risks

MDPI

Article

Correlation Metrics for Safe Artificial Intelligence

Golnoosh Babaei and Paolo Giudici *

Department of Economics and Management, University of Pavia, 27100 Pavia, Italy; golnoosh.babaei@unipv.it
* Correspondence: paolo.giudici@unipv.it

Abstract

There is a growing need to provide AI risk management models that can assess whether AI applications are safe and trustworthy, to make them responsible. To date, there are a few research papers on this topic. To fill the gap, in this paper we extend the recently proposed SAFE framework, a comprehensive approach to measure AI risks across four key dimensions: security, accuracy, fairness, and explainability (SAFE). We contribute to the SAFE framework with a novel use of the coefficient of determination (R^2) to quantify deviations from ideal behavior not only in terms of accuracy but also for security, fairness, and explainability. Our empirical findings shows the effectiveness of the proposal, which leads to a more precise measurement of risks of AI regression applications, which involve the prediction of continuous response variables.

Keywords: SAFE AI metrics; responsible AI; coefficient of determination

1. Introduction

Artificial Intelligence (AI) systems are increasingly deployed in domains where their decisions have significant societal, ethical, and economic consequences, ranging from healthcare and finance to criminal justice and education. While their adoption promises efficiency and innovation, growing evidence shows that these systems can also introduce substantial risks, including biased outcomes, adversarial vulnerabilities, and poor generalization in real-world settings. As such, the demand for responsible AI—AI that is not only performant but also secure, fair, and robust—has become a central concern in both academic and policy-making circles Floridi et al. (2018); Jobin et al. (2019).

Indeed, the widespread use of Artificial Intelligence (AI) requires us to develop advanced statistical methods that can measure its risks, in line with recently proposed regulations and standards such as the European Artificial Intelligence Act (European Commission (2022)), the American Artificial Intelligence Risk Management framework (United States National Institute of Standards and Technologies (2023)), the OECD framework for the classification of AI Systems (Organisation for Economic Cooperation and Development (2022)), and the ISO/IEC 23894 international standards (International Organization for Standardization and International Electrotechnical Commission (2023)).

From an organizational and managerial governance viewpoint, the main risks that can arise from the application of AI derive from the violation of four main principles: accuracy, explainability, fairness, and security.

First, accuracy. AI applications are complex, and consume a great deal of energy and costs. Thus, their efficiency requires accurate results. Second, explainability. AI applications typically have an intrinsic non-transparent ("black-box") nature. This is a problem in regulated industries, as authorities aimed at monitoring the risks arising from

Academic Editor: Pasquale Cirillo

Received: 18 July 2025
Revised: 5 September 2025
Accepted: 9 September 2025
Published: 12 September 2025

Citation: Babaei, Golnoosh, and Paolo Giudici. 2025. Correlation Metrics for Safe Artificial Intelligence. *Risks* 13: 178. https://doi.org/10.3390/risks13090178

the application of AI may not validate them (see, e.g., Bracke et al. 2019). Third, security. Cyber attacks and data poisoning naturally increase with the increase in AI applications. This is one of the most frequent risks related to the resilience and sustainability of AI applications. Fourth, fairness. The application of AI may pose societal problems, among which fairness and the risk of discrimination is the most relevant.

Indeed, all the recently proposed regulations and recommendations introduce compliance requirements for these four principles within a risk-based approach to AI applications. In the regulatory context, several applications of AI have high risks and, therefore, require an appropriate risk management model. To develop such a model, we need to express the regulatory requirements in terms of statistical variables, to be measured by appropriate statistical metrics.

A growing body of work has proposed frameworks for measuring and mitigating various AI-related risks. For example, frameworks such as *FAIR* (Fairness, Accountability, Interpretability, and Responsibility) and *FAT* (Fairness, Accountability, and Transparency) Barocas and Selbst (2018); Selbst et al. (2019) emphasize the ethical dimensions of AI. Others have focused on specific aspects such as robustness to adversarial examples Goodfellow et al. (2015); Szegedy et al. (2014), secure learning against data poisoning Steinhardt et al. (2017), and algorithmic fairness across demographic groups Hardt et al. (2016); Kleinberg et al. (2016). However, these frameworks often consider these dimensions in isolation, lacking a unified metric to compare and trade off between them.

A related line of research is to employ AI to measure risks. Papers in this stream of research include those by Naim (2022) and Sundra et al. (2023), who consider the application of AI in financial risk management; Sachan et al. (2020); Bussmann et al. (2020); Moscato et al. (2021); Liu et al. (2022) and Bücker et al. (2022), who employ AI to measure credit risk; Giudici and Polinesi (2021) and Giudici and Abu-Hashish (2019), who employ AI to measure contagion risks in algorithmic trading and crypto exchanges; Ganesh and Kalpana (2022), who review AI methodologies for supply chain risk management, whereas Frost et al. (2019) and Kuiper et al. (2021) do so for financial intermediation; Ainslie et al. (2017); Melancon et al. (2021); Aldasoro et al. (2022) and Giudici and Raffinetti (2021), who employ AI to measure IT risk and cyber risks; and Achakzai and Juan (2022) and Chen et al. (2018), who employ AI to detect financial frauds and money laundering.

The previously mentioned papers propose AI methodologies to measure different types of risks, often from a financial perspective. We instead propose a methodology to measure the "model" risks the AI itself generates, following the recently proposed SAFE AI risk management framework (Babaei et al. (2025)), which evaluates AI models along four fundamental dimensions: security, accuracy, fairness, and explainability. The SAFE framework is motivated by the need for a holistic and quantitative approach to AI risk assessment.

We contribute to the SAFE framework with a novel use of the coefficient of determination (R^2)—traditionally used to measure model fit in regression—as a general-purpose metric to quantify deviations from ideal behavior not only in terms of accuracy but also for security, fairness, and explainability. By extending R^2 to capture the proportion of variation in each of these domains, we offer a coherent and interpretable tool to compare models that aim to predict continuous responses.

The rest of the paper is structured as follows: Section 2 formalizes the SAFE AI dimensions and extends the R^2 metric to each of them. Section 3 presents the application of our proposal to a case study, followed by a discussion of the obtained results in Section 4. We conclude in Section 5 with a summary of our contribution.

2. Extending the SAFE AI Metrics

The security, accuracy, fairness, and explainability (S.A.F.E.) AI metrics proposed in Babaei et al. (2025) are derived from the Lorenz curve. Their core is Rank Graduation Accuracy (RGA), which is a rank-based metric to assess the accuracy of a model, a classifier or a regressor, as explained in Giudici and Raffinetti (2025). We now briefly recall them.

Consider Y^* and Y^{**}, two statistical variables (continuous, ordinal or binary). To build the Lorenz curve, we can utilize the sorted Y^* values in a non-decreasing order. For $i = 1, \ldots, n$, the Lorenz curve can be defined by the pairs $(i/n, \sum_{j=1}^{i} y^*_{r^*_j}/(n\bar{y}^*))$, where r^*_j indicates the non-decreasing ranks of Y^* and $\bar{y}^*$ indicates the mean of Y^*. Using the same variable Y^* but in a non-increasing order, we can build the dual Lorenz curve. For $i = 1, \ldots, n$, the dual Lorenz curve can be defined by the pairs $(i/n, \sum_{j=1}^{i} y^*_{r^*_{n+1-j}}/(n\bar{y}^*))$, where r^*_{n+1-j} indicates the non-increasing ranks of Y^*.

If the Y^* values are sorted according to the ranks of the Y^{**} values in a non-decreasing order, we can define the concordance curve. For $i = 1, \ldots, n$, the concordance curve is defined by the pairs $(i/n, \sum_{j=1}^{i} y^*_{r^{**}_j}/(n\bar{y}^*))$, where r^{**}_i indicates the non-decreasing ranks of Y^{**}. Regarding accuracy evaluation, Y^* and Y^{**} refer to the actual and predicted values, respectively. RGA is the ratio of the area below the last two curves (dual Lorenz curve and concordance curve) to the area between the first two curves (Lorenz and dual Lorenz curves).

Babaei et al. (2025) extended the construction behind the RGA to different types of comparison between pairs of variables that arise from the assessment of ML models. The comparison between the predicted Y^* and the predicted Y^* when the input data are modified lead to a Rank Graduation Robustness (RGR) measure, which can be employed to assess whether the output from Artificial Intelligence applications is robust under variations in the data, caused by extreme events, or by cyber attacks, which may alter the values of the explanatory variables. We remark that the SAFE framework essentially equates security with robustness and can accommodate alternative red teaming scenarios, such as prompt injection or coordinated-agent failure.

The comparison between the predicted values Y^* of a machine learning model including and excluding one or more explanatory variables leads to the Rank Graduation Explainability (RGE) measure, which can be employed to assess the interpretability of the AI output. The comparison between the predicted Y^*, conditionally and unconditionally to a protected attribute, such as gender or race, leads to a Rank Graduation Fairness (RGF) measure, which can be employed to assess the bias and degree of discrimination of AI.

The advantage of the RG divergence is that it can be used to assess, in a consistent way, the accuracy, the robustness, the explainability and the fairness for different types of response variables. Therefore, understanding a unique formula and concept, we can measure the mentioned AI principles and easily interpret them. Against the advantages, the above metrics have the disadvantage of reducing the comparison between different models to the comparison of the ranks. If two models have the same predictive ranks, RGA is the same if RGR and RGE are the same. Similarly, if two perturbed models have the same predicted ranks, RGR is the same. And, if two simplified models, with some variables removed, have the same predicted ranks, RGE is the same. This may not be ideal when the response variable is continuous. An additional disadvantage, for continuous response models, is the need to assume that the response is non-negative, implicit in the construction of the Lorenz curve.

To overcome the previous limitations, in this paper we propose to extend the SAFE AI metrics by employing the coefficient of determination R^2 in regression problems, characterized by a continuous response.

In regression analysis, R^2 is a widely used metric that quantifies the proportion of the variance in the dependent variable that is explained by the independent variables. Given a set of n observations $\{(x_i, y_i)\}_{i=1}^n$, and a regression function $\hat{y}_i = f(x_i)$, R^2 is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \tag{1}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ is the mean of the observed values.

The numerator in Equation (1) represents the *residual sum of squares* (RSS), which captures the error of the model. The denominator, known as the *total sum of squares* (TSS), quantifies the total variance in the observed data.

The notion of R^2 as a measure that describes the percentage of variability of a response variable explained by a set of explanatory independent variables can be extended to a predictive context. In this case, the n observations are partitioned in a training sample of n_1 observations, on which the regression function $\hat{y}_i = f(x_i)$ is estimated (learned), and a test sample n_2, on which the regression function is applied, to derive predictions which are compared to the ground truth values. Thus, R^2 becomes

$$R^2 = 1 - \frac{\sum_{i=1}^{n_2} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_2} (y_i - \bar{y})^2}, \tag{2}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n_2} y_i$ is the mean of the observed values in the test sample. A key difference between the R^2 in Equation (1) and that in Equation (2) is that, while in the former case the "explained variance" $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ is smaller than the total variance $\sum_{i=1}^n (y_i - \bar{y})^2$, as the fitted and error vectors are orthogonal, this may not be true in the latter case.

We remark that (2) can be extended to a classification problem, in which the response variable is binary, resorting to the Brier score (see, e.g., Rufibach 2010). To that end, given the true 0/1 responses, y_i, whose mean is the probability of 1, p, $\hat{y}_i$ becomes the predicted probability of a 1 response for the ith individual, $\hat{p}_i$. Formula (2) becomes:

$$R^2 = 1 - \frac{\sum_{i=1}^{n_2} (y_i - \hat{p}_i)^2}{\sum_{i=1}^{n_2} (y_i - p)^2}, \tag{3}$$

In this paper, we will employ the predictive R^2 in Equation (2) not only to measure predictive accuracy, but also to measure other principles in the SAFE AI framework such as security, explainability, and fairness.

Concerning robustness, which is employed as a measure of security of AI applications, we measure the resilience of a model against modifications to the data. Consider $\hat{y}_i$ and $\hat{y}_i^*$ as the original predicted values and predicted values after modifying the data. Equation (2) can be rewritten as follows to measure robustness:

$$R^{*2} = 1 - \frac{\sum_{i=1}^{n_2} (\hat{y}_i - \hat{y}_i^*)^2}{\sum_{i=1}^{n_2} (\hat{y}_i - \bar{y})^2} \tag{4}$$

In terms of explainability, we measure the contribution of a variable to the model predictions. Consider $\hat{y}_i$ and $\hat{y}_i^j$ as the original predicted values and the predicted values obtained after removing a variable x_j from the predictors for $j = 1, \ldots, J$. Equation (2) can be rewritten as follows to measure the explainability of the J available variables:

$$R^{j2} = 1 - \frac{\sum_{i=1}^{n_2} (\hat{y}_i - \hat{y}_i^j)^2}{\sum_{i=1}^{n_2} (\hat{y}_i - \bar{y})^2} \tag{5}$$

A value of R^{j2} equal to zero shows a small and negligible contribution for the considered variable x_j. The higher R^{j2}, the more important the variable.

Finally, the comparison between the R^2 values in different protected groups considering a variable (predictor) as a protected variable leads to a fairness measurement based on the model parity concept.

Formally, let R^2_A and R^2_B be two predictive R^2 values, calculated according to Equation (2) separately for all data that belong, respectively, to individuals in groups A and B. For example, A can be a group of females and B a group of males. To measure fairness we can then calculate

$$R^2_F = |R^2_A - R^2_B| \tag{6}$$

A small value of R^2_F indicates that a model achieves similar R^2 values when applied to different protected groups and, for this reason, we can consider it as a fair model. Differently, a large value of R^2_F in (6) indicates that the model has a degree of unfairness.

We conclude this section with an illustration of the practical importance of our proposal, in terms of the risk management standard established by the International Organization for Standardization and International Electrotechnical Commission (2023), in Table 1. The Table lists, for each risk management requirement, the corresponding action based on the S.A.F.E. AI framework presented in this paper.

Table 1. ISO/IEC 23894:2023 AI Risk Management controls, corresponding S.A.F.E. actions, and accountability across lifecycle.

Control Category	ISO/IEC 23894:2023 Requirement Actions	S.A.F.E. Actions	Who Measures What	When in the AI Lifecycle	Who Escalates/Rolls Back
Context Establishment	Define scope, objectives, boundaries, and stakeholders; identify intended use and potential misuse.	Define AI applications and, for each of them, response and explanatory variables.	Business owners, risk managers—define scope and variables.	Conception/ project initiation.	Governance board or project steering committee.
Risk Identification	Identify risks from data, models, processes, deployment environment, and misuse scenarios across lifecycle.	For AI application identify which S.A.F.E. metric applies.	Data scientists, domain experts—map risks and metrics.	Data collection and model design.	Risk manager with compliance team.
Risk Analysis	Assess likelihood, severity, and uncertainty of risks; consider emergent and systemic risks.	Calculate the identified metrics, equate likelihood with their complement to one, and severity with the quantity and quality of impacted stakeholders.	Data scientists, quantitative risk analysts—compute metrics.	Model development and validation.	Chief Risk Officer (CRO) or equivalent oversight function.
Risk Evaluation	Compare risks to acceptance criteria; prioritize risks for treatment.	Calculate thresholds, using appropriate statistical tests, such as Diebold and Mariano (1995).	Validation team—apply thresholds, statistical testing.	Pre-deployment validation.	Independent risk committee or model validation unit.
Risk Treatment	Apply measures such as data quality and governance checks, bias and fairness analysis, robustness testing, explainability, human oversight, security safeguards, fallback and incident response mechanisms.	Interpret deviation from threshold as potential lack of fairness, robustness and security, explainability and human oversight or fallback accuracy.	ML engineers, auditors—monitor controls and implement safeguards.	Deployment and operational readiness.	AI ethics board, compliance lead, or IT security head.
Monitoring	Continuously monitor AI performance and risks; implement feedback loops and update risk assessments.	Update model, metrics and comparison with threshold as new training data arrives.	Operations/ monitoring team—track metrics over time.	Post-deployment, continuous operation.	Incident response manager, system owner.
Communication	Engage stakeholders in risk decisions; ensure transparency and reporting of risk management outcomes.	Report S.A.F.E. metrics values, in comparison with thresholds, to stakeholders.	Risk managers, reporting officers—communicate metrics and risks.	Throughout lifecycle, periodic reviews.	Senior management, external regulators if required.

3. Application

To demonstrate the practical utility of the proposed extended SAFE AI metrics, we apply them to the `employee` dataset, available from the `stima` R package.[1] The dataset contains various features describing the characteristics of the employees of a bank, including salary, job level, and experience. Table 2 describes the variables available in the dataset.

Table 2. The available explanatory variables in the employee dataset.

Variable	Definition
salary	a numeric variable, used as response variable: current salary in US dollars
age	a numeric variable: age in years
edu	a numeric variable: educational level in years
startsal	a numeric variable: beginning salary in US dollars
jobtime	a numeric variable: months since hire
prevexp	a numeric variable: previous work experience in months
minority	a factor variable: minority classification with levels `min`, indicating minority, and `no_min`, no minority
gender	a factor variable: gender type with levels `f`, indicating female, and `m`, indicating male
jobcat	a factor variable: type of job with levels Clerical, Custodial, and Manager

We consider a regression problem, where the target variable is `salary_growth`, which represents the actual numeric increase in an employee's salary over a specified period. The presence of a continuous response allows us to assess the effectiveness of our extended SAFE AI metrics. To that aim, we fit a regression model using the available feature variables, and evaluate the model using our proposed R^2-based metrics.

In terms of accuracy, our trained linear regressor leads to a R^2 value equal to 0.4023, which shows that about 40% of the variability of the response can be explained by the considered machine learning model. The RGA metric proposed in Babaei et al. (2025) turns out to be equal to 0.8925. The comparison between the predictive R^2 and RGA emphasizes that the calibration task (predicting numerical values) is more difficult than the discrimination task (predicting ranks).

We proceed to determine explainability in terms of the proposed R^{j2} metric. The obtained values are in Table 3.

Table 3. R^{j2} values representing the contributions of the explanatory variables to the regression model.

Variable	R^{j2}
edu	0.4523
gender	0.1405
prevexp	0.0392
jobtime	0.0274
minority	0.0207
age	0.0001

From Table 3, note that the most important variable to predict the rate of growth of salaries is the education level, followed by the years of previous experience and the time spent in the company. Notice that the age is not relevant; its effect is likely absorbed by the time spent in the company, highly correlated with it. Note also that gender has some importance, indicating some bias in salary growth. The other possible protected variable, minority, has a much lower importance. For comparison with the SAFE metrics in Babaei et al. (2025), Table 4 contains the RGE values for the same model.

Table 4. R^E (Rank Graduation Explainability) values values representing the contributions of the explanatory variables to the regression model.

Variable	*RGE*
edu	0.0910
gender	0.0256
jobtime	0.0094
prevexp	0.0069
minority	0.0063
age	0.0000

Looking at the results reported in Table 4, we can see that the most important variable to predict the rate of growth of salaries is the education level, followed by gender, the time spent in the company, and the previous experience. Overall, the results from RGE are consistent with those from R^{j2} with a stronger effect of gender and, in general, smaller values of explainability.

We then consider model robustness, considering the original predicted values and the predictions obtained perturbing the input variables. As perturbation type we consider a permutation in the original data. More precisely, we perturb all input variables by swapping the 5% top with the 5% bottom observations.

The result of the robustness metric, calculated in terms of our proposed R^{*2} statistics, is equal to 0.0389. This indicates that the model has a low degree of robustness, consistently with the limited sample size. Note that the application of the RGR metric in Babaei et al. (2025) leads to an RGR value of 0.7621. This is a higher value than that obtained using R^{*2}, consistent with the fact that statistics based on ranks are more robust than those based on actual values.

We remark that R^{*2} is rather general, and can be applied to different types of perturbations, which simulate different types of adversarial attack, for example within a red teaming framework, such as prompt injection, see, e.g., Pathade (2025), or coordinated-agent failure, see, e.g., Manheim and Garrabrant (2018). To this aim, it is sufficient to know the data, as modified by the perturbation, and apply R^{*2} to the comparison of the model output before and after the perturbation.

We also remark that an important aspect of the proposed R^{*2} statistics is that, being a function of a mean squared error, it is easy to derive test statistics, such as Diebold and Mariano's (Diebold and Mariano (1995)). For any given type I error probability level (such as 5%) an escalation threshold value (such as 1.96) can be derived and, thus, employed to escalate and prompt a deployment action.

Finally, we have obtained that the model imparity, employing gender as a protected variable, is equal to 0.0277. Note that the application of the SAFE AI framework in Babaei et al. (2025) leads to a model imparity equal to 0.0462. Both the original SAFE metric and our extension indicate a limited degree of bias, in line with the small value of explainability of the gender variable obtained with our proposed metric (but not using RGE).

Benchmarking

For a further comparison of our extension with the original SAFE metrics, in this section we consider a classification problem, in which the objective is to predict whether an employee's salary has doubled.

We define a binary target variable, `doubling_salary`, which takes the value 1 if an employee's salary has increased by a factor of two or more, and 0 otherwise. This classification problem can be evaluated using the classic SAFE AI framework, or more

traditional metrics, such as accuracy, the Area Under the Curve (AUC), the $F1$ score and classic group-based fairness measures.

To this aim, we train a simple binary classifier, logistic regression, on the dataset. The resulting AUC, F1 score, and accuracy of the classifier are equal to 0.7523, 0.6763, and 0.6831, respectively, all indicating acceptable (but not excellent) levels.

We now consider the application of the classic SAFE metrics. RGA is equal to AUC, in this case 0.7523.

In terms of explainability, we consider the original predicted values and the predictions after removing each of the variables to find the contributions of the variables, in terms of the RGE metric in Babaei et al. (2025). Table 5 contains the results.

Table 5. *RGE* values as the explanatory scores representing the contributions of the explanatory variables to a classification model.

Variable	*RGE*
age	0.3329
jobtime	0.1386
prevexp	0.0009
gender	0.0005
edu	0.0001
minority	0.0000

Looking at the results in Table 6, note that the most important variables for the doubling of the salary are the age of an employee and their time spent in the company. This is quite consistent with a traditional company, as the bank whose data is considered in this paper.

For comparison, we now consider the application of our proposed R^{j2} metric to assess explainability for the classification problem. Table 6 contains the results.

Table 6. R^{*2} values as the explanatory scores representing the contributions of the explanatory variables to a classification model.

Variable	R^{*2}
age	0.9992
jobtime	0.3684
prevexp	0.0091
gender	0.0016
edu	0.0008
minority	0.000

Comparing Table 6 with Table 5, note that the ranks of the R^{*2} values are consistent with those of the RGE values, with age and jobtime as the most important variables.

In terms of robustness, the application of the RGR metric in the SAFE AI approach leads to a value equal to 0.7892. The application of our proposed R^{*2} value leads to a value equal to 0.0593. Both results are in line with those obtained for the regression problem, with our metric indicating low robustness, consistent with the relatively small amount of data considered.

Finally, in terms of fairness, the RGE statistics leads to a model imparity equal to 0.0317. The same imparity, calculated with our proposed R^{*2} statistics, is equal to 0.0529. Both cases indicate little bias, consistent with that obtained for the regression problem.

Overall, the results from the benchmarking exercise indicate once more that the calibration task is different from the discrimination task. While in the latter the explainability

and fairness metrics give similar results, in the former case they do not, and show that the proposed R^2 metrics are more precise than the RG metrics in Babaei et al. (2025), as they compare values rather than their ranks.

4. Discussion

The obtained results and empirical findings indicate that the proposed R^{*2} safe Artificial Intelligence metrics are well suited to evaluate machine learning models aimed at predicting continuous response variables within a calibration framework.

The comparison with the Rank Graduation metrics of Babaei et al. (2025) shows that our proposed metrics lead to values of accuracy, robustness and explainability that are consistent with the nature of the calibration problem, more difficult and less robust than the discrimination problem considered in classification problems, where the RG metrics of Babaei et al. (2025) are more useful.

To summarize, the R^{*2} and the RG metrics should be seen as complementary. While the latter are universal, and can be used for all types of responses, the former are more precise, but can be used, and should indeed be used, when the objective of the study is the prediction of point values and not of ranks.

5. Conclusions

We have presented a set of novel SAFE AI risk management metrics, based on a predictive R^2 measure. The metrics extend those proposed using the rank-based measures in Babaei et al. (2025). While the latter are more general than the metrics proposed in this paper, as they can be applied to all types of variables, the metrics considered here apply only to continuous response variables but are more informative, in the same fashion as a point prediction is more informative than a categorical prediction.

We have tested our proposal on a real and well known dataset, concerning the growth of the salaries of a set of employees. Our results indicate that the model has a moderate accuracy, and a limited robustness, consistent with the considered small sample size. In terms of explainability, to predict growth of the salary in continuous time, what matters most is the education level, the years of experience and the time spent in the company. Finally, the model shows a reasonable level of fairness.

The potential users of the proposed solution include all stakeholders that participate to the AI lifecycle: providers, deployers, users, consumers and authorities that control and monitor AI applications. We also remark that the proposed solution is quite general and can be applied to alternative frameworks, not only based on tabular data, but also text and other unstructured data.

Indeed, the main limitation of this study is that is focused on tabular data, and the application to the financial sector. Further research work should include the extension of the work to unstructured data, including generative AI models, and to other sectors, such as healthcare and manufacturing. Further work may also include extending the security metrics to include simulation scenarios based on different types of read teaming.

Author Contributions: Conceptualization, G.B. and P.G.; methodology, P.G.; software, G.B.; validation, P.G.; formal analysis, G.B.; investigation, G.B. and P.G.; writing—original draft preparation, G.B.; writing—review and editing, P.G.; supervision, P.G. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by: the European Union-NextGenerationEU, in the framework of the GRINS- Growing Resilient, INclusive and Sustainable (GRINS PE00000018).

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://examples.rpkg.net/packages/stima/reference/employee.ob (accessed on 10 September 2025).

Acknowledgments: The authors thank the discussant and the participants at the SAFE machine learning workshop in Pavia on 19 June 2025, where the paper was presented, and where G.B. received the prize for the best presentation from the MDPI journal *Risks*. The authors also thank the editor and the three anonymous reviewers for their constructive and accurate remarks.

Conflicts of Interest: The authors declare no conflict of interest.

References

Achakzai, Muhammad Atif Khan, and Peng Juan. Using machine learning meta-classifiers to detect financial frauds. *Finance Research Letters* 48: 102915. [CrossRef]

Ainslie, Russell, John McCall, Sid Shakya, and Gilbert Owusu. 2017. Predicting service levels using neural networks. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*. Berlin/Heidelberg: Springer, pp. 411–416. [CrossRef]

Aldasoro, Iñaki, Leonardo Gambacorta, Paolo Giudici, and Thomas Leach. 2022. The drivers of cyber risk. *Journal of Financial Stability* 60: 100989. [CrossRef]

Babaei, Golnoosh, Paolo Giudici, and Emanuela Raffinetti. 2025. A rank graduation box for safe AI. *Expert Systems with Applications* 259: 125239. [CrossRef]

Binns, Reuben. 2018. Fairness in machine learning: Lessons from political philosophy. Paper presented at 2018 Conference on Fairness, Accountability and Transparency, PMLR, New York, NY, USA, 23–24 February 2018; pp. 149–159.

Bracke, Philippe, Anupam Datta, Carsten Jung, and Shayak Sen. 2019, August. *Machine Learning Explainability in Finance: An Application to Default Risk Analysis*. Working Paper 816. London: Bank of England. [CrossRef]

Bücker, Michael, Gero Szepannek, Alicja Gosiewska, and Przemyslaw Biecek. 2022. Transparency, auditability, and explainability of machine learning models in credit scoring. *Journal of the Operational Research Society* 73: 70–90. [CrossRef]

Bussmann, Niklas, Paolo Giudici, Dimitri Marinelli, and Jochen Papenbrock. 2020. Explainable machine learning in credit risk management. *Computational Economics* 57: 2013–16. [CrossRef]

Chen, Zhiyuan, Le Dinh Van Khoa, Ee Na Teoh, Amril Nazir, Ettikan Kandasamy Karuppiah, and Kim Sim Lam. 2018. Machine learning techniques for anti-money laundering (aml) solutions in suspicious transaction detection: A review. *Knowledge and Information Systems* 57: 245–85. [CrossRef]

Diebold, Francis X., and Roberto S. Mariano. 1995. Comparing predictive accuracy. *Journal of Business & Economic Statistics* 13: 253–63. [CrossRef]

European Commission. 2022. Artificial Intelligence Act. Available online: https://artificialintelligenceact.eu (accessed on 18 August 2025).

Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, and et al. 2018. Ai4people—An ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. *Minds and Machines* 28: 689–707. [CrossRef]

Frost, J., L. Gambacorta, Y. Huang, and P. Zbindnen. 2019. Bigtech and the changing structure of financial intermediation. *Economic Policy* 34: 761–99. [CrossRef]

Ganesh, A. Deiva., and P. Kalpana. 2022. Future of artificial intelligence and its influence on supply chain risk management—A systematic review. *Computers and Industrial Engineering* 169: 108206. [CrossRef]

Giudici, Paolo, and Emanuela Raffinetti. 2021. Explainable ai in cyber risk management. *Quality and Reliability Engineering International 38*: 1318–26. [CrossRef]

Giudici, Paolo, and Emanuela Raffinetti. 2025. RGA: a unified measure of predictive accuracy. *Advances in Data Analysis and Classification 19(1)*: 67–93 [CrossRef]

Giudici, Paolo, and Gloria Polinesi. 2021. Crypto price discovery through correlation networks. *Annals of Operations Research* 299: 443–57. [CrossRef]

Giudici, Paolo, and Iman Abu-Hashish. 2019. What determines bitcoin exchange prices? A network var approach. *Finance Research Letters* 28: 309–18. [CrossRef]

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. Paper presented at International Conference on Learning Representations (ICLR), San Diego, CA, USA, May 7–9.

Hardt, Moritz, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems* 29: 3315–23.

International Organization for Standardization and International Electrotechnical Commission. 2023. *Information Technology—Artificial Intelligence—Guidance on Risk Management*. ISO/IEC 23894:2023. London: British Standards Institution.

Jobin, Anna, Marcello Ienca, and Effy Vayena. 2019. The global landscape of ai ethics guidelines. *Nature Machine Intelligence* 1: 389–99. [CrossRef]

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv* arXiv:1609.05807. [CrossRef]

Kuiper, O., M. Van den Berg, J. Van der Burgt, and S. Leijnen. 2021. Exploring explainable ai in the financial sector: Perspectives of banks and supervisory authorities. In *Artificial Intelligence and Machine Learning*. Berlin/Heidelberg: Springer. [CrossRef]

Liu, Wanan, Hong Fan, and Meng Xia. 2022. Credit scoring based on tree-enhanced gradient boosting decision trees. *Expert Systems with Applications* 189: 116034. [CrossRef]

Manheim, David, and Scott Garrabrant. 2018. Multiparty dynamics and failure modes for machine learning systems. *arXiv* arXiv:1810.10862.

Melançon, Gabrielle Gauthier, Philippe Grangier, Eric Prescott-Gagnon, Emmanuel Sabourin, and Louis-Martin Rousseau. 2021. A machine learning-based system for predicting service-level failures in supply chains. *Informs Journal on Applied Analytics* 51: 200–12. [CrossRef]

Moscato, Vincenzo, Antonio Picariello, and Giancarlo Sperlì. 2021. A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications* 165: 113986. [CrossRef]

Naim, Arshi. 2022. Role of artificial intelligence in business risk management. *American Journal of Business Management, Economics and Banking* 1: 55–66.

Organisation for Economic Cooperation and Development. 2022. Framework for the Classification of AI Systems. Available online: https://oecd.ai/en/ai-principles (accessed on 18 August 2025).

Pathade, Chetan. 2025. Red teaming the mind of the machine: A systematic evaluation of prompt injection and jailbreak vulnerabilities in llms. *arXiv* arXiv:2505.04806. [CrossRef]

Rufibach, Kaspar. 2010. Use of brier score to assess binary predictions. *Journal of Clinical Epidemiology* 63: 938–39. [CrossRef]

Sachan, Swati, Jian-Bo Yang, Dong-Ling Xu, David Eraso Benavides, and Yang Li. 2020. An explainable ai decision-support-system to automate loan underwriting. *Expert Systems with Applications* 144: 113100. [CrossRef]

Selbst, Andrew D, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. Paper presented at Conference on Fairness, Accountability, and Transparency, Rio de Janeiro, Brazil, June 3–6. pp. 59–68.

Steinhardt, Jacob, Pang Wei Koh, and Percy Liang. 2017. Certified defenses for data poisoning attacks. *Advances in Neural Information Processing Systems* 30: 3517–29.

Sundra, B. M., K. Sathiyamurthi, and G. Subramanian. 2023. Critical evaluation of applying machine learning approaches for better handling bank risk management in the post-modern era. *Scandinavian Journal of Information Systems* 35: 1228–31.

Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. *arXiv* arXiv:1312.6199. [CrossRef]

United States National Institute of Standards and Technologies. 2023. Ai Risk Management Framework. Available online: https://www.nist.gov/itl/ai-risk-management-framework (accessed on 18 August 2025).