduce feature reliance. Furthermore, feature complexity can be penalized through L1 regularization (*reg_alpha*), which encourages sparsity by eliminating less important features, and L2 regularization (*reg_lambda*), which shrinks feature weights to reduce overfitting while maintaining generalization ([17]).

Finally, interpretability measures such as *SHAP (SHapley Additive Explanations)* and *LIME (Local Interpretable Model-Agnostic Explanations)* can be leveraged to identify and eliminate weakly contributing features, thereby reducing model complexity without sacrificing performance. These interpretability techniques are described in detail in the next section 3.3.3.

If the above-described analyses highlight that less complex models can provide comparable levels of performance (i.e., within a sufficiently limited interval), then the complexity of the ML model is not justified, suggesting room for an improvement in performance or, alternatively, a simplification of the modelling techniques.

3.3.3 Interpretability

The main advantage of machine learning models is that they are extremely powerful, as they are able to uncover complex non-linear relationships between variables that are not identified by traditional, simpler, models. If on the one hand this feature increases the accuracy of ML-based predictive models, on the other hand it could complicate the interpretation of the identified relationships, making the model's results opaque and more difficult to explain. Different from traditional regression models, where regression coefficients allow a direct grasp on what is the relative impact of a single model feature on the overall prediction, machine learning models do not have explicit parameters and coefficients. To address this challenge, various interpretability techniques have been recently developed, which allow the contribution of each model variable on its predictions to be assessed and identified.

Generally, the greater the complexity of the ML technique adopted, the lower the transparency of results; for instance, the outcomes of a complex neural network are far less interpretable than those of a (rather simple) decision tree or a random forest ([18]). It is important to emphasize that there is a distinction between machine learning (ML) models and black boxes. Not all ML models are inherently black boxes, even though there is no strict boundary between simple and advanced models. Generally, there is a heuristic trade-off between interpretability and accuracy — more interpretable models tend to have lower accuracy, while highly accurate models are often less interpretable. Figure 4 below provides a graphical representation of such trade-off:

Figure 4 - The relationship between interpretability and accuracy (source: Hottenhuis (2022))

[15] Leveraging on the "German Credit Data", we implemented an example on Recursive Feature Elimination (RFE, using *RFE* method from python package *sklearn.feature_selection*) to evaluate its effectiveness in feature selection. Specifically, we initialized an RFE object with a LightGBM classifier as the base model to iteratively rank feature importance. At each step, the least important feature was removed until the number of remaining features matched *n_features_to_select*. The names of the most important features were then retrieved for further analysis. However, it is important to note that the dataset used contained a limited number of features, which may not provide significant improvements in model performance when applying RFE.

[16] Leveraging on the same "German Credit Data", we also combined Recursive Feature Elimination (RFE) with an early-stopping approach (using *early_stopping* method from python package *lightgbm.callback*) to optimize feature selection and model performance. Specifically, we defined a validation dataset *eval_set=[(x_valid, y_valid)]* to monitor model performance during training. The AUC score was used as the evaluation metric (*eval_metric="auc"*), and early stopping was configured with *callbacks=[early_stopping(50)]*. This setup ensures that training automatically stops if the AUC score fails to improve for 50 consecutive iterations.

[17] The previous hyperparameters (e.g., *max_depth*) are usually optimized before the final training phase using the *RandomizedSearchCV* function from python package *sklearn.model_selection*, which searches for the best combination through randomized sampling and cross-validation.

[18] It should be noted that a model which is excessively complex can lead, in addition to interpretability issues, also to overfitting problems, which are described in section 0.

The interpretability (or explainability, these two terms are usually used interchangeably) of ML models' results is then a key area of concern, especially in a context, that of credit risk, where the ability to identify the key drivers behind a certain credit decision is of utmost importance (e.g., explain why the loan application of a certain borrower has been refused). This is relevant also from a regulatory perspective; indeed, art. 15(1)(h) of the European General Data Protection Regulation (GDPR) ([19]) provides for the data subjects (and thus customers) "*the right to obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed, and, where that is the case, access to the personal data and the following information: [...] the existence of automated decision-making and [...] meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject*". According to Banca d'Italia (2022), this request for "meaningful information on the logic involved" implies an "*obligation for intermediaries to provide so-called "local" explanations to applicants, inclusive of the details of the main variables that contributed to a specific outcome regarding the loan approval or denial*". In this regard, in a recent case concerning fully automated credit scoring, the Advocate General of the European Court of Justice expressed its opinion regarding the level of disclosure and transparency that should be guaranteed to customers in case of automated decision making. According to this opinion, data subjects have the right to receive details about methods and criteria used, including information on the parameters and input variables used in the determination of the rating and their influence on the calculated rating ([20]).

To effectively understand and explain the relationships identified by a Machine Learning (ML) model, it is crucial to have suitable explainability techniques available.

This requirement is integral to any audit framework, which should include controls to assess whether model developers have applied appropriate explainability techniques. These techniques allow auditors and stakeholders to identify the variables that most significantly impact the model's results and to check this estimated contribution against economic and business expectations.

We believe that the evaluation of model interpretability during audit activities can proceed along the following steps:

- Verify business relevance: ensure that explainability analyses conducted during the model development phase (or during validation activities, if performed by the Internal Validation function) align with business logic and expectations.
- Assess the appropriateness of techniques: if the outcomes of the previous step align with expectations, verify that the explainability methods used are suitable and consistent with established practices in the relevant literature. Consider the previous discussion and any cited references to confirm methodological soundness.
- Monitor result stability: review whether the outcomes of explainability analyses have been consistently monitored and documented. Assess the stability of these results over time during both the development and validation phases.
- Challenger model testing: Since post-hoc techniques are being used, consider testing a challenger model as a benchmark to further validate the reliability of the explanations provided.

The explainability of Machine Learning models – commonly known as *XAI* (*eXplainable Artificial Intelligence*), *XML* (*eXplainable Machine Learning*), and *IML* (*Interpretable Machine Learning*) – usually refers to model-agnostic post-hoc explanations (Molnar, 2022). Two widely used explainability techniques, particularly relevant in credit risk modelling, are SHAP and LIME. These techniques are both local, model-agnostic explainability methods. "Local" implies that they provide insights into individual predictions for specific observations, while "model-agnostic" indicates they can explain predictions from any ML model type (e.g., random forest, XGBoost, neural networks). Notably, SHAP can also be applied to global explainability, offering insights into how specific model features influence predictions across a testing set.

SHAP

SHAP (*SHapley Additive exPlanations*) is a method to explain individual predictions, based on Shapley values. The theory underlying Shapley values comes for coalitional game theory, which tells us how to fairly distribute the payout of a game (the prediction of the model in our case) among the different players (the features). In the context of model interpretability, Shapley values highlight how much each model's feature has contributed to the single (local) prediction compared to the average prediction. The Shapley value of a specific feature of the model is the average marginal contribution of the feature value across all possible coalitions. In pseudo-code it works as follows ([21]):

1. Create the set of all possible feature combinations (called coalitions).
2. Calculate the average model prediction.
3. For each coalition, calculate the difference between the model's prediction without the feature *i* and the average prediction.
4. For each coalition, calculate the difference between the model's prediction with *i* and the average prediction.
5. For each coalition, calculate how much *i* changed the model's prediction from the average (i.e., step 4 – step 3). This represents the marginal contribution of feature *i*.
6. Shapley value is equal to the average of all the values calculated in step 5 (i.e., the average of *i*'s marginal contributions).

[19] *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data [...] (General Data Protection Regulation).*
[20] See Opinion of Advocate General Richard De La Tour, delivered on 12 September 2024, Case C-203/22.
[21] For further details on the pseudo code, please see the web resources Shapley Values - A Gentle Introduction | H2O.ai.

It should be noted that the exact calculation of the Shapley values for a feature *i* involves the evaluation of all possible coalitions of feature values, with and without the *i-th* feature to calculate the exact Shapley value. For more than a few features, the exact solution to this problem might become problematic as the number of possible coalitions exponentially increases as more features are added. For this reason, in the most common packages used to calculate the Shapley values they are actually computed via approximation (such as Monte-Carlo sampling).

More about the theoretical fundamentals of SHAP and Shapley values can be found in the Annex, as well as in Lundberg and Lee (2017) and Molnar (2022).

Here below we propose a SHAP interpretability analysis, applied to an illustrative credit scoring model trained on the publicly available dataset "German Credit Data" ([22]).

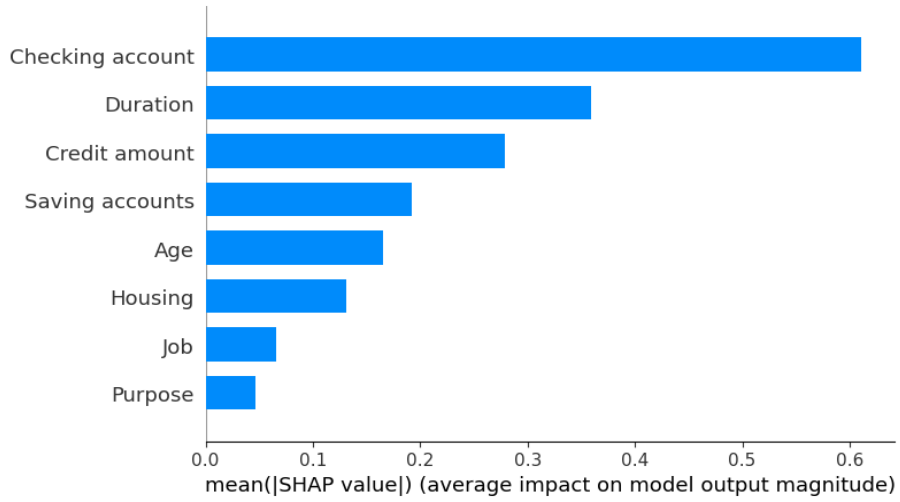

Figure 5 - SHAP waterfall plot for the illustrative credit scoring model (source: elaboration of the Authors)

One of the features of SHAP is that it allows the assessment of the "global feature importance" of a model, i.e., to understand which are the most important features that are driving the model's prediction in a given testing set. The above chart presented in Figure 5 represents the variable impact sorted for importance within our illustrative model. As can be noticed, the variable importance has a good level of business sense, as the most important variables appear to be "Checking account", "Duration" and "Credit Amount".

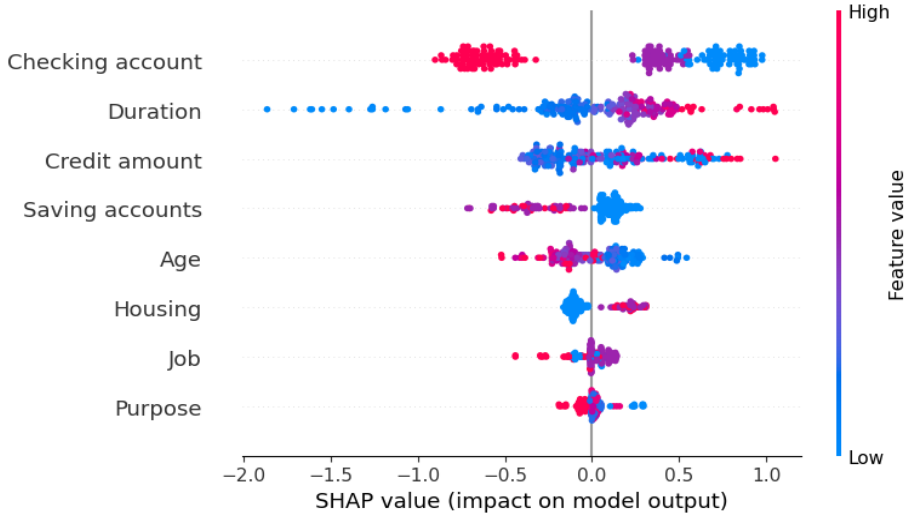

Figure 6 - SHAP beeswarm plot for the illustrative credit scoring model (source: elaboration of the Authors)

In addition, SHAP allows an assessment of the estimated impact of each variable on the model prediction is coherent with economic business sense. The above chart presented in Figure 6 visualizes the impact of each feature on the model's prediction of the probability of default in our illustrative model. In detail, the x-axis shows the SHAP values, which represent the influence each feature has on pushing the prediction toward a higher or lower probability of default. Values to the left indicate a decrease in the probability of default, while values to the right indicate an increase. As one can see:

- The "Checking account" feature appears to have a strong negative impact. This suggests that individuals with stable checking accounts are perceived as lower risk.
- The "Duration" seems to have a moderate positive impact. An increase of the loan term might be associated with higher default risk.

[22] The credit scoring model considered is the same already described in the "Overfitting" section.

- The "Credit amount" is a significant driver of the model's prediction. This aligns with intuition: larger loan amounts often correlate with higher risk.

It is important to highlight that SHAP analysis could reveal some non-business sense effects in the variables, for example in the variable “Credit amount” there is a mixing in the risk sense between “low” and “high” values. Situations like the previous one should raise attention in the phase of feature creation, suggesting for instance the need for a better treatment of missing values.

LIME

Another technique that can be used for auditing ML model’s interpretability is *LIME*, which stands for *Local Interpretable Model-agnostic Explanations*. The main idea behind LIME is to train a local interpretable model (for instance, a linear model) in proximity of on observation of interest (local region), which acts as a simpler and explainable model (surrogate) to identify the features that drive the prediction of the model for this observation, i.e., to get a local explanation. The explanation provided by the simpler model does not represent a global explanation of the complex model predictions, but it represents a good explanation of the local prediction, as it is trained on the local region in proximity of the observation of interest. How is the surrogate model trained? Data points around the observation of interest are perturbed and weighted according to their proximity to it, then the original model's prediction on the perturbed data points is used to train a simpler model that approximates the model accurately in the vicinity of such observation. Then, the local explanation is obtained by interpreting the local model (e.g., in case of a linear regression model used as surrogate, by looking at the fitted weights of the regression).

Below we provide the main steps for training a local surrogate model ([23]):

1. Select the observation of interest for which you want to have an explanation of its black box prediction.
2. Perturb the dataset and get the black box predictions for these new points.
3. Weight the new samples according to their proximity to the observation of interest.
4. Train a weighted, interpretable model on the dataset with the variations.
5. Explain the prediction by interpreting the local model.

For further details about the theoretical foundations of the LIME approach, please refer to Ribeiro *et al* (2016), Molnar (2022) and Juscafresa (2022).

To raise awareness, we conduct a LIME (Local Interpretable Model-agnostic Explanations) analysis based on the previous illustrative credit scoring model ([24]). The purpose of this analysis is to provide insights into the model's prediction by identifying the features that contribute the most positively or negatively to the specific outcome. By highlighting these influential features, we aim to enhance our understanding of how the model arrives at its decision for a given input data of the test sample.

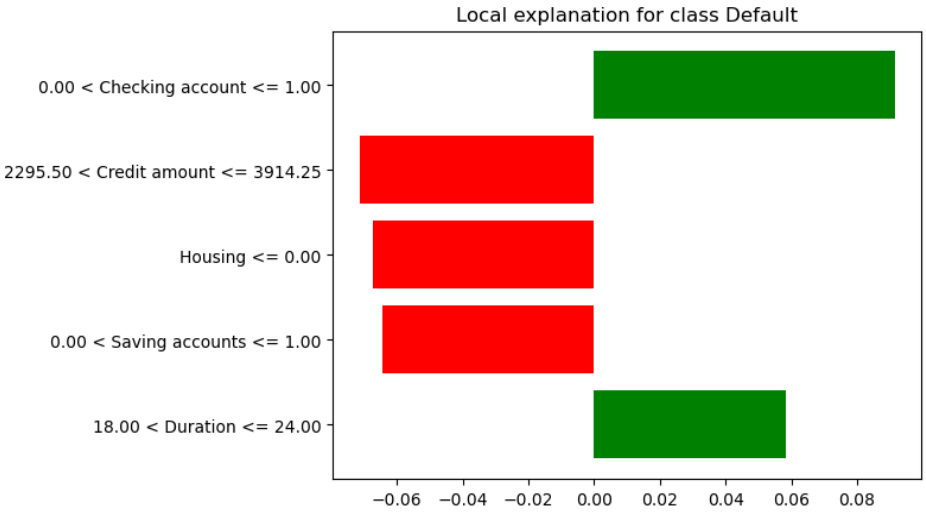

Figure 7 - Example of local explanation with LIME (source: elaboration of the Authors)

In Figure 7 above, red bars represent features that push the “local” prediction (i.e., the prediction for a single observation of the sample) toward “No-Default”, while green bars represent features that push the prediction toward "Default." The key features influencing the prediction are as follows:

- “Checking Account” (green): This is the most influential feature (largest green bar). A low checking account value strongly increases the likelihood of the individual being classified as “Default”.
- “Credit Amount” (red): This feature has a significant negative impact, meaning individuals with credit amounts in this range are less likely to be classified as Default.

[23] For further details please see 9.2 Local Surrogate (LIME) | Interpretable Machine Learning.
[24] For further details on LIME analysis please contact the authors.

- "Duration" (green): This feature has a positive effect, with longer loan durations increasing the likelihood of the individual being classified as "Default."

The results of the LIME approach, in terms of most important variables and the sign of their impact, are substantially aligned with that resulting from the previous SHAP analysis, confirming the robust business sense of these model variables.

In this regard, empirical evidence showed that LIME can serve as a powerful tool for assessing the explanatory capabilities of an audited model within specific data clusters. For example, consider a scenario in which a subset of customers characterized by outlier data has been identified. In this case, it may be of interest to investigate whether LIME's explanations align with those provided by SHAP. This comparison is particularly relevant to assess whether the model's predictors maintain an "explainable relationship" with the target variable within specific subpopulations. Consequently, a meaningful approach could involve comparing global SHAP values with LIME explanations calculated only for the target subpopulation. However, being LIME explanations calculated for each single instance, to make the comparison consistent, results from these two XAI techniques must be ranked ([25]).

To operationalize this approach, one could extract, for each instance within the selected cluster, the top five variables with the highest explanatory power according to LIME. By counting the frequency with which these variables appear in the top five rankings, a ranking of variables based on their explanatory consistency across instances can be generated. If the frequency-based ranking derived from LIME aligns with the ranking of average SHAP values calculated over the whole population, this would indicate that the model's explanatory capability remains stable and reliable within the specific cluster under analysis.

Feature	SHAP value (average)	LIME top 5 explainer
Checking account	0.37	85%
Credit amount	0.34	75%
Housing	0.32	69%
Saving accounts	0.22	62%
Duration	0.2	55%

Table 7 - Comparison between top 5 average SHAP values and top 5 frequency-based LIME explanations (source: elaboration of the Authors)

It is important to highlight that implementing the previously discussed metrics requires careful consideration of the computational burden, especially when analyzing large datasets. To mitigate this challenge, both SHAP and LIME can be applied to subsamples, reducing computational costs without compromising the robustness of the insights. For instance, SHAP values can be computed only for the top percentage of predictors based on their discriminatory power, as well as on specific clusters/subsamples of the development sample, without compromising the robustness of test information. Additionally, when applying these tools, it is crucial to recognize their potential limitations and shortcomings, which may impact their reliability or even lead to misleading conclusions. Some of these challenges include:

- For SHAP:
    - The exact computation of SHAP is complex, therefore their calculation for many instances or for global interpretability purposes might result time and computationally intensive.
    - SHAP ignores feature dependence, as it replaces feature values with random instances, and therefore might lead to misleading results in case of highly correlated variables.
- For LIME:
    - LIME calculation is computationally burdensome for large datasets, as it requires generating multiple perturbed samples and making multiple predictions.
    - LIME explanations are instable, thus hampering their reliability. As shown by Alvarez-Melis and Jaakkola (2018), very small perturbations in the feature values, that have minimal or no effect on the underlying model's predictions, might produce significant effects on the explanations given by the surrogate model.
    - LIME explanations heavily depend on the setting of LIME hyperparameters, i.e. the definition of the "local region" (neighborhood), the surrogate model used, how the dataset is perturbed (sampling), etc.

For additional details on possible shortcoming of SHAP and LIME, see Molnar (2022) as well as the web resources referred to in the footnote ([26]). In conclusion, speaking more generally of post-hoc XAI techniques it is important to consider the following points:

- As acknowledged by EBA (2021), XAI typically offers only a partial understanding of a model. A potential solution is to place greater emphasis on inherently interpretable models, also referred to in the literature as white box models, where "explainability/interpretability" is designed in. At present, it appears that the financial industry is not considering white box models for credit risk, instead focusing on tree ensemble models such as Random Forest and XGBoost.
- There is no established procedure to compare LIME and SHAP analyses. The proposed approach is an innovative methodology introduced within a real audit engagement, marking the first instance in which interpretability issues in AI-based models were addressed alongside computational challenges. A comparison between these two XAI approaches at the

[25] A similar approach based on rankings was proposed in Shi *et al* (2023). Our approach proposed in this paper does not aim to make LIME explanations global; rather, it seeks a common-sense method for comparing the alignment of results from two different XAI techniques.
[26] *From Opaque to Transparent: Understanding Machine Learning Models with LIME | Medium* and *What's Wrong with LIME | Medium*.

local level was presented by Giudici and Gramegna (2021) using both an unsupervised approach (K-means clustering) and a supervised approach, with the final result indicating that SHAP performs better.

3.3.4 Fairness

A widely discussed topic in the realm of AI-based decision-making applications with impact on individual persons is related to ethical risks, i.e., the risk of discrimination against individuals or groups of individuals based on sensitive personal attributes (such as age, disability, gender, sexual or political orientation etc.). The property of non-discrimination of outputs produced by ML models is generally referred to as *fairness*.

One of the main causes of unfair outputs produced by Machine Learning models is bias, whose concept may be distinguished as:

- Statistical or representation bias: this form of distortion arises when the data is not representative of the true population it refers to. It can be caused by forms of selection bias, where the observations within the training data are not a random selection of the true population. One example is that of data related to loan repayment capacity, which is observed and available only for individuals who have actually been granted credit in the past, and therefore not representative of individuals who have applied for credit but have been denied it.
- Historical or societal bias: even if the data is not affected by statistical bias, there may be a form of distortion in the data reflecting distorted behaviors or decisions occurred in the past. This can happen when there is a bias in label assignment, for instance when past decisions are systematically favorable/unfavorable towards certain groups of individuals, and such labels are used to train a ML model.

Fairness is a widely and deeply discussed topic in the current literature concerning the ethical implications of AI-based models, including credit scoring models.

The first step of analysis when dealing with ML-fairness is defining the characteristics for which a model can be deemed fair. As reported in Hurlin *et al* (2024), the main definitions of fairness can be summarized as below:

- *Demographic parity* (also defined as *independence* or *statistical parity*): a model is said to satisfy this property if the predicted outcome $\hat{Y}$ is independent of the sensitive attribute(s) D, i.e., if $\hat{Y} \perp D$.
  In the context of credit scoring, and considering gender as an example of sensitive protected attribute (e.g., $D = 1$ for males, $D = 0$ for females), this property implies that all applicants should have an equivalent opportunity of obtaining a good outcome (loan approval ([27]), $\hat{Y} = 1$), regardless of their gender, i.e., $Pr(\hat{Y} = 1|D = 1) = Pr(\hat{Y} = 1|D = 0) = Pr(\hat{Y} = 1)$.
  Looking at the above definition, and its practical implications, it is quite straightforward to highlight its flaws, as it imposes an equality of treatment for all individuals, thus not allowing to take into account their creditworthiness, which a good credit scoring model should instead reflect.
- *Conditional demographic parity*: a model is said to satisfy this property if the predicted outcome $\hat{Y}$ is independent of the sensitive attribute(s) D, conditional to the value of non-sensitive attributes (X_C), i.e., if $\hat{Y} \perp D \,|X_C$. The conditioning attributes, X_C, should be non-sensitive variables that have a direct impact on the creditworthiness of individuals, such as their income. Practically, in the credit scoring-gender example, this definition implies that males and females with a similar level of income ($X_C = x$), should have an equivalent opportunity of obtaining a loan approval ($\hat{Y} = 1$), while different outcomes between males and females should be justified by different level of incomes: $Pr(\hat{Y} = 1 \mid D = 1, X_C = x) = Pr(\hat{Y} = 1 \mid D = 0, X_C = x) = Pr(\hat{Y} = 1 \mid X_C = x)$.
- *Separation* (also defined as *equality of odds*): a model is said to satisfy this property if the predicted outcome $\hat{Y}$ is independent of the sensitive attribute(s) D, conditional to the value of the actual outcome Y, i.e., if $\hat{Y} \perp D \mid Y$. This definition implies that all applicants with similar credit standing ($Y = y$) should have the same opportunity of obtaining a loan approval ($\hat{Y} = 1$), regardless of their sensitive attribute characterisation; in other terms, it requires the model to have equal True Positive Rates (TPR) and False Positive Rates (FPR) across sensitive groups: $Pr(\hat{Y} = 1 \mid D = 1, Y = y) = Pr(\hat{Y} = 1 \mid D = 0, Y = y) = Pr(\hat{Y} = 1 \mid Y = y)$, with $y \in \{0, 1\}$.
- *Sufficiency*: a model is said to satisfy this property if the actual outcome Y is independent of the sensitive attribute(s) D, conditional to the value of the prediction $\hat{Y}$ made by the model. In practice, this definition implies that all applicants receiving the same prediction $\hat{Y}$ by the scoring model, should have the same probability of having a certain credit type realisation (good/bad). From a rating model standpoint, this definition can be seen as requiring that individuals within the same rating class (i.e., with predicted probability of default within a specific bound) have a similar level of observed riskiness (e.g., in terms of default rate): $Pr(Y = 1 \mid D = 1, R = C_k) = Pr(Y = 1 \mid D = 0, R = C_k) = Pr(Y = 1 \mid R = C_k)$, where C_k defines the specific value of the attributed rating class R.

For a more extensive and deeper overview of fairness definitions and their practical implications, we refer to Hurlin *et al* (2024), Barocas *et al* (2023), Adebayo (2012), Castelnovo *et al* (2021), Castelnovo *et al* (2022).

The integration of fairness into existing model development workflows requires a structured approach and involves a bias assessment and mitigation across the entire machine learning modelling pipeline, i.e. from data pre-processing to the evaluation of the model's results. As a matter of fact, as also detailed in Castelnovo *et al* (2021), the bias mitigation and fairness assessment can be embedded in each of the following stages of the model development workflow:

[27] Here the focus is on the "equality of treatment", for this reason the condition is imposed on the outcome of the loan approval process, rather than on the actual default of the applicant.

- Pre-processing: methods that can be used in this stage of the modelling pipeline are based on the idea of removing potential unfair biases directly from the training dataset. Among these methods, the simplest one is to remove possible sensitive attributes from the dataset prior to the model training (so-called *Fairness Through Unawareness*).
- In-processing: methods used in this phase consist mainly in enforcing a model to produce fair outcomes by adding constraints or penalties to its optimization problem ("fairness-aware cost functions"), thus imposing fairness at training time.
- Post-processing: these strategies are focused on mitigating potential unfair outcomes of a Machine Learning model which has already been trained. This can be achieved by defining a new classifier as a function of the (potentially biased) outcomes of the unmitigated model, optimizing some cost function over false positives and false negatives subject to some fairness constraint.

It is important to highlight that the above-described stages of fairness assessment play a key role in the activities of model development, while from an audit perspective, which by its nature is based on a downstream assessment of the model, the evaluation of fairness aspects is more strictly oriented towards the final outputs produced by the model.

In this context, we believe that the most suitable definitions of fairness, to be assessed as part of a sound audit framework, are those of separation or sufficiency, better if paired with a verification on the *blindness* of the model, i.e., ensuring that no sensitive attributes are used as explanatory variables within the model, a situation which would lead to a direct risk of discrimination.

We summarize below the key steps to conduct an assessment on the model fairness from an audit perspective:

- Verify whether fairness was taken into account during the development phase, and if fairness assessments were conducted during this phase (or during validation activities, if performed by the Internal Validation function).
- Assess the appropriateness of the definitions and techniques applied, by verifying that they are aligned with business best practices and suited for the model and the use case under analysis (e.g., statistical parity is not well suited for credit risk models). In addition, verify if the list of sensitive attributes identified is in line with applicable regulatory requirements and industry best practices.
- If possible or necessary, challenge the fairness assessment, for instance analyzing additional fairness metrics, or assessing the non-discrimination properties of the model on specific sub-buckets of the population. If not done during the development or the validation phase of the model, two useful fairness tests could be the following:
    1. Blindness: verify that the model does not contain explanatory variables based on sensitive attributes or related direct transformations.
    2. Sufficiency (group fairness): verify that the individuals for which the model gives a similar prediction, are actually similar in terms of observed riskiness. To do so, we propose the application of a two-sample Z-test for proportions, a statistical test used to determine whether two proportions are different from each other, with the null hypothesis that the two proportions are equal. Such a test is generally used for testing the homogeneity of individuals belonging to different buckets within the same rating class, where the proportion of each bucket is represented by its default rate. The formula for the two-sample Z-test for proportions is the following:

$$Z = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

    where $\widehat{p_1}$ is the proportion (the observed default rate) in the first sub-bucket and n_1 the related number of observations, $\widehat{p_2}$ is the proportion (the observed default rate) in the second sub-bucket and n_2 the related number of observations, and $\hat{p}$ is the average proportion of the two sub-buckets (the observed default rate of the union of the two sub-buckets). In practice, for each bucket of individuals which are similar in terms of model prediction (e.g., within the same rating class or within a limited bound of estimated PD values), we check that the actual riskiness of such individuals is similar despite their characterization of the sensitive attributes. To give a more practical example, suppose we have a model with N rating classes: for each class, split the borrowers between groups based on the sensitive attribute (e.g., males vs. females) and verify that their riskiness, represented by the observed default rate, is not statistically different, using the above defined Z-test. If the null hypothesis of "equality of default rates" within a certain bucket is not rejected (i.e., the p-value associated to the Z-statistic is above the chosen significance level), we can conclude that the model is "fair" for that bucket, as it treats similar individuals in a similar way, despite their sensitive characterization.

Below we provide an example of practical application of the fairness assessment described above, based on a simple credit scoring model trained for illustrative purposes ([28]). For the considered model, the information about the borrower's gender (male vs. female) is available. We did not use this information as an explanatory variable for training the model, hence the property of blindness is satisfied. However, we want to check also for sufficiency, that is we want to answer to the following question: is the model treating similarly males and females in our sample, in relation to their observed riskiness?

To conduct such test, we clustered our testing sample in 2 buckets (classes): observations with a predicted PD above a certain cut-off fall within a "bad" class, while below the PD cut-off they are assigned to a "good" class. We want to verify if males and females inside each bucket are treated similarly; the Z-test allows us to assess if borrowers within each cluster are similar also in terms of

[28] The illustrative credit risk model considered here is the same already described in the previous paragraphs regarding the overfitting and interpretability assessments.

observed riskiness. We then split each of the two buckets in two sub-buckets, based on gender, distinguishing males from females, and via the Z-test we verified if these two sub-buckets are homogeneous in term of default rate. The results of the assessment are summarized in Table 8 below. As it can be noticed, for both buckets the p-value is above the level of confidence (set in our example to 5%), therefore the null hypothesis of "equality of default rates" ([29]) between sub-buckets of each bucket cannot be rejected. Based on these results, we can conclude that our model is "fair", as in addition to the blindness property it also satisfies the definition of sufficiency.

Class	Gender	Count all [bucket]	Count all [sub-bucket]	Avg. PD [bucket]	Avg. PD [sub-bucket]	Default Rate [bucket]	Default Rate [sub-bucket]	Z-score	p-value	Null hypothesis (DR_M=DR_F) ([30])
Good	**Male**	592	426	13.139%	13.086%	7.095%	7.277%	0.277	0.782	**NOT REJECTED**
	Female		166		13.274%		6.627%			
Bad	**Male**	408	264	55.393%	55.115%	63.235%	60.606%	-1.491	0.136	**NOT REJECTED**
	Female		144		55.903%		68.056%			
TOTAL		**1,000**	**1,000**	**30.379%**	**30.379%**	**30.000%**	**30.000%**			

Table 8 - Results of the fairness assessment on the illustrative credit scoring model (source: elaboration of the Authors)

In addition to what discussed above, when dealing with the fairness of AI-based credit risk models, a relevant point to be considered is the trade-off between the model's performance and the fairness of results that such model produces. As a matter of fact, it is reasonable to expect that more complex models ([31]), which can ensure the achievement of higher levels of accuracy and performance, might be more prone to exploit potential biases embedded in the training data, thus exacerbating the risk of producing discriminatory results.

While this trade-off should be carefully evaluated during the modelling phase, we propose here a possible check that could be performed from the audit standpoint to assess the balance between the two dimensions of performance and fairness. The idea of the proposed test is to verify whether, given a specific model under assessment, it is possible to obtain an alternative challenger model which ensures similar levels of accuracy but at the same time being "fairer" than the one under evaluation. This can be done by first identifying possible sensitive features included in the current model, and then challenge it through an alternative model which does not rely on such features.

To give a more practical example, we leverage on the same illustrative credit scoring model already considered above. Recall that this model includes "Age" as an explanatory variable: even though such information has been so far widely considered and included in credit risk models and rating systems, one could argue that its inclusion in the model specification might lead to potentially discriminatory results against certain age groups of the population. In this context, our proposed test involves verifying if an alternative model, equal to the audited one but with the exclusion of the "Age" variable, can achieve an accuracy which is sufficiently similar (i.e. not too much lower) to that of the model under investigation.

Model specification	AUC
Including "Age"	69.295%
Excluding "Age"	68.015%
Delta	**-1.280%**

Table 9 - AUC (test sample) of the credit scoring model with and without the "Age" variable (source: elaboration of the Authors)

As it can be noticed from the table above, the exclusion of the sensitive variable "Age" leads only to a minor and surely negligible drop in the model performance (measured by its AUROC, which slightly decreases from 69% to 68%). In this context, the audit test would suggest refusing the original model specification and prefer the alternative one, as it reduces the risk of possible bias and, eventually, discrimination, stemming from the reference to sensitive characteristics of the individuals such as their age.

3.3.5 Model Implementation

Model implementation is not a prerogative area of Machine Learning models, as Internal Audit checks on these aspects are expected to be conducted also for models based on traditional statistical techniques.

When it comes to Machine Learning models, it should be noted that the implementation of such models ([32]) is usually done via a dedicated Data Analytics platform, that allows the management of the comprehensive end-to-end process of model development, testing, deployment and implementation within a unique platform. By leveraging specific and advanced data analytics platforms, it is possible to deploy a model in a production environment without the need to refactor the entire model development code for the new environment. Instead, a simple "dump" of the model can be performed, ensuring perfect alignment in the model's functioning operation across all environments (development, testing, and production). In light of this, the incisiveness of the audit controls required on these aspects of model implementation is limited.

[29] Similar to other validation tests that compare estimated PD and default rates, it is important to note that the estimated PD and corresponding rating classes are based on data from time t-1, while the performance window for counterparties in a specific rating class spans from t-1 to t. This means that the test can only assess whether the model was meeting the fairness condition 12 months ago.

[30] When the number of observations within a class is low, it is advisable to also assess the power of the test. This ensures that, in case of failing to reject the null hypothesis, the difference you are investigating does not actually exist (Type II Error).

[31] Complexity is intended here both as in terms of the algorithm employed as well as in terms of the explanatory variables considered for the model training.

[32] This is primarily due to the complexity of ML/AI algorithms, such as ensemble tree methods, which rely on multiple decision trees and are therefore challenging to deploy manually."

Nevertheless, it should be taken into account that, as highlighted in section 2, specific requirements for the providers of high-risk systems are foreseen by EU AI Act, covering also the area of model implementation. Indeed, in addition to the accuracy, robustness and cybersecurity requirements set out by art. 15, the Regulation also provides the requirement of record keeping of logs issued by the model over its lifetime (art. 12), to ensure the traceability of its functioning, as well as the obligation to report any serious incidents regarding the high-risk system to market surveillance authorities (art. 73).

From an audit perspective, it is thus useful to introduce some audit checks on the implementation phase of the model, assessing the level of technical robustness of the implemented model and the framework for record keeping and incidents reporting set out.

3.3.6 Model Use, Monitoring and Review

As depicted in Table 4, this area of our audit framework encompasses three phases, namely model use, model monitoring and model review.

Each of these three phases is relevant from an audit perspective, especially when dealing with ML models, which introduce elements of specificity within each of these stages.

Regarding model use, it is important that people involved in the final use of the model's outcomes within credit processes have appropriate knowledge of how the model works, how it makes predictions and what these predictions are driven by. This is relevant both from a regulatory compliance standpoint (in particular, art. 14 of the EU AI Act requires a sufficient level of human oversight during the model functioning to ensure that the model is functioning as intended), but also from an internal one, as it ensures that the model's results are used consciously and appropriately. In addition, it allows the detection of potential model malfunctions and interventions with appropriate corrective actions in case of anomalous results.

It is therefore important, from an audit perspective, to check that all these elements are in place to allow for a correct and responsible use of the model and an adequate level of human oversight. Practical audit checks on this aspect could include the following controls:

- Is there a complete and transparent documentation in place that clearly describes the model functioning and its main drivers?
- Does the documentation describe the limitations of the model, and/or situations where its outcomes should be considered and used with a lower level of reliability?
- Is there a well-defined framework of escalation / remediation in cases where the application of model's outcomes in the credit process has highlighted potential anomalies?

Model monitoring aims at ensuring that the model functions smoothly and correctly over time. To do so, it is important that there is an adequate framework for the periodic monitoring and maintenance of the model, for instance with a periodic analysis and assessment of its predictive performance and established thresholds to trigger model maintenance interventions. More practically, the audit framework in this area could include the following checks:

- Assess whether a structured monitoring framework has been defined to verify the adequate functioning of the model in production over time. The monitoring framework should include at least tests in the areas of model performance, stability, interpretability and fairness, with related thresholds to highlight potential anomalies and analyze their trend over time.
- Assess whether the monitoring framework in place allows for a timely identification of malfunctioning and anomalies (i.e., if the controls are executed over a short time frame after the periodic model runs), and if results of the monitoring checks are interpreted correctly and communicated clearly to all relevant stakeholders.

An additional element to be considered in the model monitoring area is the presence of an appropriate log record-keeping system, that allows to track and detect potential anomalies in the functioning of the system and ensure compliance with EU AI Act requirements.

Lastly, concerning the model review, it is important that there is an adequate framework in place providing clear instructions in terms of periodic model review and the conditions for a model update, as well as an ex-ante definition of the situations and conditions that might lead to a decommissioning of the model. Examples of practical checks that could be performed in this area are the following:

- Verify if the development function has defined, also in conjunction with the framework for model monitoring, a set of rules, metrics and thresholds to identify situations in which the review or the decommission of the model should be evaluated, and if such rules, metrics and thresholds are sound and adequate.
- Verify if situations like those defined in the framework have occurred in the past and if the defined rules, metrics and thresholds have been applied consistently.

4. Conclusions

In recent years, the increasing availability of IT resources capable of managing and analyzing massive amounts of data, along with the potential consequences of Basel IV's regulatory model rollback, has driven banks and financial institutions to develop and adopt a new class of cutting-edge models based on machine learning technology. These models aim to enhance competitiveness and efficiency in creditworthiness assessment. Furthermore, recent reports from the European Banking Authority (EBA) and the announcement that the upcoming version of the ECB Model Guide will include specific supervisory expectations on the use of machine learning highlight the growing regulatory focus on credit risk models incorporating AI techniques. These developments also foresee a crucial role for audit functions in providing assurance on these models by ensuring their reliability, regulatory compliance, and alignment with best practices in risk governance.

In this light, this paper offers a structured audit approach for assessing and testing AI-based credit risk models, systematically identifying model risks throughout the model life-cycle phases defined in Model Risk Management (MRM). This approach aligns with both current and emerging regulations while integrating established methodological references. The proposed framework tries to provide a comprehensive method for identifying risks associated with Machine Learning models, addressing both risks directly linked to the credit risk model—such as obligor misclassification—and broader risks related to the model's use, governance, monitoring, and maintenance. Additionally, the framework incorporates considerations for ethical risks.

The proposed tests and their practical implementation present a potential approach for auditing AI-based credit risk models. While not exhaustive or definitive, this approach aims to reflect the inherent complexity and multidimensional nature of the subject, emphasizing the need for flexibility and adaptability, recognizing the evolving challenges in this area.

Moreover, the proposed framework is designed to accommodate potential future extensions to recognize emerging aspects; as a matter of fact, Artificial Intelligence and Machine Learning science is continuously evolving, and this evolution can provide new instruments to address challenges of AI-based credit risk models. An example of possible future extension of our framework is related to causality, a prominent emerging topic in Machine Learning literature: causality, also known as Causal Inference or Causal AI, consists of a set of techniques which reveal the causal relationships that are embedded in the data, making it possible to back-test predictive ML models and analyze the robustness of causal relationship identified. Further detail on causality is presented in the Annex.

A notable aspect of this framework is its consideration of future regulatory requirements, particularly those related to the forthcoming EU AI Act. By integrating potential future obligations, the framework positions itself as forward-looking, helping organizations stay compliant with evolving standards and best practices in AI governance. This proactive approach is crucial in an era where regulations around AI and Machine Learning are becoming increasingly stringent and impactful. However, in our proposed practical tests, we recommend evaluating a "transition period" prior to the full implementation of the EU AI Act.

Finally, it is important to emphasize that this work is intended as a starting point for further refinement and discussion. The dynamic nature of both the regulatory landscape and advancements in AI technology means that ongoing development and adjustment will be necessary. The aim here is to foster a broader dialogue on how to effectively manage model risk while balancing regulatory compliance, ethical considerations, and technological innovation. Future research and industry input will be vital to expanding and refining the approaches discussed in this paper.

5. Annex

Theoretical fundamentals on SHAP

SHAP and Shapley values allow to identify what is the contribution of a specific feature i for a specific prediction of our model. To calculate the Shapley we first create all possible coalitions of the other features of the model, and re-calculate the model prediction for every coalition, with and without the feature of interest, and calculate the difference between this prediction and the average prediction. By comparing the change with respect to the average prediction, with and without the feature of interest, we get the marginal contribution. The value of variables that are not in the coalition is replaced with a random value extracted from the available dataset. This marginal contribution is then averaged across all possible coalitions to get the average marginal contribution.

In formulas:

$$\phi_i\,(N, v) = \frac{1}{|N|!} \sum_{S \subseteq \setminus N\{i\}} |S|!\,(|N| - |S| - 1)!\;[v(S \cup i) - v(S)]$$

where:

- the symbol $|.|$ is used to indicate the cardinality (i.e., the numerosity) of a set of features (it is therefore not to be intended as the "absolute value");
- N is the set of all features;
- $\frac{1}{|N|!}\sum_{S \subseteq \setminus N\{i\}}$ identifies the average of the marginal contributions of feature i over all possible coalitions;
- $S \subseteq \setminus N\{i\}$ is the subset of features that does not include feature i;
- $|S|!\,(|N| - |S| - 1)!$ identifies the weight, that is the product of the number of permutations of S and the number of permutations of the complement of S, where S is a subset of features used in the model;
- $[v(S \cup i) - v(S)]$ is the marginal contribution of feature i to the subset S.

It should be noted that the exact calculation of the Shapley values for a feature i involves the evaluation of all possible coalitions of feature values, with and without the *i-th* feature to calculate the exact Shapley value. For more than a few features, the exact solution to this problem might become problematic as the number of possible coalitions exponentially increases as more features are added. For this reason, in the most common packages used to calculate the Shapley values they are actually computed via approximation (such as Monte-Carlo sampling).

More about the theoretical fundamentals of SHAP and Shapley values can be found in Lundberg and Lee (2017) and Molnar (2022).

Theoretical fundamentals on LIME

Here we briefly introduce with the mathematical details of LIME which, as we will see, also depend on the interpretable model which is chosen as surrogate.

$$\hat{g} = \underset{g \in G}{arg\,min}\, \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

where:

- f: the complex model being explained.
- g: the interpretable model (e.g., a linear model) that approximates f locally around *x*.
- G: the set of all possible interpretable models.
- $\mathcal{L}(\mathrm{f}, \mathrm{g}, \pi_x)$: a loss function that measures the difference between f and g in the neighborhood of x, weighted by π_x.
- π_x: a locality measure, which assigns higher weights to data points closer to x.
- $\Omega(g)$: a regularization term that enforces simplicity in g to ensure interpretability.

The goal of LIME is to minimize with an interpretable model $\hat{g}$ the loss function $\mathcal{L}$ around x, with a penalty $\Omega(g)$ that keeps g simple and interpretable. For further details about the theoretical foundations of the LIME approach, please refer to Ribeiro *et al* (2016), Molnar (2022) and Juscafresa (2022).

Causality

Machine Learning models leverage their power to recognize and exploit complex correlation patterns among variables and use such correlations to make predictions.

However, as it is commonly said, correlation does not imply causation. Indeed, two variables that appear related based on their observational data, might instead be the outcome of a confounding variable, that is a (usually unobserved) variable that influences both observed features and leads to an apparent causal relationship that is actually spurious. Prediction models that leverage spurious correlation relationships might provide poor performances and inaccurate predictions when such relationships shift or break down.

To mitigate this problem, a prominent area of analysis is currently being developed, known as Causal Inference or Causal AI. Causal AI does not merely look at correlations, but it rather allows study of the causal relationships that are embedded in the data. Leveraging these tools, it is possible to back-test predictive ML models analyzing the robustness of causal relationship identified.

An example of Causal AI is the usage of causal graphs, that represent graphically the assignment structure of a model, and allow the identification of direct and indirect causal relationships in a model. With causal graphs, one can identify and visually represent *confounding* variables, as well as *mediator* variables. An example of causal graphs for this type of analysis is reported in figures 8 and 9.

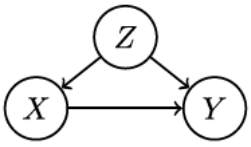

Figure 8 - Example of causal graph with a confounding variable (source: Barocas et al (2018))

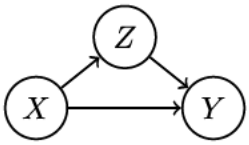

Figure 9 - Example of causal graph with a mediator variable (source: Barocas et al (2018))

In Figure 8, we can see that variable Z acts as a confounder for the relationship between X and Y, as it produces a positive correlation of X and Y which is a mere result of confounding, while X and Y are not linked by a causal relationship. A model that draws conclusions on the $X \rightarrow Y$ relationship without considering the effect of the variable Z would most likely lead to inaccurate predictions. A similar effect, but due to a different situation, is that of a mediator variable, represented in Figure 9: here we can notice that the relationship between X and Y is two-fold: a direct relationship $X \rightarrow Y$, and an indirect relationship $X \rightarrow Z \rightarrow Y$. Also in this case, a model that attempts to model the relationship between X and Y and make related predictions, should appropriately capture the effect of the mediator variable Z. For additional details on Causal AI theory, and on causal graphs, please refer to Barocas *et al* (2018).

References

- Adebayo, Julius. "*FairML: Toolbox for diagnosis bias in predictive modeling*." Master's thesis, Massachusetts Institute of Technology, 2016. https://dspace.mit.edu/handle/1721.1/108212
- Alvarez-Melis, David, and Tommi S. Jaakkola, "*On the Robustness of Interpretability Methods*", 2018. https://doi.org/10.48550/arXiv.1806.08049
- Aniceto, Cardoso Maisa, Flavio Barboza, and Herbert Kimura. "*Machine learning predictivity applied to consumer creditworthiness*". Future Business Journal, Springer, vol. 6(1), pages 1-14, December. https://doi.org/10.1186/s43093-020-00041-w
- Babaei, Goolnosh, Paolo Giudici, and Emanuela Raffinetti. "*A Rank graduation box for SAFE AI*", Expert systems with applications, Vol. 259, 125239, January 2025. http://dx.doi.org/10.1016/j.eswa.2024.125239
- Barocas, Solon, Moritz Hardt, and Arvind Narayanan. "*Fairness and Machine Learning. Limitations and opportunities*", 2018. https://api.semanticscholar.org/CorpusID:113402716
- Board of Governors of the Federal Reserve System. "*Supervisory Guidance on Model Risk Management*. SR Letter 11-7. 2011. https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm
- Bonaccorsi di Patti, Emilia, Filippo Calabresi, Biagio De Varti, Fabrizio Federico, Massimiliano Affinito, Marco Antolini, Francesco Lorizzo, Sabina Marchetti, Ilaria Masiani, Mirko Moscatelli, Francesco Privitera, and Giovanni Rinna. "*Intelligenza artificiale nel credit scoring. analisi di alcune esperienze nel sistema finanziario italiano*". Banca d'Italia, Questioni di Economia e Finanza, 721, 2022. https://www.bancaditalia.it/pubblicazioni/qef/2022-0721
- Castelnovo Alessandro, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Claudio Andrea Cosentini. "*A clarification of the nuances in the fairness metrics landscape*", Sci Rep 12, 4209, 2022. https://doi.org/10.1038/s41598-022-07939-1
- Castelnovo, Alessandro, Riccardo Crupi, Giulia Del Gamba, Greta Greco, Aisha Naseer, Daniele Regoli, and San Miguel Gonzalez Beatriz. "*Befair: Addressing fairness in the banking sector*", IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 2020, pp. 3652-3661. https://doi.org/10.48550/arXiv.2102.02137
- Clark, Andrew. "*The Machine Learning Audit - CRISP-DM Framework*". ISACA Journal, 1, Volume 1, 2018. https://www.isaca.org/resources/isaca-journal/issues/2018/volume-1
- European Banking Authority. "*Discussion Paper on Machine Learning for IRB Models (EBA/DP/2021/04)*", 2021. https://www.eba.europa.eu/discussion-paper-machine-learning-irb-models
- European Banking Authority. "*Follow-up Report on the use of Machine Learning for IRB models (EBA/REP/2023/28)*", 2023. https://www.eba.europa.eu/publications-and-media/press-releases/eba-publishes-follow-report-use-machine-learning-internal
- European Central Bank. "*ECB Guide to Internal Models*", 2024. https://www.bankingsupervision.europa.eu/ecb/pub/pdf/ssm.supervisory_guides202402_internalmodels.en.pdf
- European Court of Justice. "*Opinion of Advocate General Richard De La Tour, delivered on 12 September 2024, on Case C-203/22*". https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:62022CC0203
- European Parliament and European Council. "*Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*". http://data.europa.eu/eli/reg/2016/679/oj
- European Parliament and European Council. "*Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) 300/2008, (EU) 167/2013, (EU) 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)*". http://data.europa.eu/eli/reg/2024/1689/oj
- European Parliament and European Council. "*Regulation (EU) 575/2013 of 26 June 2013 on prudential requirements for credit institutions and investment firms and amending Regulation (EU) 648/2012 (Capital Requirements Regulation)*". http://data.europa.eu/eli/reg/2013/575/oj
- Giudici, Paolo, and Emanuela Raffinetti. "*SAFE Artificial Intelligence in Finance*". Finance Research letters, Vol. 56, 104088, September 2023. https://doi.org/10.1016/j.frl.2023.104088
- Giudici, Paolo, Mattia Centurelli and Stefano Turchetta. "*Artificial Intelligence risk measurement*". Expert systems with applications, Vol. 235, 121220, January 2024. https://doi.org/10.1016/j.eswa.2023.121220
- Giudici, Paolo. "*Safe machine learning*". Statistics, 58(3), 473–477, 2024. https://doi.org/10.1080/02331888.2024.2361481
- Giudici, Paolo. "*Artificial Intelligence, risk management and the financial sector: the Safe model to assess the risks of AI*". Bancaria, November 2024, No. 11.
- Gramegna, Alex, and Paolo Giudici. "*SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk*". Frontiers in Artificial Intelligence, 4, 2021. https://doi.org/10.3389/frai.2021.752558
- Hottenhuis, Wouter. "*Inherently interpretable Machine Learning for probability of default estimation in IRB models*". Master's thesis, University of Twente, 2022. http://essay.utwente.nl/91965/
- Hurlin, Cristophe, Cristophe Perignon, and Sebastien Saurin. "*The Fairness of Credit Scoring Models*", 2024. https://doi.org/10.48550/arXiv.2205.10200
- Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker, Barnes. 2020. "*Closing the AI Accountability Gap: Defining an End-to-End*". In Conference on Fairness, Accountability, and Transparency, January 27–30, 2020.

- Juscafresa, Aleix Nieto. "*An introduction to explainable artificial intelligence with LIME and SHAP*". Master's thesis, Universitat de Barcelona, 2022. http://hdl.handle.net/2445/192075
- Lundberg, Scott M., and Su-In Lee. "*A Unified Approach to Interpreting Model Predictions*". Advances in Neural Information Processing Systems, 2017. https://doi.org/10.48550/arXiv.1705.07874
- Molnar, Christoph. *Interpretable Machine Learning*. 2nd edition, 2022. https://christophm.github.io/interpretable-ml-book
- Moscatelli, Mirko, Fabio Parlapiano, Simone Narizzano, and Gianluca Viggiano. "*Corporate default forecasting with Machine Learning*". Banca d'Italia, Temi di discussione, 1256, 2019. https://doi.org/10.1016/j.eswa.2020.113567
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "*Why Should I Trust You?: Explaining the Predictions of Any Classifier*". Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. https://doi.org/10.48550/arXiv.1602.04938
- Sandu, Iuliana, Menno Wiersma, and Daphne Manichand. "*Time to audit your AI algorithms*". Maandblad voor Accountancy en Bedrijfseconomie, 96 (7/8): (253-265), 2022. https://doi.org/10.5117/mab.96.90108
- Satish, Garla, and Dhillon Sukhbir. "*Best Practices for Effective Model Risk Management*", 2016. SAS Institute Inc., Cary, NC. support.sas.com/resources/papers/proceedings16/SAS6485-2016.pdf
- Shi, Li, Redoan Rahman, Esther Melamed, Jacek Gwizdka, Justin F. Rousseau, and Ying Ding. "*Using Explainable AI to Cross-Validate Socio-economic Disparities Among Covid-19 Patient Mortality*". AMIA Jt Summits Transl Sci Proc. 2023 Jun 16. https://doi.org/10.48550/arXiv.2302.08605
- The Institute of Internal Auditors (IIA). "*Auditing Model Risk Management*", 2018. www.theiia.org
- The Institute of Internal Auditors (IIA). "*The Definition of Internal Auditing*", 2022. www.theiia.org