Quality & Quantity
https://doi.org/10.1007/s11135-023-01613-y

A Rank Graduation Accuracy measure to mitigate Artificial Intelligence risks

Emanuela Raffinetti[1]

Accepted: 10 January 2023

Abstract
A key point to assess the applications of machine learning models in Artificial Intelligence (AI) is the evaluation of their predictive accuracy. This because the "automatic" choice of an action crucially depends on the made prediction. While the best model in terms of fit to the observed data can be chosen using a "universal" - and therefore automatable - criterion, based on the models' likelihood, such as AIC and BIC, this is not the case for the best model in terms of predictive accuracy. To fill the gap, we propose a Rank Graduation Accuracy (*RGA*) measure which evaluates the concordance between the ranks of the predicted values and the ranks of the actual values of a series of observations to be predicted. We apply the *RGA* to a use-case that concerns the measurement of the financial risks that arise from crypto assets. The *RGA* appears as a "universal" alternative predictive model selection criterion that, differently from standard measures, such as the Root Mean Squared Error, is robust to the presence of outlying observations.

Keywords Predictive accuracy · Robustness · Financial risk · Crypto assets

1 Introduction

The growing availability of data and computational power has allowed innovative developments in the field of Artificial Intelligence (AI). If, on the one hand, AI has the potential to yield economic and societal benefits, on the other hand, the consideration of the possible negative consequences on strongly impacting actions has led policy makers and regulators to a degree of suspicion towards AI applications. This because AI methods are based on a "black-box", where input data are transformed through complex processes without controlling and monitoring the deriving risks.

In particular, "black-box" AI is not suitable in insurance services, where platform and cyber risks can arise (see e.g., Aldasoro et al. 2022). Indeed, as AI includes a wide set of technologies and methods which are disrupting the insurance domain, the main challenge for insurers is resorting to suitable regulatory actions to mitigate operational, reputational

✉ Emanuela Raffinetti
emanuela.raffinetti@unipv.it

[1] Department of Economics and Management, University of Pavia, Via San Felice al Monastero 5, 27100 Pavia, Italy

and strategic risks (see e.g., Ceylan 2022; Eling et al. 2022; Mullins et al. 2021). To face this issue, AI models have to be trustworthy and reliable, providing details or reasons to make their functioning clear or easy to understand.

In line with the requirement of a trustworthy AI, the European Commission has proposed, on 21 April 2021, an AI act (https://artificialintelligenceact.eu), which has become a template for both the European and other countries in the world. The act assigns applications of AI to three risk categories. First, applications and systems that create an unacceptable risk are banned. Second, high-risk applications, such as those which involve ranking of individuals or of organisations, are permitted but conditionally on compliance requirements. Third, low-risk applications, not explicitly banned or listed as high-risk, are largely left unregulated.

To fulfill trustworthiness, AI methods have to be safe. A safe application of AI must satisfy four basic key-principles, summarised as: sustainability, accuracy, fairness and explainability. "Sustainability" means that AI methodologies have to be robust, both in terms of data and computation. "Fairness" implies that AI methods should not discriminate by age, ethnicity, gender or other population groups. The "Explainability" key-principle requires that AI models are interpretable in terms of their drivers (see e.g., Bracke et al. 2019).

In this paper, we focus on predictive accuracy and explainability. Concerning the latter, researchers have recently addressed the issue of how machine learning models can be made explainable. Existing papers may be divided in two main approaches: global explanations and local explanations. While global explanations describe the model as a whole, in terms of which explanatory variables most determine the predictions, for all units, local explanations aim at interpreting individual predictions, at the single unit level (see e.g., Aas et al. 2021; Joseph 2019; Molnar 2022).

Among the local explanation methods, the Shapley value approach, originally introduced in Shapley (1953) and implemented by Lundberg and Lee (2017) and Strumbelj and Kononenko (2010), is gaining importance due to its remarkable properties. According to the Shapley value procedure, the total change in prediction is divided among the features in a way which is fair to their contributions across all the possible sets of features. A measure of the contribution, associated with each predictor, to each point prediction of a machine learning model is then provided.

Several research papers are currently addressed to the use of the Shapley value-based approach to improve explaianbility of AI applications. As discussed for instance by Bussmann et al. (2020), the Shapley values can be employed by resorting to the SHAP (SHapley Additive exPlanations) computational framework. This procedure differs from the GAM (Generalized Additive Models) developed by Lou et al. (2012), where the model is decomposed into linear combinations of simple models, trained by a single explanatory variable, instead of decomposed into linear combinations of all the model configurations, trained by all the possible combinations of the available explanatory variables. In order to extend Shapley values to the global setting, Song et al. (2016) provided a global decomposition based on the (euclidean) variance decomposition.

In a recent research paper, Giudici and Raffinetti (2021b) suggested to combine the interpretability power of the local Shapley value approach, with a more robust global approach. To this aim, the Shapley value game theoretic approach was applied to the Lorenz Zonoid model accuracy tool, proposed by Giudici and Raffinetti (2020). The main contributions of their work are, in summary: a) the introduction of a novel global explainable AI framework, based on the combination of Lorenz Zonoids with the Shapley value approach; b) the mathematical derivation of the exact expression of a novel Shapley-Lorenz decomposition,

that can explain any machine learning model in terms of the contribution of each explanatory variable to the Lorenz Zonoid predictive accuracy.

It is worth mentioning that both the Shapley values and Shapley-Lorenz values are computationally intensive, especially when dealing with huge datasets composed of a wide set of predictors. With the aim of reducing the time-consuming and the computational effort, a more parsimonious model, able to ensure a satisfactory degree of predictive accuracy, should be selected. It follows that, besides the interpretability requirement, a further important challenge for machine learning methods is the construction of predictive accuracy tools that can evaluate and monitor the quality of the predictions. For a review, see for example Hand and Till (2001); Gneiting (2011); Kang et al. (2021); Petropoulos et al. (2022) and the references therein.

The traditional paradigm compares statistical models within a model selection procedure, in which a model is chosen through a sequence of pairwise comparisons, based on the comparison of the likelihoods (or of the posterior probabilities) of the models being compared. These criteria are to be preferred to those measures, like the Brier's score (see, e.g. Brier 1950), that do not penalise forecasts, which predict a zero probability, strongly enough when they are wrong. This may lead to conclusions which appear opposite to intuition (see, e.g. Redelmeier et al. 1991).

Nevertheless, the same likelihood-based criteria are not generally applicable, when an underlying probabilistic model is not specified, as in neural networks and random forest models.

These considerations suggest that classical model comparison is not sufficient to compare the models that can be learned from the data. Indeed, the last few years have witnessed the growing importance of model comparison methods based on the comparison between the predicted and the actually observed cases, typically within cross-validation methods. In cross-validation, the data is split in two sets, with a "training" set, used to fit a model, and a "test" set, used to compare the predictions made by the fitted model with the actual observed values.

Our aim is to compare different models, in terms of predictive accuracy. For example, suppose we would like to build an AI tool which allows an insurer to predict, on the basis of all available data, the premium amount an individual has to pay. In this case, the response variable, corresponding to the premium amount to be paid, can be expressed on a continuous scale and the reliability of the predictions can be evaluated through the Root Mean Squared Error (RMSE) measure. But the insurer can also decide to rely on the prediction of whether tomorrow the premium is above or below a certain threshold and evaluate the reliability of the tool using the AUROC measure. How can an insurer decide which response to predict? It would be desirable if the AI itself solves this problem. Comparing the p-values is not a solution, as they depend on two different models. It is necessary to develop a more general predictive accuracy measure that is model agnostic, not only with respect to the type of model - function of the explanatory variables - to employ, but also with respect to the type of response variable to be predicted.

In this paper, we contribute to solve the problem proposing a new predictive accuracy measure, called Rank Graduation Accuracy (*RGA*), which is based on the ranks and that generalises the predictive accuracy problem to all ordered variable scales. The *RGA* measure evaluates point predictions in terms of their ranks, rather than in terms of their values, gaining robustness.

We remark that the *RGA* is appropriate when the aim of the research is to determine the ordering of the observed response variable values (whether binary, ordinal or continuous), induced by the corresponding predicted values generated by the model. Thus, the *RGA*

measure differs from the Lorenz Zonoid tool described in Giudici and Raffinetti (2020), where instead the predicted values of the model are directly used for the assessment of the related predictive accuracy.

In addition, a further key-benefit derived from the *RGA* measure employment is related to the possibility of selecting less complex models, then fulfilling the parsimony principle, reducing the computational effort typically requested by the local and global explanation-based approaches and allowing to only involve those predictors which provide a significant explainability of the target variable.

The remainder of the paper is organized as follows. Section 2 illustrates, as a background, the notion of the *C* concordance curve; Section 3 introduces our main proposal, the *RGA* measure and its properties; Section 4 proposes a significance test for the *RGA* measure; Section 5 presents a simulation study aimed at showing the robustness of the *RGA* measure for real-valued predictions; Section 6 introduces the application of the proposed rank accuracy measure to an insurance problem that concerns the risks arising from crypto asset prices; Section 7 concludes with a final discussion.

2 The predictive role of the *C* concordance curve

Let D be the available data, a matrix with $h+1$ columns, corresponding to h explanatory variables and a response variable; $N = n^* + n$ rows, corresponding to all the joint observations of Y and $X_1, X_2, \dots, X_h$, partitioned into a training set D_{train}, of dimension $n^* \times (h+1)$, from which the unknown parameters of a statistical model can be estimated; and a test set D_{test}, of dimension $n \times (h+1)$, which can be used to obtain the n-dimensional vector $\hat{y}$ of the predicted values whose distance from the n observed values y will measure the predictive accuracy of the model.

When the Y variable is at least ordinal: continuous, ordered categorical, or binary, the Y values can be used to build the Lorenz curve (see e.g., Lorenz 1905), L_Y, arranging the Y values in a non-decreasing sense. More formally, for $i = 1, \dots, n$, the Lorenz curve is defined by the pair: $(i/n, \sum_{j=1}^{i} y_{r_j}/(n\bar{y}))$, where r_j indicates the non-decreasing ranks of Y and $\bar{y}$ indicates the mean of Y.

The same Y values can also be used to build the dual Lorenz curve, $L_Y^{'}$, ordering the Y values in a non-increasing sense. More formally, for $i = 1, \dots, n$, the dual Lorenz curve is defined by the pair: $(i/n, \sum_{j=1}^{i} y_{r_{n+1-j}}/(n\bar{y}))$, where r_{n+1-j} indicates the non-increasing ranks of Y.

A similar reasoning can be employed to order the predicted values $\hat{Y}$. Let $\hat{r}_i$, for $i = 1, \dots, n$, indicate the non-decreasing ranks of $\hat{Y}$. Giudici and Raffinetti (2011) suggested to build a C concordance curve by ordering the Y values not in terms of their ranks, but with respect to $\hat{r}_i$, the ranks of the predicted $\hat{Y}$ values. Formally, for $i = 1, \dots, n$, the concordance curve is defined by the pairs: $(i/n, \sum_{j=1}^{i} y_{\hat{r}_j}/(n\bar{y}))$, where $\hat{r}_i$ indicates the non-decreasing ranks of $\hat{Y}$.

To visually describe the concordance curve, Figure 1 reports, for a given test set D_{test}, the Lorenz curve, the dual Lorenz curve and the C concordance curve, together with the 45-degree line. From Fig. 1, note that the Lorenz curve and its dual are symmetric around the 45-degree line, and that the concordance curve lies between them (as shown in Raffinetti and Giudici 2012). When $\hat{r}_i = r_i$, for all $i = 1, \dots, n$, we have a perfect concordance: the concordance curve is equal to the Lorenz curve. When $\hat{r}_i = r_{n+1-i}$, for all $i = 1, \dots, n$, we have perfect discordance: the concordance curve is equal to the

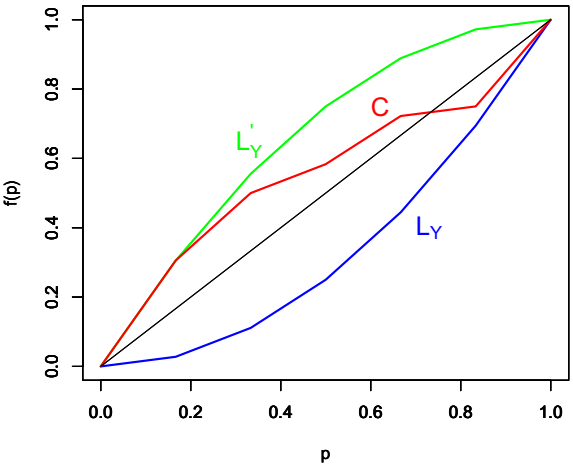

Fig. 1 The L_Y and $L_Y^{'}$ Lorenz curves and the C concordance curve, where p (on the x-axis) and $f(p)$ (on the y-axis) are the cumulative values of the x and y coordinates of the L_Y, $L_Y^{'}$ and C curves

dual Lorenz curve. In general, for any given point, the distance between the concordance curve and the Lorenz curve reveals how the rank of the predicted value differs from that of the best case, which is equal to the rank of the observed value. And, for any given point, the distance between the concordance curve and the dual Lorenz curve reveals how the rank of the predicted value differs from that of the worst case, which is equal to the rank of the inversely ordered value.

The number of points on which the C curve in Fig. 1 is constructed is equal to the number of observations n. When the response variable is continuous, the observed and predicted values can take all possible real values. When the response variable is ordinal, Y and $\hat{Y}$ can be replaced by the corresponding ranks R and $\hat{R}$, as illustrated in Giudici and Raffinetti (2021a). When the response variable is binary, taking one of two possible outcomes, corresponding to the presence ($Y = 1$) or the absence ($Y = 0$) of an attribute of interest, the predicted values take all possible real values in the interval [0, 1], which estimate the probability that $Y = 1$.

The C curve is a graphical plot of the predictive accuracy of the model predictions $\hat{y}_i \in \mathbb{R}$ for an ordered response $y_i \in \mathbb{R}$, for $i = 1, \ldots, n$. The C curve is obtained joining n points, which correspond to the observed values, ordered by the non-decreasing magnitude of the predictions.

Before moving to a summary measure, it is useful to associate the C curve behaviour with the main reference scenarios that occur in model comparison: the best case: a perfectly concordant model; the worst case: a perfectly discordant model; the random case, in which predictions are generated randomly and, finally, a generic "intermediate" case.

It results that:

(i) the best case occurs when the ordering of the Y response variable values corresponds to the ordering of the predicted values, with the C curve perfectly overlapping the Lorenz curve L_Y;
(ii) the worst case occurs when the ordering of the Y response variable values is in inverse correspondence with the ordering of the predicted values, with the C curve perfectly overlapping the dual Lorenz curve $L_Y^{'}$;
(iii) in the random case, the C curve overlaps the 45-degree line;

(iv) in the generic case, the C curve lies in the area between the Y response variable Lorenz curve, L_Y, and its dual, L_Y'. The distance between C and the 45-degree line measures how a model improves over random predictions.

3 The Rank Graduation Accuracy measure (*RGA*)

Drawing on item iv) of the last section, a summary measure for the C curve of a model could be obtained considering the area between the dual Lorenz curve and the concordance curve, and dividing it by its maximum possible value: the area between the dual Lorenz curve and the Lorenz curve.

More formally, we define a Rank Graduation Accuracy (*RGA*) measure with the following expression:

$$RGA = \frac{\sum_{i=1}^{n} \left\{ \frac{1}{n\bar{y}} \left(\sum_{j=1}^{i} y_{r_{n+1-j}} - \sum_{j=1}^{i} y_{\hat{r}_j} \right) \right\}}{\sum_{i=1}^{n} \left\{ \frac{1}{n\bar{y}} \left(\sum_{j=1}^{i} y_{r_{n+1-j}} - \sum_{j=1}^{i} y_{r_j} \right) \right\}}. \tag{1}$$

Remark 1 When tied predictions occur, it may be unclear how to order the observed values in the expression of *RGA*. In this case, we suggest to follow Ferrari and Raffinetti (2015), who proposed to replace the observed response values corresponding to the same predictions with their mean values.

We now present some important properties of the *RGA* measure.

Property 1 - *Simplification*

Through some algebraic manipulations, the RGA measure can be simplified as follows:

$$RGA = \frac{\sum_{i=1}^{n} iy_{\hat{r}_i} - \sum_{i=1}^{n} iy_{r_{n+1-i}}}{\sum_{i=1}^{n} iy_{r_i} - \sum_{i=1}^{n} iy_{r_{n+1-i}}}. \tag{2}$$

The proof of Property 1 is reported in the Appendix.

Property 2 - *Normalisation*

In general, $0 \leq RGA \leq 1$, *with* $RGA = 1$ *in the best case of a perfectly concordant model*; $RGA = 0$ *in the worst case of a perfectly discordant model*; $RGA = 0.5$ *in the case of random predictions*.

4 A significance test for the *RGA* measure

To evaluate whether the *RGA* of a model significantly differs from that of another model, a statistical test is necessary. To this aim, the *RGA* measure can be expressed in terms of covariance operators, as in the following Proposition.

Proposition 1 *When the response variable is continuous*:

$$RGA = \frac{cov(Y_{r(\hat{Y})}, F(Y)) + cov(Y, F(Y))}{2cov(Y, F(Y))}, \tag{3}$$

where $Y_{r(\hat{Y})}$ represents the Y variable re-ordered according to the ranks of the corresponding predictions $\hat{Y}$ and F is the cumulative continuous distribution function of Y.

The proof of Proposition 1 is reported in the Appendix.

Note that, through some mathematical computations, Eq. (3) can be re-expressed as

$$RGA = \frac{1}{2}\frac{cov(Y_{r(\hat{Y})}, F(Y))}{cov(Y, F(Y))} + \frac{1}{2} \tag{4}$$

and, therefore, the *RGA* is a linear function of the ratio:

$$\psi(Y, \hat{Y}) = cov(Y_{r(\hat{Y})}, F(Y))/cov(Y, F(Y)). \tag{5}$$

The ratio in (5) was originally introduced by Schechtman and Yitzhaki (1987) to evaluate the correlation between total income and its sources. Here we employ it to derive statistics for the *RGA* measure. Intuitively, the smaller $\psi(Y, \hat{Y})$, the greater the distance between a model and the best case.

More formally, given two alternative models (Mod_1 and Mod_2), the statistic in (5) will be used to test following hypotheses:

$$H_0 : \psi(Y, \hat{Y}_{Mod_1}) = \psi(Y, \hat{Y}_{Mod_2}) \quad \text{vs} \quad H_1 : \psi(Y, \hat{Y}_{Mod_1}) \neq \psi(Y, \hat{Y}_{Mod_2}), \tag{6}$$

where $\psi(Y, \hat{Y}_{Mod_1}) = cov(Y_{r(\hat{Y}_{Mod_1})}, F(Y))/cov(Y, F(Y))$ and $\psi(Y, \hat{Y}_{Mod_2}) = cov(Y_{r(\hat{Y}_{Mod_2})}, F(Y))/cov(Y, F(Y))$ are functions that derive from the application of (5), respectively to RGA_{Mod_1} and RGA_{Mod_2}. To derive a test statistic for the hypotheses in (6), note that the estimator of $\psi(Y, \hat{Y}_{Mod_1})$ can be expressed as a function of two dependent U-statistics, denoted with U_1 and U_2:

$$\hat{\psi}(Y, \hat{Y}_{Mod_1}) = \frac{U_1}{U_2} = \frac{\frac{1}{4\binom{n}{2}} \sum_{i=1}^{n} (2i - 1 - n) Y_{r(\hat{Y}_{i_{Mod_1}})}}{\frac{1}{4\binom{n}{2}} \sum_{i=1}^{n} (2i - 1 - n) Y_{r(Y_i)}}. \tag{7}$$

Similarly, the estimator of $\psi(Y, \hat{Y}_{Mod_2})$ can be defined as a function of two dependent U-statistics, U_3 and U_2:

$$\hat{\psi}(Y, \hat{Y}_{Mod_2}) = \frac{U_3}{U_2} = \frac{\frac{1}{4\binom{n}{2}} \sum_{i=1}^{n} (2i - 1 - n) Y_{r(\hat{Y}_{i_{Mod_2}})}}{\frac{1}{4\binom{n}{2}} \sum_{i=1}^{n} (2i - 1 - n) Y_{r(Y_i)}}. \tag{8}$$

It follows that $\delta = \psi(Y, \hat{Y}_{Mod_1}) - \psi(Y, \hat{Y}_{Mod_2})$ can be estimated by $\hat{\delta}$, a function of three dependent U-statistics:

$$\hat{\delta} = \hat{\psi}(Y, \hat{Y}_{Mod_1}) - \hat{\psi}(Y, \hat{Y}_{Mod_2}) = \frac{U_1}{U_2} - \frac{U_3}{U_2}. \tag{9}$$

According to Hoeffding (1948), a function of several dependent U-statistics has a normal distribution, provided that the sample size is large enough. Thus, the estimator in Eq. (9) has a limiting normal distribution, whose variance $Var(\hat{\delta})$ can be estimated by means of the Jackknife method (see e.g., Efron and Stein 1981) each time omitting the pairs $(Y, \hat{Y}_{Mod_1})$ and $(Y, \hat{Y}_{Mod_2})$.

Therefore, the test statistic for testing the null hypothesis $H_0 : \psi(Y, \hat{Y}_{Mod_1}) = \psi(Y, \hat{Y}_{Mod_2})$ becomes:

$$Z = \frac{\hat{\delta}}{\sqrt{\widehat{Var(\hat{\delta})}}} \rightarrow N(0, 1), \tag{10}$$

where $\widehat{Var(\hat{\delta})} = \frac{n-1}{n} \sum_{i=1}^{n} (\hat{\delta}_{(-i)} - \bar{\delta})^2$, where $\hat{\delta}_{(-i)}$ are the values of $\hat{\delta}$ by omitting one pair $(Y, \hat{Y})$ at a time and where $\bar{\delta}$ is the average of the values $\hat{\delta}_{(-i)}$, for $i = 1, \ldots, n$.

For a fixed significance level α, a rejection region for the test corresponds to the region $|Z| \geq z_{\alpha/2}$. If the test statistic falls in this region, Mod_1 and Mod_2 are significantly different from each other.

Remark 2 It is worth noting that the proposed test can be extended, without loss of generality, to all types of ordinal variables. The continuity constraint of the joint distribution can be preserved replacing tied observations with their mean value. This adjustment gives rise to a continuous variable which, together with $\hat{Y}$, provides a continuous joint distribution, satisfying the assumptions in Proposition 1.

5 Robustness of the *RGA* measure: a simulation study

It is important that the measurement of predictive accuracy is not affected by outlying observations, which may bias model comparison. For this reason, in this section we aim at assessing the robustness of the *RGA* measure by means of a simulation study.

Without loss of generality, let X and Z be two independent continuous random variables with $X \sim U(0, 10)$ and $Z \sim N(0, 1)$ and let $Y = 5 + 3X + Z$, from which we can simulate a set of observations.

To assess robustness, we replace the obtained left and right tail observations of the X distribution with outliers in the tails of the distribution. Without loss of generality, we consider six alternative replacements, as follows:

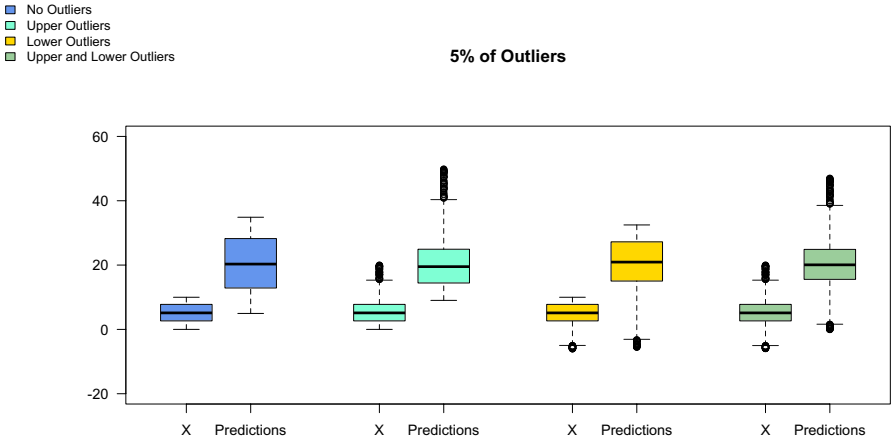

Fig. 2 Distribution of X and $\hat{Y}$ for scenarios: a) upper 5% outliers; b) lower 5% outliers; c) upper and lower 5% outliers; and in the case of no outliers

(a) the observations greater than the 95% percentile are replaced by observations sampled from a $U(15, 20)$ distribution;
(b) the observations lower than the 5% percentile are replaced by observations sampled from a $U(-10, -5)$ distribution;
(c) the observations greater than the 95% percentile are replaced by observations sampled from a $U(15, 20)$ distribution and the observations lower than the 5% percentile are replaced by observations sampled from a $U(-10, -5)$ distribution;
(d) the observations greater than the 90% percentile are replaced by observations sampled from a $U(15, 20)$ distribution;
(e) the observations lower than the 10% percentile are replaced by observations sampled from a $U(-10, -5)$ distribution;
(f) the observations greater than the 90% percentile are replaced by observations sampled from a $U(15, 20)$ distribution and the observations lower than the 10% percentile are replaced by observations sampled from a $U(-10, -5)$ distribution.

For each of the six modified samples of X, we estimate a linear model and evaluate its predictive accuracy, using the same training and test samples.

The resulting distributions of the predicted Y values, along with that of the sampled X variable are plotted in Fig. 2 (scenarios (a), (b), (c)) and 3 (scenarios (d), (e), (f)). For comparison, we also report, to the left of both the figures, the distribution of X and of the predictions when no outliers replace the originally sampled observations.

We then proceed to calculate the RMSE and the *RGA* for all scenarios, whose distributions are presented in Fig. 2 and 3 . The results are reported in Table 1.

From Table 1, note that while the RMSE increases its value with respect to the case of no outliers, highlighting a decreased predictive accuracy, the *RGA* preserves its value, showing a superior robustness to the presence of outliers.

The results in Table 1 allow to conclude that, when real-valued forecasts are considered, the *RGA* measure is more robust than the RMSE, the standard measure in most predictive applications. A similar result can be obtained in the case of ordered categorical variables, replacing continuous measurement with the corresponding ranks, and binary variables.

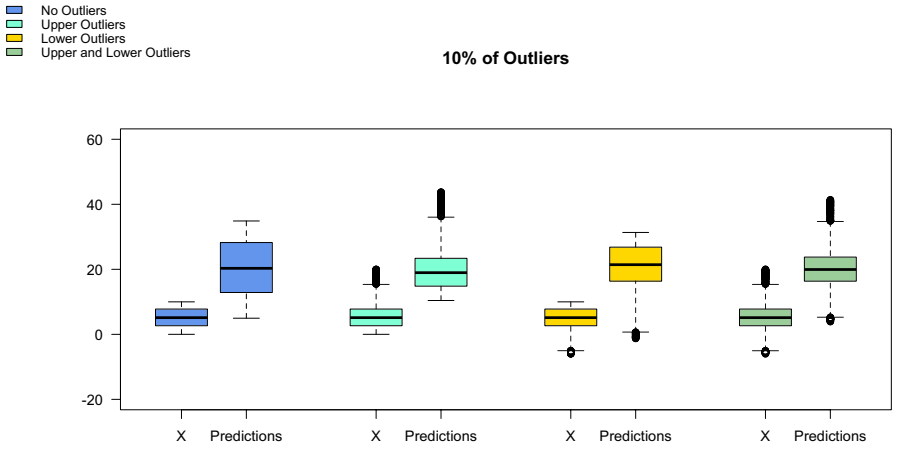

Fig. 3 Distribution of X and $\hat{Y}$ for scenarios: d) upper 10% outliers; e) lower 10% outliers; f) upper and lower 10% outliers; and in the case of no outliers

6 Application: financial risk

Cryptocurrencies have rapidly gained great popularity among investors, companies and regulators (see e.g., Angerer et al. 2021). Nevertheless, criticisms about the evolution of cryptocurrencies arise in relation to the management of the deriving risks, such as volatility and liquidity, together with the risk of cybercrime, and to the measurement of their effects on society (see e.g, Feng et al. 2018). In particular, the majority of crypto assets are underinsured or uninsurable according to the existing insurance standards, as there is no insurance deposit for this asset class, implying a lack in terms of investor security. Precisely in this perspective, the development of statistical tools to appropriately measure the accuracy of machine learning models emerges as a crucial issue in the risk management domain.

In this section we apply our proposal to the measurement of the financial risks generated by crypto assets. We will show how to employ the *RGA* measure and the associated statistical test to compare the accuracy of forecasts of bitcoin prices on a rolling horizon basis.

The data consist of several time series of financial prices. Among them, the series of the daily bitcoin prices in the Coinbase exchange, from 18 May 2016 to 30 April 2018, will be used as a response variable to be predicted. Whereas the daily prices of classical assets, such as oil, gold and sp500, together with the exchange rates (dollar/yuan and dollar/euro), for the same period of time, will be considered as candidate predictors.

An overview of the main summary statistics of the used data are reported in Table 2.

Our aim is to compare the model selection performance of the *RGA* against that of the RMSE. To this aim, we consider four different monthly time windows of bitcoin prices to be forecast: from January 2018 to April 2018. For each of them, we compare alternative linear regression models based on the time series of the considered five predictors for the previous year.

Specifically, for each possible model dimension, ranging from 1 to 5, we choose the best model using the Akaike Information Criterion (AIC), in order to detect the five candidate best models. The selected models are specified in Table 3.

Table 1 Predictive accuracy under different outlier configurations

Predictive accuracy measures	*RMSE*	*RGA*
Without outliers	***0.981***	***0.997***
Scenario a) (upper 5% outliers)	3.769	0.997
Scenario b) (lower 5% outliers)	3.695	0.997
Scenario c) (upper and lower 5% outliers)	4.119	0.997
Scenario d) (upper 10% outliers)	4.163	0.996
Scenario e) (lower 10% outliers)	4.018	0.996
Scenario f) (upper and lower 10% outliers)	4.027	0.996

Table 2 Summary statistics for bitcoin prices, classic asset prices and exchange rates: mean value; standard deviations (*sd*); coefficient of variation (*cv*); minimum and maximum values

Variables	Mean	*sd*	*cv*	Min.	Max.
bitcoin	3919.10	4318.98	1.10	438.38	19650.01
sp500	2399.17	212.31	0.09	2000.54	2872.87
gold	1275.58	52.34	0.04	1128.42	1366.38
oil	49.36	3.37	0.07	39.51	57.20
exchange rate dollar/euro	0.88	0.04	0.05	0.80	0.96
exchange rate dollar/yuan	6.68	0.19	0.03	6.27	6.96

Table 3 Results from the stepwise model selection on the daily time series data from May 2016 to April 2018

Model	**Variables**
Model 1	sp500
Model 2	sp500, exchange rate dollar/yuan
Model 3	sp500, gold, oil
Model 4	sp500, gold, oil, exchange rate dollar/euro
Model 5	sp500, gold, oil, exchange rate dollar/euro, exchange rate dollar/yuan

To predict bitcoin prices referred to the first term of the year 2018, we follow a rolling window procedure. Specifically, the model is trained on rolling windows including data related to year 2017. To do this, we first trained a model with a sliding window of 1 year (from 1st January 2017 to 31 December 2017) and then we predicted over the next month (January 2018). Then, we shifted 1 month, re-trained the model and predicted the next month (February 2018) and so on.

The rolling window procedure can be summarised as follows:

- models are trained using only data between 1st January 2017 and 31 December 2017. Forecasts are derived for a time window (first window) that starts on 1st January 2018 and ends at 31 January 2018;
- models are trained using only data between 1st February 2017 and 31 January 2018. Forecasts are derived for a time window (second window) that starts on 1st February 2018 and ends at 28 February 2018;

- models are trained using only data between 1st March 2017 and 28 February 2018. Forecasts are derived for a time window (third window) that starts on 1st March 2018 and ends at 31 March 2018;
- models are trained using only data between 1st April 2017 and 31 March 2018. Forecasts are derived for a time window (fourth window) that starts on 1st April 2018 and ends at 30 April 2018.

For each time window, the *RGA* and the RMSE are computed and the related results are displayed in Fig. 4. To understand the difficulty of the forecasting exercise, the standard deviations (*sd*) of the bitcoin prices to be predicted in the four time windows are provided in the caption of the figure. Moreover, with the aim of improving the readability of Fig. 4, in Table 4 the RMSE and *RGA* values are reported for each time window.

The first time window in Fig. 4 is the most difficult to predict, as it has the largest standard deviation of bitcoin prices. Consistently with the difficulty, neither model reaches a good predictive accuracy: the *RGA* takes low values, and the RMSE takes high values, in comparison with the other time windows. Specifically, the *RGA* selects Model 4 as the best model, while the RMSE chooses Model 2. To check whether the dimension of the model selected according to the *RGA* can be reduced, we have applied the *RGA* test to evaluate whether the difference in predictive accuracy between Model 2 and Model 4 is significant. The resulting p-value is largely higher than 5%, leading to choose Model 2 also with the *RGA* measure.

The *RGA* and RMSE model selection differ (also for the second and third time window). The highest value for the *RGA* is achieved by Model 4, while the lowest RMSE is reached by Model 2. However, also for these windows, the *RGA* based test does not reject the simplification to Model 2.

Finally, for the fourth time window, the easiest to predict as it has the smallest deviation of bitcoin prices, the *RGA* and RMSE select the same model (Model 2).

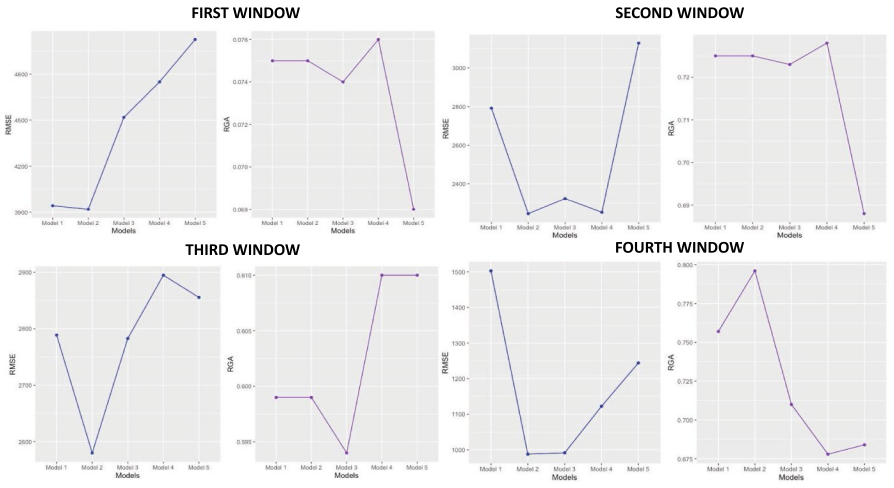

Fig. 4 RMSE and *RGA* behaviours across the four considered windows. For each time window, the standard deviation of the prices to be predicted are as follows. First window: $sd = 2046.594$; second window: $sd = 1159.363$; third window: $sd = 1290.287$; fourth window: $sd = 986.027$

Table 4 Predictive accuracy across different time windows (*RGA* against RMSE)

Models	RMSE	*RGA*
First windows		
Model 1	3942.024	0.075
Model 2	3919.056	0.075
Model 3	4517.832	0.074
Model 4	4748.649	0.076
Model 5	5024.660	0.068
Second windows		
Model 1	2791.945	0.725
Model 2	2245.867	0.725
Model 3	2323.669	0.723
Model 4	2252.900	0.728
Model 5	3128.814	0.688
Third windows		
Model 1	2788.690	0.599
Model 2	2579.766	0.599
Model 3	2782.661	0.594
Model 4	2894.766	0.610
Model 5	2855.433	0.610
Fourth windows		
Model 1	1502.564	0.757
Model 2	987.583	0.796
Model 3	991.221	0.710
Model 4	1122.098	0.678
Model 5	1243.969	0.684

The previous considerations can be summarised by a general finding: the *RGA* always selects more complex models than the RMSE, and it especially does so for the most volatile responses. This behaviour is correlated with the higher robustness of the *RGA*, less affected by outliers in the predictor variable.

While the *RGA* leads to select more complex models than the RMSE, the application of the *RGA* test allows to simplify the chosen model to the same dimension chosen with the RMSE.

To better strength the *RGA* role to the evaluation of predictive accuracy, we suppose that bitcoin prices are only available on an ordinal scale based on five categories encoded by 1, 2, 3, 4, and 5. Following Giudici and Raffinetti (2021a), we deal with the ordinal nature of the response by resorting to the rank regression model. Formally, the generic variable Y assuming M ordered categories is transformed into ranks by assigning rank $r_1 = 1$ to the smallest ordered category of Y and rank $r_m = r_{m-1} + n_{m-1}$ to the m-th ordered category of Y (where $m = 1, \ldots, M$; r_{m-1} is the rank related to the $(m-1)$-*th* ordered category of Y and n_{m-1} is the associated frequency).

Given the ordinal nature of the response variable, the analysis is led without applying a rolling window procedure which instead is recommended to adjust the forecast in order to accommodate recent changes or trends.

For the sake of simplicity, the model is trained on data referred to year 2017 and tested on data referred to year 2018. As well as for the bitcoin price prediction, we select the best

Table 5 Results from the stepwise model selection on ordered categorised bitcoin prices from May 2016 to April 2018

Model	Variables
Model 1	sp500
Model 2	sp500, exchange rate dollar/yuan
Model 3	sp500, exchange rate dollar/yuan, oil
Model 4	sp500, exchange rate dollar/yuan, oil, gold
Model 5	sp500, exchange rate dollar/yuan, oil, gold, exchange rate dollar/euro

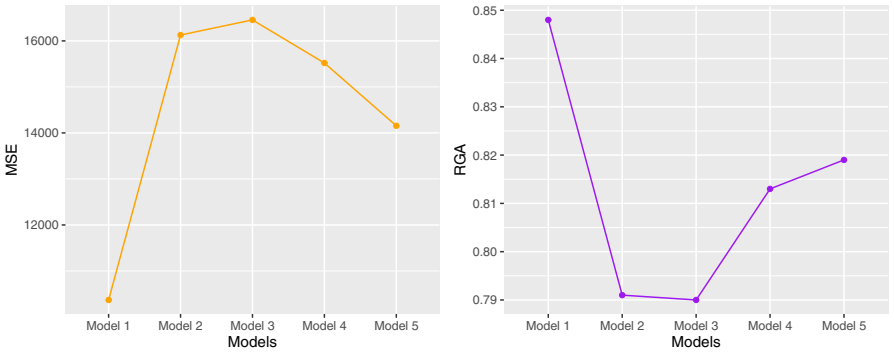

Fig. 5 MSE and *RGA* behaviours

model configurations using the Akaike Information Criterion (AIC). The resulting models are specified in Table 5.

Gaudette and Japkowicz (2009) compared several metrics to assess the ordinal classification accuracy. They showed that both the RMSE and MSE (Mean Squared Error) perform well for ordinal data converted to small integers. Therefore, we consider the MSE as a competitor of the *RGA* and we graphically report in Fig. 5 their behaviours across the different model dimensions together with the corresponding values in Table 6.

From Fig. 5, it results that the best model (characterised by the highest *RGA* and the lowest MSE) is the simplest model with only the explanatory variable sp500. For more complex model configurations, the predictive accuracy generally worsens, both in terms of *RGA* and MSE. An improvement is provided by the full model, although the *RGA* and MSE values are smaller and higher than those associated with the simplest model. It follows that *RGA* and MSE are coherent in selecting the most accurate model.

Beside the issue of the bitcoin price prediction, also the forecast of the returns derived from cryptocurrencies becomes a crucial topic, especially for those investors who are interested in measuring the potential gains or losses. In order to cover this

Table 6 Predictive accuracy (*RGA* against MSE)

Models	MSE	*RGA*
Model 1	10369.55	0.848
Model 2	16127.49	0.791
Model 3	16456.40	0.790
Model 4	15520.03	0.813
Model 5	14155.12	0.819

Table 7 Results from the stepwise model selection on gains or losses from May 2016 to April 2018

Model	Variables
Model 1	Gold
Model 2	Gold, exchange rate dollar/euro
Model 3	Gold, exchange rate dollar/euro, oil
Model 4	Gold, exchange rate dollar/euro, oil, sp500
Model 5	Gold, exchange rate dollar/euro, oil, sp500, exchange rate dollar/yuan

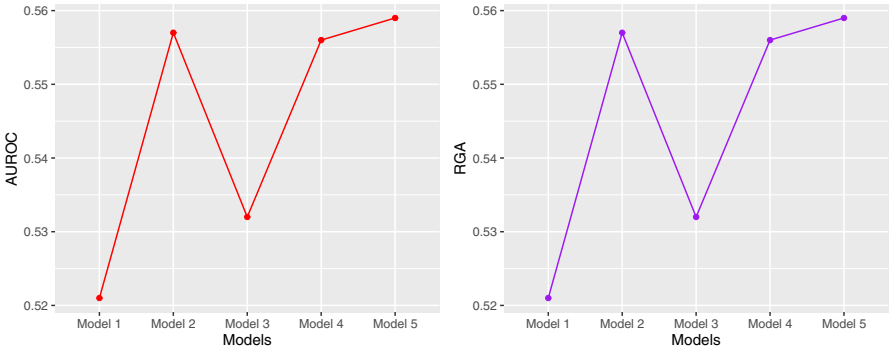

Fig. 6 AUROC and *RGA* behaviours

perspective, we first transform the bitcoin prices into returns and then we proceed to their binarisation by assigning value equal to 1 to the negative returns (losses) and value equal to 0 to the positive returns (gains).

Given the binary nature of the target variable, our purpose is to compare the model selection performance of the *RGA* against that of the AUROC. As well as for the case of the ordinal response, the analysis can be simplified without following a rolling window procedure.

A logistic regression model is trained on data referred to year 2017 and tested on data referred to year 2018 and the best model configurations are determined by using the Akaike Information Criterion (AIC). The chosen models are specified in Table 7.

To evaluate the predictive accuracy, the *RGA* and AUROC behaviours across the different model dimensions are depicted in Fig. 6 and the related values are provided in Table 8.

From both Fig. 6 and Table 8, it arises that the *RGA* and AUROC show the same performance. Specifically, the logistic regression model does not achieve a good predictive accuracy degree, being the values associated with the *RGA* and AUROC very close to 0.5. Thus, they almost perform as well as a random model. This because although the logistic regression model is characterised by a "white-box" that allows an easy interpretability of the results, it provides a limited predictive accuracy. Resorting to more complex machine learning models, such as neural network and random forest models, would on one side improve the accuracy of predictions but on the other side worsen the interpretability of the results.

Table 8 Predictive accuracy (*RGA* against AUROC)

Models	AUROC	*RGA*
Model 1	0.521	0.521
Model 2	0.557	0.557
Model 3	0.532	0.532
Model 4	0.556	0.556
Model 5	0.559	0.559

7 Discussion

To improve the compliance of AI applications in finance and insurance, in the paper we have proposed a new tool to evaluate the predictive accuracy and robustness of a machine learning model: the *RGA*.

The adoption of the *RGA* measure can assess, monitor and improve the quality of machine learning models in terms of their accuracy (difference between the observed and the predicted values) and their robustness (stability of the estimated model with respect to variations in the data).

By means of a simulation study, and the application of the measure to a crypto asset dataset, we have shown that the proposed measure is quite consistent in model choice and also more robust than classical predictive accuracy measures, such as the RMSE. The *RGA* can thus be proposed as an additional toolkit for machine learning.

From a methodological viewpoint, the *RGA* measure provides a rather general accuracy statistic, applicable in the same manner to all ordered response variables. It is preferable to other measures when the aim is to predict the correct ordering of a point response, regardless of whether such response is binary, ordinal or continuous.

Future research extensions should consider the case of a multivariate response variable, which would require to generalise the concordance curve to a multidimensional setting.

From the applied side, the proposed measure should be employed to other comparison settings and, in particular, to those involving applications of machine learning models.

Appendix

In this appendix, the proof of Property 1 is reported together with the proof of Proposition 1.

Proof (*Property 1*) We prove that Eq. (1), can be re-written as in Eq. (2).

Based on Eq. (1), it results that:

$$RGA = \frac{\sum_{i=1}^{n} \left\{ \frac{1}{n\bar{y}} \left(\sum_{j=1}^{i} y_{r_{n+1-j}} - \sum_{j=1}^{i} y_{\hat{r}_j} \right) \right\}}{\sum_{i=1}^{n} \left\{ \frac{1}{n\bar{y}} \left(\sum_{j=1}^{i} y_{r_{n+1-j}} - \sum_{j=1}^{i} y_{r_j} \right) \right\}},$$

which can be re-expressed as

$$RGA = \frac{\sum_{i=1}^{n}\left\{\sum_{j=1}^{i} y_{r_{n+1-j}} - \sum_{j=1}^{i} y_{\hat{r}_j}\right\}}{\sum_{i=1}^{n}\left\{\sum_{j=1}^{i} y_{r_{n+1-j}} - \sum_{j=1}^{i} y_{r_j}\right\}} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{i} y_{r_{n+1-j}} - \sum_{i=1}^{n}\sum_{j=1}^{i} y_{\hat{r}_j}}{\sum_{i=1}^{n}\sum_{j=1}^{i} y_{r_{n+1-j}} - \sum_{i=1}^{n}\sum_{j=1}^{i} y_{r_j}}. \quad (11)$$

As $\sum_{i=1}^{n}\sum_{j=1}^{i} y_{r_{n+1-j}} = n(n+1)\bar{y} - \sum_{i=1}^{n} i y_{r_{n+1-i}}$, $\sum_{i=1}^{n}\sum_{j=1}^{i} y_{\hat{r}_j} = n(n+1)\bar{y} - \sum_{i=1}^{n} i y_{\hat{r}_i}$ and $\sum_{i=1}^{n}\sum_{j=1}^{i} y_{r_j} = n(n+1)\bar{y} - \sum_{i=1}^{n} i y_{r_i}$ (see e.g., Marshall et al. 2011), Eq. (11) becomes

$$\begin{aligned} RGA &= \frac{n(n+1)\bar{y} - \sum_{i=1}^{n} i y_{r_{n+1-i}} - n(n+1)\bar{y} + \sum_{i=1}^{n} i y_{\hat{r}_i}}{n(n+1)\bar{y} - \sum_{i=1}^{n} i y_{r_{n+1-i}} - n(n+1)\bar{y} + \sum_{i=1}^{n} i y_{r_i}} \\ &= \frac{\sum_{i=1}^{n} i y_{\hat{r}_i} - \sum_{i=1}^{n} i y_{r_{n+1-i}}}{\sum_{i=1}^{n} i y_{r_i} - \sum_{i=1}^{n} i y_{r_{n+1-i}}}. \end{aligned}$$

□

Proof (*Proposition 1*) To prove Proposition 1, we first show that the *RGA* can be written, through the covariance formulation, as in Eq. (3).

By exploiting Property 1, Eq. (1) can be re-written as

$$RGA = \frac{\frac{1}{n\bar{y}}\sum_{i=1}^{n} i y_{\hat{r}_i} - \frac{1}{n\bar{y}}\sum_{i=1}^{n} i y_{r_{n+1-i}}}{\frac{1}{n\bar{y}}\sum_{i=1}^{n} i y_{r_i} - \frac{1}{n\bar{y}}\sum_{i=1}^{n} i y_{r_{n+1-i}}}. \quad (12)$$

By adding and subtracting, at both the numerator and denominator, the term $\sum_{i=1}^{n}\frac{i}{n}$, we derive that

$$RGA = \frac{\left(\frac{1}{n\bar{y}}\sum_{i=1}^{n} i y_{\hat{r}_i} - \sum_{i=1}^{n}\frac{i}{n}\right) + \left(\sum_{i=1}^{n}\frac{i}{n} - \frac{1}{n\bar{y}}\sum_{i=1}^{n} i y_{r_{n+1-i}}\right)}{\left(\frac{1}{n\bar{y}}\sum_{i=1}^{n} i y_{r_i} - \sum_{i=1}^{n}\frac{i}{n}\right) + \left(\sum_{i=1}^{n}\frac{i}{n} - \frac{1}{n\bar{y}}\sum_{i=1}^{n} i y_{r_{n+1-i}}\right)}. \quad (13)$$

As $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$, through some mathematical manipulations, Eq. (13) can then be expressed as

$$RGA = \frac{\frac{1}{\bar{y}}\left\{\frac{1}{n}\sum_{i=1}^{n} i y_{\hat{r}_i} - \frac{n+1}{2}\bar{y}\right\} - \frac{1}{\bar{y}}\left\{\frac{1}{n}\sum_{i=1}^{n} i y_{r_{n+1-i}} - \frac{n+1}{2}\bar{y}\right\}}{\frac{1}{\bar{y}}\left\{\frac{1}{n}\sum_{i=1}^{n} i y_{r_i} - \frac{n+1}{2}\bar{y}\right\} - \frac{1}{\bar{y}}\left\{\frac{1}{n}\sum_{i=1}^{n} i y_{r_{n+1-i}} - \frac{n+1}{2}\bar{y}\right\}}. \quad (14)$$

As shown by Giudici and Raffinetti (2020), $\bar{i} = \frac{(n+1)}{2}$, where $\bar{i}$ represents the mean of i intended as the rank assigned to the Y values (i.e., $i = r(y_i)$). Based on this assumption, the terms in (14) can be translated into covariance operators as

$$\frac{1}{\bar{y}}\left\{\frac{1}{n}\sum_{i=1}^{n} i y_{\hat{r}_i} - \frac{(n+1)}{2}\bar{y}\right\} = \frac{1}{\bar{y}} c\hat{o}v(Y_{r(\hat{Y})}, r(Y)), \quad (15)$$

$$\frac{1}{\bar{y}}\left\{\frac{1}{n}\sum_{i=1}^{n} i y_{r_{n+1-i}} - \frac{(n+1)}{2}\bar{y}\right\} = -\frac{1}{\bar{y}}\hat{cov}(Y, r(Y)) \tag{16}$$

and

$$\frac{1}{\bar{y}}\left\{\frac{1}{n}\sum_{i=1}^{n} i y_{r_i} - \frac{(n+1)}{2}\bar{y}\right\} = \frac{1}{\bar{y}}\hat{cov}(Y, r(Y)). \tag{17}$$

Thus, Eq. (14) is equivalent to

$$RGA = \frac{\hat{cov}(Y_{r(\hat{Y})}, r(Y)) + \hat{cov}(Y, r(Y))}{\hat{cov}(Y, r(Y)) + \hat{cov}(Y, r(Y))} = \frac{\hat{cov}(Y_{r(\hat{Y})}, r(Y)) + \hat{cov}(Y, r(Y))}{2\hat{cov}(Y, r(Y))}. \tag{18}$$

Moving from the sample to the population version, and denoting with μ the expected value of Y, $\frac{1}{\bar{y}}\hat{cov}(Y_{r(\hat{Y})}, r(Y))$, $-\frac{1}{\bar{y}}\hat{cov}(Y, r(Y))$ and $\frac{1}{\bar{y}}\hat{cov}(Y, r(Y))$ are equivalent to $\frac{1}{\mu}cov(Y_{r(\hat{Y})}, r(Y))$, $-\frac{1}{\mu}cov(Y, r(Y))$ and $\frac{1}{\mu}cov(Y, r(Y))$, respectively.

As stated by Lerman and Yitzhaki (1984), the $r(Y)/n$ terms are the empirical representation of the cumulative function $F(Y)$. It follows that, by multiplying the covariances by n, one can re-express $\frac{1}{\mu}cov(Y_{r(\hat{Y})}, r(Y))$, $-\frac{1}{\mu}cov(Y, r(Y))$ and $\frac{1}{\mu}cov(Y, r(Y))$ as $\frac{n}{\mu}cov(Y_{r(\hat{Y})}, F(Y))$, $-\frac{n}{\mu}cov(Y, F(Y))$ and $\frac{n}{\mu}cov(Y, F(Y))$, respectively.

The *RGA* measure in terms of covariance operators is finally equal to:

$$RGA = \frac{cov(Y_{r(\hat{Y})}, F(Y)) + cov(Y, F(Y))}{cov(Y, F(Y)) + cov(Y, F(Y))} = \frac{cov(Y_{r(\hat{Y})}, F(Y)) + cov(Y, F(Y))}{2cov(Y, F(Y))}. \tag{19}$$

□

Funding Open access funding provided by Università degli Studi di Pavia within the CRUI-CARE Agreement. The Author acknowledges support from the European xAIM (eXplainable Artificial Intelligence in healthcare Management) project supported by the CEF Telecom under Grant Agreement No. INEA/CEF/ICT/A2020/2276680.

Declarations

Competing Interests The Author declares she has not competing financial interests.

References

Aas, K., Jullum, M., Løland, A.: Explaining individual predictions when features are dependent: more accurate approximations to shapley values. Artif. Intell. **298**, 1–24 (2021). https://doi.org/10.1016/j.artint.2021.103502

Aldasoro, I., Gambacorta, L., Giudici, P., Leach, T.: The drivers of cyber risk. J. Financ. Stabil. **60**, 100989 (2022). https://doi.org/10.1016/j.jfs.2022.100989

Angerer, M., Hoffmann, C.H., Neitzert, F., Kraus, S.: Objective and subjective risks of investing into cryptocurrencies. Financ. Res. Lett. **40**, 101737 (2021). https://doi.org/10.1016/j.frl.2020.101737

Bracke, P., Datta, A., Jung, C., Hayak, S.: Machine learning explainability in finance: an application to default risk analysis. Staff Working Paper No. 816, Bank of England. https://www.bankofengland.co.uk/-/media/boe/files/working-paper/2019/machine-learning-explainability-in-finance-an-application-to-default-risk-analysis.pdf?la=en &hash=692E8FD8550DFBF5394A35394C00B1152DAFCC9E (2019). Accessed 26 September 2022

Brier, G.: Verification of forecasts expressed in terms of probabilities. Mon. Weather Rev. **78**, 1–3 (1950). https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Bussmann, N., Giudici, P., Marinelli, D., Papenbrock, J.: Explainable AI in credit risk management. Front. Artif. Intell. **3**, 1–5 (2020). https://doi.org/10.3389/frai.2020.00026

Ceylan, E.I.: The Effects of Artificial Intelligence on the Insurance Sector: Emergence, Applications, Challenges, and Opportunities. In: Bozkuş Kahyaoğlu, S. (eds.) The impact of artificial intelligence on governance, economics and finance Vol. 2. Accounting, finance, sustainability, governance & fraud: theory and application. Springer, Singapore (2022)

Efron, B., Stein, C.: The jackknife estimate of variance. Ann. Stat. **9**, 586–596 (1981). https://doi.org/10.1214/aos/1176345462

Eling, M., Nuessle, D., Staubli, J.: The impact of artificial intelligence along the insurance value chain and on the insurability of risks. Geneva Pap. Risk. Insur. Issues Pract. **47**, 205–241 (2022). https://doi.org/10.1057/s41288-020-00201-7

Feng, W., Wang, Y., Zhang, Z.: Can cryptocurrencies be a safe haven: a tail risk perspective analysis. Appl. Econ. **50**, 4745–4762 (2018). https://doi.org/10.1080/00036846.2018.1466993

Ferrari, P.A., Raffinetti, E.: A different approach to dependence analysis. Multivar. Behav. Res. **50**, 248–264 (2015). https://doi.org/10.1080/00273171.2014.973099

Gaudette, L., Japkowicz, N.: Evaluation Methods for Ordinal Classification. In: Gao Y., Japkowicz N. (eds) Advances in artificial intelligence, Canadian AI 2009. Lecture notes in computer science, 5549. Springer, Berlin & Heidelberg (2009)

Giudici, P., Raffinetti, E.: On the Gini measure decomposition. Stat. Probabil. Lett. **81**, 133–139 (2011). https://doi.org/10.1016/j.spl.2010.10.005

Giudici, P., Raffinetti, E.: Lorenz model selection. J. Classif. **37**, 754–768 (2020). https://doi.org/10.1007/s00357-019-09358-w

Giudici, P., Raffinetti, E.: Cyber risk ordering with rank-based statistical models. AStA-Adv. Stat. Anal. **105**, 469–484 (2021). https://doi.org/10.1007/s10182-020-00387-0

Giudici, P., Raffinetti, E.: Shapley–Lorenz explainable artificial intelligence. Exp. Syst. Appl. **105**, 114104 (2021). https://doi.org/10.1016/j.eswa.2020.114104

Gneiting, T.: Making and evaluating point forecasts. J. Am. Stat. Assoc. **106**, 746–762 (2011). https://doi.org/10.1198/jasa.2011.r10138

Hand, D., Till, R.J.: A simple generalisation of the area under the ROC curve for multiple class classification problem. Mach. Learn. **45**, 171–186 (2001). https://doi.org/10.1023/A:1010920819831

Hoeffding, W.: A class of statistics with asymptotically normal distribution. Ann. Math. Stat. **19**, 293–325 (1948). https://doi.org/10.1214/aoms/1177730196

Joseph, A.: Shapley regressions: a framework for statistical inference in machine learning models. Working paper No. 2019/7, King's College London. https://www.kcl.ac.uk/business/assets/pdf/dafm-working-papers/2019-papers/shapley-regressions-a-framework-for-statistical-inference-on-machine-learning-models.pdf (2019). Accessed 26 September 2022

Kang, T.-H., Sharma, A., Marshall, L.: Assessing goodness of fit for verifying probabilistic forecasts. Forecasting **3**, 763–773 (2021). https://doi.org/10.3390/forecast3040047

Lerman, R., Yitzhaki, S.: A note on the calculation and interpretation of the Gini index. Econ. Lett. **15**, 363–368 (1984). https://doi.org/10.1016/0165-1765(84)90126-5

Lorenz, M.O.: Methods of measuring the concentration of wealth. Publ. Am. Stat. Assoc. **9**, 209–219 (1905). https://doi.org/10.2307/2276207

Lou, Y., Caruana, R., Gehrke, J.: Intelligible models for classification and regression. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining, pp. 150-158 (2012)

Lundberg, S., Lee, S.: A unified approach to interpreting model predictions. Adv. Neur. In. **30**, 4765–4774 (2017)

Marshall, A.W., Olkin, I., Arnold, B.C.: Inequalities: theory of majorization and its applications. Springer, New York, Dordrecht, Heidelberg & London (2011)

Molnar, C.: Interpretable machine learning. A guide for making black box models explainable. 2nd Edn (2022)

Mullins, M., Holland, C.P., Cunneen, M.: Creating ethics guidelines for artificial intelligence and big data analytics customers: the case of the consumer European insurance market. Patterns **10**, 1–14 (2021). https://doi.org/10.1016/j.patter.2021.100362

Petropoulos, F., Apiletti, D., Assimakopoulo, V., et al.: Forecasting: theory and practice. Int. J. Forecast. **38**, 705–871 (2022). https://doi.org/10.1016/j.ijforecast.2021.11.001

Raffinetti, E., Giudici, P.: Multivariate Ranks-Based Concordance Indexes. In: Di Ciaccio, A., Coli, M., Ibanez, J.M.A (eds.) Advanced statistical methods for the analysis of large data-sets, series: studies in theoretical and applied statistics. Springer, Berlin & Heidelberg, pp. 465–473 (2012)

Redelmeier, D.A., Bloch, D.A., Hickam, D.A.: Assessing predictive accuracy: how to compare brier scores. J. Clin. Epidemiol. **44**, 1141–1146 (1991). https://doi.org/10.1016/0895-4356(91)90146-Z

Schechtman, E., Yitzhaki, S.: A measure of association based on Gini's mean difference. Commun. Stat.-Theor. M. **16**, 207–231 (1987). https://doi.org/10.1080/03610928708829359

Shapley, L.: A value for n-person games. In: Kuhn, H., Tucker, A. (eds.) Contributions to the theory of games II, pp. 307–317. Princeton University Press, Princeton (1953)

Song, E., Nelson, B., Staum, J.: Shapley effects for global sensitivity analysis: theory and computation. SIAM/ASA J. Uncert. Quantif. **4**, 1060–1083 (2016). https://doi.org/10.1137/15M1048070

Strumbelj, E., Kononenko, I.: An efficient explanation of individual classifications using game theory. J. Mach. Learn. Res. **11**, 1–18 (2010). https://doi.org/10.1145/1756006.1756007