# SAFE AI platform: A Platform for Continuous Evaluation of Accuracy, Fairness, Explainability and Robustness in Machine Learning Systems

Golnoosh Babaei[1,2,*], Marco Brandirali[2] and Paolo Giudici[1]

[1]University of Pavia, Via S. Felice Al Monastero, 5, 27100 Pavia, Italy
[2]Cashberry, Via Aosta, 20122 Milan, Italy

**Abstract**

The deployment of machine learning models in real-world environments requires continuous monitoring beyond predictive performance. In high-impact domains such as financial forecasting, regulatory compliance demands transparency, fairness, and stability over time. In this paper, we introduce SAFE-AI, a modular evaluation platform designed to systematically assess model *Accuracy*, *Fairness*, *Explainability*, and *Robustness*. The platform provides both metric-level diagnostics and aggregated risk indicators through a unified SAFE Score. Such a holistic evaluation approach is particularly valuable in financial applications, where model decisions can directly impact merchants, customers, and operational strategies. Consequently, the proposed platform supports more transparent, stable, and responsible use of machine learning models in transaction-driven environments. The proposed platform is validated through a real-world transaction forecasting use case using a LightGBM model.

**Keywords**

Artificial Intelligence, AI regulation, Model compliance, Financial Technology

## 1. Introduction

Machine Learning (ML) models are increasingly used in high-stakes decision-making contexts. In addition to the opportunities these models provide, they impose many risks. Therefore, we need to evaluate the models and their predictions to ensure reliability of the decisions. Regarding the model evaluation, While predictive accuracy remains a fundamental requirement, it is insufficient for guaranteeing responsible deployment [1]. The EU AI Act is the first European regulatory framework dedicated to regulating Artificial Intelligence (AI) systems [2]. Its objective is to ensure the safe, transparent, and compliant use of AI by introducing specific obligations based on the level of risk posed by AI systems. Therefore, based on the modern AI governance frameworks, ML models should be evaluated not only based on predictive accuracy but also based on fairness, explainability, and robustness [3].

With respect to the EU AI Act, AI applications are classified into different risk levels: Prohibited AI practices such as practices that distort human behavior, exploit vulnerabilities, generalize remote biometric identification for law enforcement. High risk AI practices which are allowed but require a risk management system. They include product safety components (e.g. in health and automotive) and personal services such as credit lending. Limited risk AI practices which are allowed, possibly under transparency requirements. The EU AI Act requires companies to be able to demonstrate the compliance of AI systems, particularly with regard to model reliability, transparency, fairness and robustness. Therefore, companies need scalable and integrated tools to manage AI compliance, reduce regulatory risks, and sustainably support innovation. For systems subject to regulatory obligations such as AI applications in finance, it is necessary to: 1. Measure and document model performance (e.g., accuracy and robustness) 2. Evaluate the transparency and explainability of algorithmic decisions 3. Monitor

*Late-breaking work, Demos and Doctoral Consortium, colocated with the 4th World Conference on eXplainable Artificial Intelligence: July 01–03, 2026, Fortaleza, Brazil*

*Corresponding author.
golnoosh.babaei@unipv.it (G. Babaei); marco@cashberry.it (M. Brandirali); paolo.giudici@unipv.it (P. Giudici)
0000-0001-6132-2271 (G. Babaei); 0009-0004-8787-5252 (P. Giudici)

fairness and potential biases [4]. These regulations will have a direct impact on companies that develop, use, or distribute AI systems, particularly in regulated sectors such as financial services. The challenge is not only to meet the requirements, but to be able to demonstrate them in a continuous and structured manner [5]. However, organizations often lack an integrated tool capable of monitoring these dimensions simultaneously [6]. To address this gap, we developed SAFE-AI platform, a platform that provides continuous evaluation across four complementary dimensions: Accuracy, Fairness, Explainability, and Robustness. To show the effectiveness of our proposed platform, we apply it to a Point-of-Sale (POS) transactional forecasting problem. POS transactional data analysis has become increasingly important for financial institutions, payment providers, and merchants, as it enables monitoring of economic activity, demand forecasting, risk assessment, and operational planning [7]. POS data typically exhibits high volume, strong temporal dynamics, and significant heterogeneity across merchants, sectors, and geographic regions. Machine learning models are therefore widely adopted to extract predictive insights from such data. However, despite their predictive capabilities, these models may introduce new risks related to reliability, stability, and potential biases across different merchant groups. In high-stakes environments such as banking and financial services, understanding and monitoring these risks is essential for ensuring responsible deployment of machine learning systems [8].

The rest of the paper is as follows; next section presents state of the art. Section 3 represents the structure and design of the proposed platform, followed by different explicit sections for the modules available on the SAFE AI platform.

## 2. State of the Art

Recent research in machine learning evaluation has increasingly emphasized the need to go beyond predictive accuracy and consider additional dimensions such as fairness, explainability, and robustness. Traditional evaluation approaches primarily rely on error-based metrics such as Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE), which provide a direct measure of predictive performance. However, these metrics alone are insufficient to assess the overall impact and trustworthiness of models, particularly in real-world decision-making contexts [9]. In parallel, fairness has gained significant attention, with various metrics proposed to quantify disparities across groups or temporal segments [10]. These approaches aim to identify biases and ensure equitable model behavior, especially in sensitive applications. Explainability has also become a central topic, with model-agnostic methods such as SHAP (SHapley Additive exPlanations) widely adopted to interpret model predictions by estimating feature contributions [11]. Robustness evaluation focuses on assessing the stability of model performance under varying conditions, such as temporal shifts or noisy inputs. Metrics based on variability (e.g., standard deviation or coefficient of variation) are commonly used to quantify performance consistency over time.

Despite these advances, most existing approaches address these dimensions in isolation. Only a limited number of works [12] attempt to integrate multiple evaluation criteria into a unified framework, and even fewer focus on practical, end-to-end platforms tailored to forecasting tasks. In this context, the SAFE-AI platform aims to bridge this gap by providing a unified evaluation framework that combines accuracy, fairness, robustness, and explainability into a single, practitioner-oriented assessment pipeline.

## 3. Platform Architecture

SAFE-AI platform is structured into independent evaluation modules that operate on model predictions across multiple temporal periods. Each module produces: 1. Per-period diagnostic metrics 2. Aggregated indicators 3. A normalized score in the range $[0, 1]$. The platform intentionally relies on widely adopted and interpretable metrics (e.g., MAE, MAPE, SHAP) to ensure transparency, reproducibility, and ease of adoption in real-world industrial settings. In the following sections, we consider a specific regression use case to show the functionality of this platform. However, the proposed SAFE-AI platform is applicable to different types of systems such as classification applications.

## 4. Use Case: POS Transaction Forecasting

We applied SAFE-AI to a LightGBM model predicting merchant POS transaction volumes. The dataset used in this experiment is provided by a FinTech company called Cashberry [1]. This dataset includes transactional information of various merchants that own POS terminals from different Italian banks.

Figure 1 represents the user interface of SAFE-AI platform. Since this platform is designed for Cashberry, the language of the platform represented in this paper is Italian. User can select one of the four available AI concerns to evaluate model in that environment. With the aim of assisting users to understand the results, we set an option to see the explanations provided by a selected Large Language Model, in this study Llama 3 [13]. The SAFE score section, calculates the unique indicator of reliability of the model. Finally, the button called "PDF report" produces a PDF file that includes all the results, which can be downloaded easily.

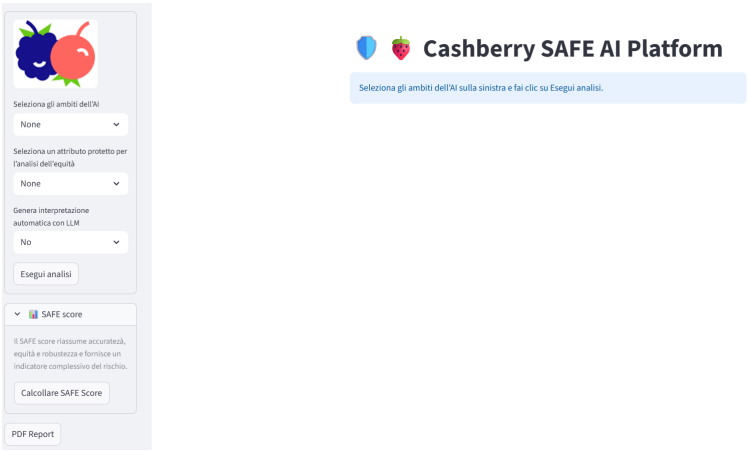


**Figure 1:** SAFE-AI user interface.

In the next sections, we will see the results of each module one by one.

## 5. Accuracy Module

The Accuracy module evaluates predictive performance using the following two standard metrics:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \quad MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{1}$$

These metrics are computed across multiple time periods to monitor performance drift. Moreover, a unique indicator, Accuracy Score, which is a normalized score is derived as:

$$AccuracyScore = \frac{MAE\_mean - MAE\_min}{MAE\_max - MAE\_min} \tag{2}$$

Taking into account this accuracy score, we can monitor the performance of the model to see how the model was able to improve the average performance.

Using the POS transaction data and the lgb model, the results of the accuracy module are represented in Figure 2:

As can be seen, to help users understand the results and discover the importance of accuracy evaluation, we present some preliminary definitions and we introduce the metrics that we calculate.

As an extra assistant for users, they have this option to see the comments from LLM-based intelligent agents. For example, in Figure 3, user selected to use AI agents. Therefore, in addition to the calculated results, he/she received the illustrated comprehensible comments originally by Llama 3.

[1] https://cashberry.it/

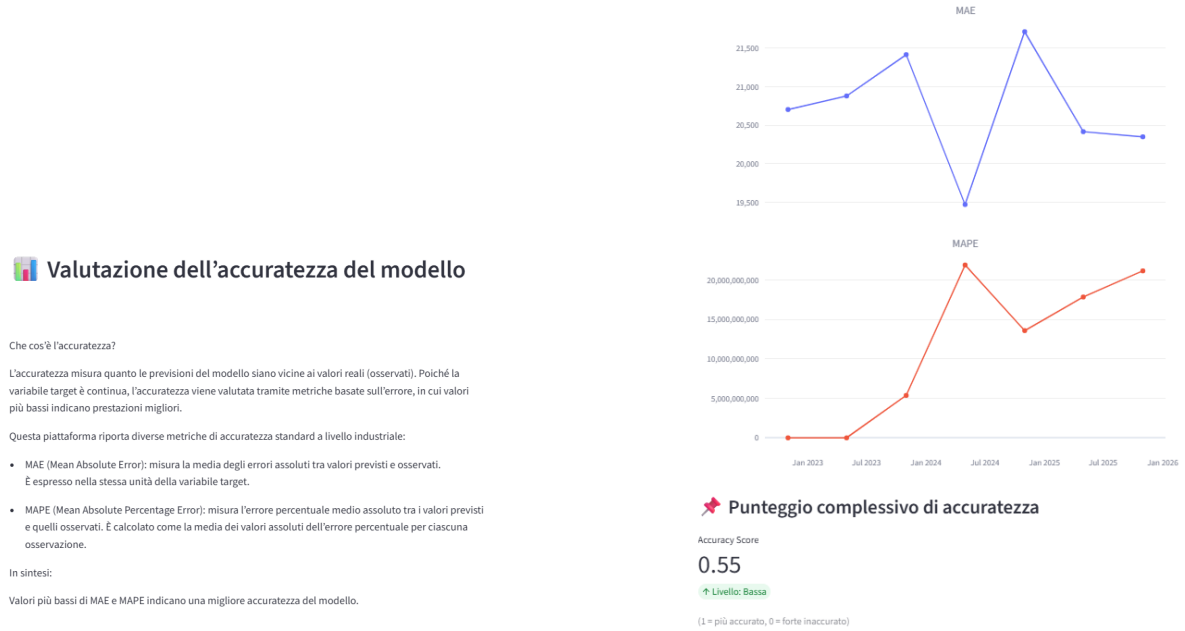


**Figure 2:** Accuracy module results.

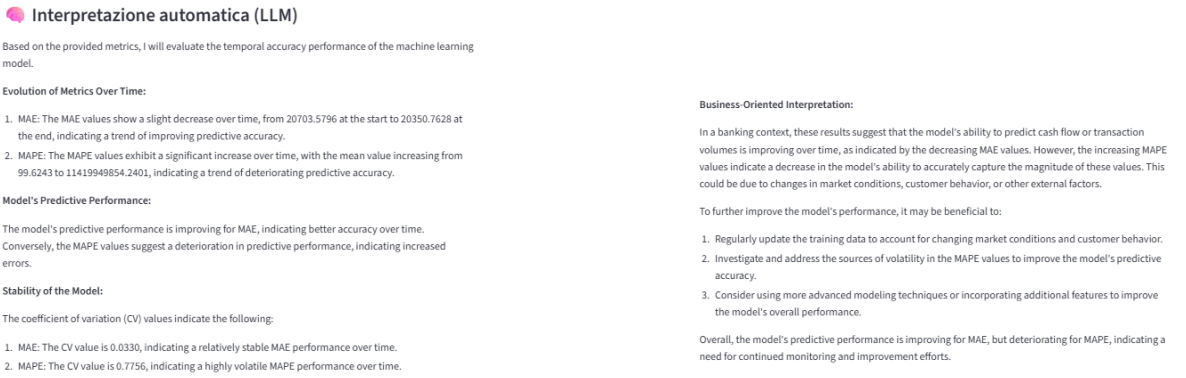


**Figure 3:** Interpretation of the accuracy results generated by LLM.

## 6. Explainability Module

To ensure understandability of the decisions given by the model, SAFE-AI platform integrates SHAP (SHapley Additive exPlanations) [14]. For this purpose, in each period the importance level of variable $j$ is as follows:

$$Importance_j = \mathbb{E}[|SHAP_j|] \tag{3}$$

In this module, we evaluate the mean importance of the variables over time to track how their importance changes in time. Regarding stability of the importance values, Coefficient of variation (CV) is utilized. Moreover, we look for the frequency of appearance in top-K features. This allows identification of: 1. Stable predictors 2. Emerging influential variables 3. Structural model shifts

Figure 4 represents the temporal importance of the most significant variables in the POS transaction dataset:

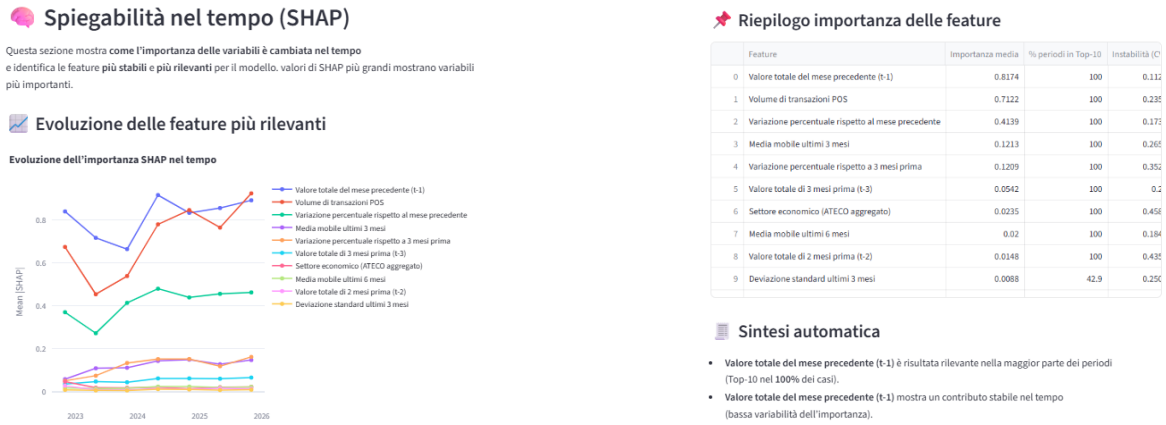


**Figure 4:** Explainability module results.

## 7. Fairness Module

The Fairness module evaluates group-level equity by analyzing disparities in predictive error across groups defined by a selected protected attribute [15]. For each temporal split, the Mean Absolute Error (MAE) is computed separately for every group with sufficient sample size. Within each period, the minimum and maximum MAE values across groups are identified, and the *MAE gap* is defined as:

$$\text{MAE gap} = \text{MAE}_{\max} - \text{MAE}_{\min}. \tag{4}$$

This quantity measures the performance disparity between the best-performing and worst-performing groups. Smaller MAE gaps indicate more equitable model behavior, as predictive performance is more uniform across protected groups.

To capture fairness dynamics over time, the MAE gap is computed for each temporal period and summarized into a normalized *Fairness Score* in the interval $[0, 1]$. The score is obtained by normalizing the variability of the MAE gap across periods, such that higher values correspond to smaller and more stable disparities. In addition, the module reports the MAE aggregated per group (weighted by sample size) to provide an overall ranking of group-level performance. This design enables both temporal monitoring of fairness and structural analysis of persistent performance differences among protected groups. Figure 5 represents the results of the platform:

## 8. Robustness Module

The Robustness module evaluates the stability of the predictive model across different time periods. Given multiple temporal splits of test data and corresponding predictions, the module computes standard regression error metrics for each period, namely Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE). These metrics are calculated independently for each temporal slice and aggregated into a time-ordered performance table.

To quantify stability over time, the module employs the *Coefficient of Variation* (CV), defined as:

$$CV = \frac{\sigma}{\mu} \tag{5}$$

where $\sigma$ denotes the standard deviation and $\mu$ the mean of a metric across time periods. The CV provides a normalized measure of dispersion, enabling comparison of variability across metrics with different scales. Lower CV values indicate greater temporal stability and therefore higher robustness.

An overall robustness score is computed by averaging the CV values of MAE and MAPE, and transforming the result into the interval $[0, 1]$:

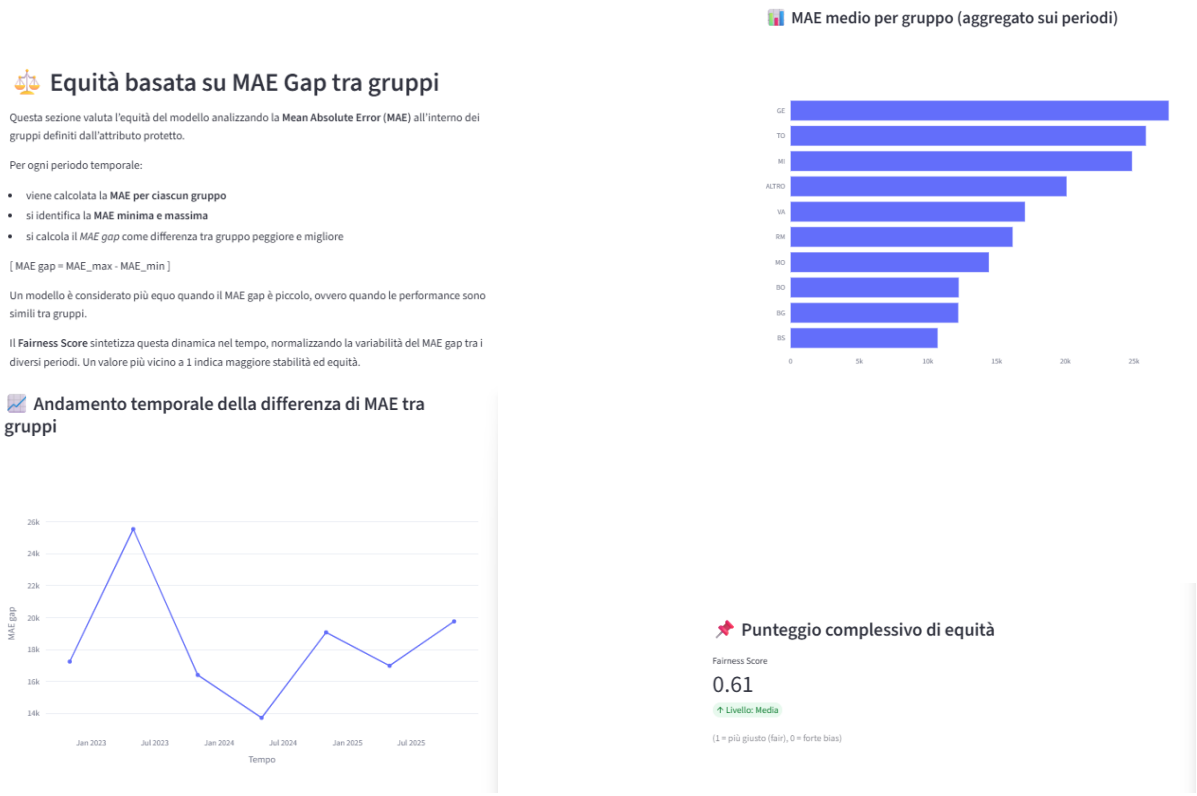


**Figure 5:** Fairness module results.

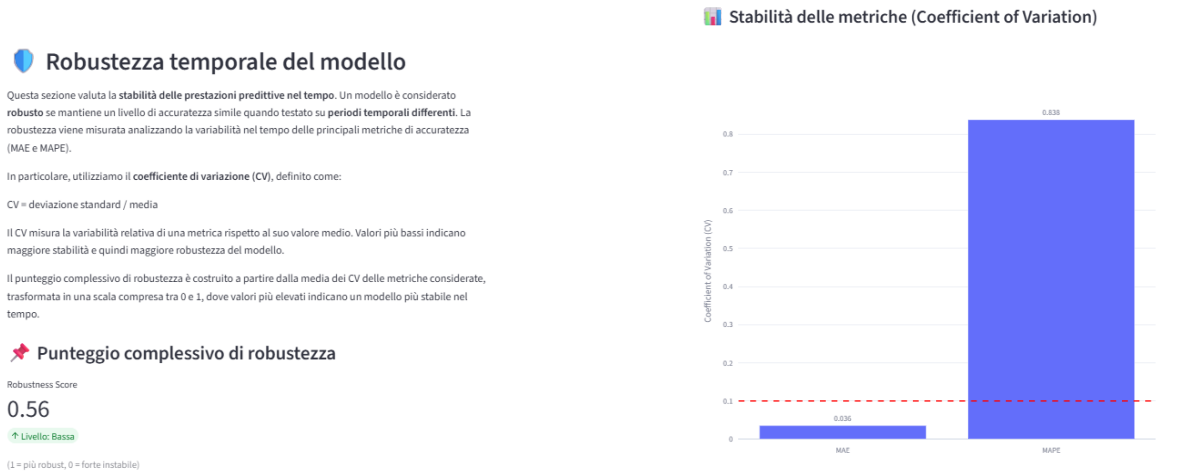


**Figure 6:** Robustness module results.

$$\text{Robustness Score} = \max(0, 1 - \overline{CV}) \tag{6}$$

where $\overline{CV}$ represents the mean coefficient of variation across the considered metrics. Values closer to 1 indicate high temporal stability, while values closer to 0 indicate strong performance variability over time. Figure 6 illustrates the stability analysis. The bar chart reports the CV values of MAE and MAPE, together with a reference threshold at $CV = 0.1$. In the presented case, all metrics exhibit CV values well below the threshold (MAE $\approx 0.051$, MAPE $\approx 0.001$), indicating limited temporal variability. Consequently, the overall robustness score of approximately 0.97 confirms a high level of temporal robustness.

## 9. SAFE Score

To provide a unified and interpretable assessment of model compliance, we introduce the SAFE Score, a composite indicator that aggregates three key dimensions of trustworthy machine learning: *accuracy*, *robustness*, and *fairness*. Each component is normalized to a bounded scale in $[0, 1]$, where higher values indicate better performance (i.e., lower risk). The SAFE Score is defined as the arithmetic mean of the available components:

$$\text{SAFE} = \frac{1}{K} \sum_{k=1}^{K} S_k,$$

where $S_k$ represents the normalized score of each dimension and $K$ is the number of valid components. We selected this design because simple aggregation facilitates adoption in real-world monitoring settings where complex weighting schemes may reduce transparency. We acknowledge that alternative aggregation strategies (e.g., weighted scoring, multi-objective optimization, or non-linear combinations) could better capture trade-offs between dimensions. Exploring such approaches is based on the clients' request and part of future work.

Figure 7 represents a holistic view of the model reliability and supports risk-aware governance in the high-stakes domain of POS transaction forecasting:


**Figure 7:** SAFE score module results.

Finally, a PDF report can be produced by selecting the "PDF report" button in the platform. All the four modules and the SAFE score components are summarized in this report.

## 10. Conclusion

SAFE-AI platform provides an integrated evaluation framework that moves beyond traditional performance monitoring. By combining accuracy, fairness, explainability, and robustness into a unified system, the platform supports Responsible AI governance, Regulatory compliance, Early detection of model degradation, and Transparent risk communication. By combining these perspectives into a unified framework and summarizing them through the SAFE Score, the platform provides an interpretable and systematic assessment of model reliability when applied to POS transactional data. Our empirical analysis for the specific POS transaction prediction use case revealed a stable predictive performance across periods. We also concluded high fairness gaps across economic sectors and provinces. Moreover, lag-based features showed Consistent importance levels in the considered time periods.

Future work includes extending the fairness module to evaluate also individual fairness and utilizing other statistical metrics and methods to evaluate accuracy, explainability and robustness. For example, evaluating the quality of explanations remains an open challenge and could be added to the explainability module. Moreover, future extensions of SAFE-AI could investigate trade-offs and potential correlations between evaluation dimensions.

## Acknowledgments

We acknowledge funding from the Italian PRIN2020 project FIN4GREEN - Finance for a sustainable, green and resilient society.

## Declaration on Generative AI

During the preparation of this work, the authors used Chatgpt in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

## References

[1] A. Zhuk, Ethical implications of ai in the metaverse, AI and Ethics 5 (2025) 5643–5654.
[2] L. Edwards, The eu ai act: a summary of its significance and scope, Artificial Intelligence (the EU AI Act) 1 (2021) 25.
[3] J. Laux, S. Wachter, B. Mittelstadt, Trustworthy artificial intelligence and the european union ai act: On the conflation of trustworthiness and acceptability of risk, Regulation & Governance 18 (2024) 3–32.
[4] C. Wijethilake, Ethical challenges in utilizing artificial intelligence in banking operations: A systematic review (2025).
[5] E. Prem, From ethical ai frameworks to tools: a review of approaches, AI and Ethics 3 (2023) 699–716.
[6] G. Babaei, P. Giudici, A statistical package for safe artificial intelligence: G. babaei, p. giudici, Statistical Methods & Applications 34 (2025) 499–517.
[7] M. Kholod, A. Celani, G. Ciaramella, The analysis of customers' transactions based on pos and rfid data using big data analytics tools in the retail space of the future, Applied Sciences 14 (2024) 11567.
[8] FINRA, Artificial intelligence in the securities industry, Financial Industry Regulatory Authority Report (2020).
[9] D. Amodei, et al., Concrete problems in ai safety, arXiv preprint arXiv:1606.06565 (2016).
[10] S. Barocas, A. D. Selbst, Big data's disparate impact, California Law Review 104 (2016) 671–732.
[11] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in Neural Information Processing Systems (NeurIPS) (2017).
[12] G. Babaei, P. Giudici, E. Raffinetti, A rank graduation box for safe ai, Expert systems with applications 259 (2025) 125239.
[13] Z. Mai, J. Zhang, Z. Xu, Z. Xiao, Financial sentiment analysis meets llama 3: A comprehensive analysis, in: Proceedings of the 2024 7th International Conference on Machine Learning and Machine Intelligence (MLMI), 2024, pp. 171–175.
[14] K. E. Mokhtari, B. P. Higdon, A. Başar, Interpreting financial time series with shap values, in: Proceedings of the 29th annual international conference on computer science and software engineering, 2019, pp. 166–172.
[15] K. K. Gokmenoglu, A. Amir, The impact of perceived fairness and trustworthiness on customer trust within the banking sector, Journal of Relationship Marketing 20 (2021) 241–260.