1. Introduction

Machine Learning (ML) models are boosting Artificial Intelligence applications in many domains, such as finance and health care. This is mainly due to their advantage, in terms of predictive accuracy, with respect to "classic" statistical models. However, while complex ML models can reach high predictive performance, they have an intrinsic black-box nature.

This is a problem in regulated industries, as authorities aimed at monitoring the risks arising from the application of Artificial Intelligence (AI) methods may not validate them (see, e.g. Joseph (2019) and Bracke et al. (2019)). For example, the application of AI to credit lending may lead to automated decisions that can classify a company at risk of default, without explaining the underlying rationale and, therefore, impeding remedial actions.

Accuracy and explainability are not the only desirable characteristics of a ML model. The recently proposed European regulation on Artificial Intelligence, the AI Act (European Commission, 2020), attempts to regulate the use of AI by means of a set of integrated requirements.

The AI Act introduces a risk-based approach to AI applications, defining an AI risk taxonomy with four risk categories: unacceptable risk, high risk (the main focus of this paper), limited risk, and minimal risk. The requirements established for high-risk applications include sustainability, accuracy, fairness and explainability, which need a set of metrics that can establish not only whether but also how much the requirements are satisfied over time. To the best of our knowledge, there exists no such set of

metrics, yet.

While accuracy, fairness and explainability are concepts relatively well defined in the literature, “sustainability” is more ambiguous, as it is mostly associated with environmental aspects. In this paper we align with the English language, so that “sustainability” means that an AI application will be able to continue over a period of time, producing output that are impacted neither by extreme data events nor by cyber data manipulations. But we also mean that an AI application does not create damage to the environment, through excessive energy consumption and related CO2 emissions.

In this paper, we propose a set of four main metrics, aimed at measuring Sustainability, Accuracy, Fairness and Explainability (S.A.F.E. in brief), which have the advantage of being all based on one unifying statistical method: the Lorenz curve. The Lorenz curve is a well known robust statistical tool, which has been employed to measure, on one hand, predictive accuracy and, on the other, income and wealth inequalities. It thus appears as a natural candidate on which to build an integrated set of trustworthy AI measurement metrics.

Indeed, a recent work by Giudici and Raffinetti (2021) has shown how to measure Accuracy and Explainability, using the notion of Lorenz Zonoids, based on the Lorenz Curve. The result is a metric that can, differently from other measures, such as Shapley values, jointly measure accuracy and explainability; a metric that is also robust to data variations, being based on the mutual variability, instead on the variability from the mean, as other measures, such as the Root Mean Squared Error.

In this paper we extend (Giudici and Raffinetti, 2021) to the measurement of Fairness and Sustainability, providing an overall joint measure for all S.A.F.E. AI requirements.

The explainability requirement is fulfilled "by design" through classic statistical models, such as logistic and linear regression. However, in complex data analysis problems, classical statistical models may have a limited predictive accuracy, in comparison with "black-box" ML models, such as neural networks and random forests. This suggests to empower ML models with post-modelling tools that can "explain" them.

Recent attempts in this direction, based on the cooperative game theory work of Shapley (1953), have led to promising applications of explainable AI methods in finance, among which Bracke et al. (2019) and Bussmann et al. (2020).

Shapley values have the advantage of being agnostic: independent on the underlying model with which classifications and predictions are computed; but have the disadvantage of not being normalised and, therefore, difficult to interpret and compare. To overcome this limitation, Giudici and Raffinetti (2021) proposed Shapley–Lorenz values, which combine Shapley values with Lorenz Zonoids, obtaining a measure of the contribution of each explanatory to the predictive accuracy of the response, rather than to the value of the predictions, as is the case for standard Shapley values.

In this paper we extend Giudici and Raffinetti (2021) and employ Lorenz Zonoids to build methods useful to measure not only Accuracy and Explainability, but also Sustainability and Fairness. The extension will allow to develop an integrated

measurement model for Sustainability, Accuracy, Fairness and Explainability, and a unified score of AI SAFEty.

The requirement of sustainability implies the model results are stable under variations in the data and, in particular, when extreme data, resulting from stressed scenarios and/or from cyber data manipulations, are inserted into the observed data.

To measure the sustainability of AI applications we propose to extend variable selection methods, available for probabilistic models, to non-probabilistic models, such as random forests and neural network models, using statistical tests based on the comparison between the Lorenz Zonoids of the predictions. The extension provides a model selection criterion for (non-probabilistic) ML models, not available at the moment. The criterion will lead to the choice of a parsimonious model, more sustainable than a complex one. The extension will also allow to compare the selected model with a model that would be obtained when extreme data are artificially injected into the underlying data.

The condition of fairness requires that the results of AI applications do not present biases among different population groups.

To measure the fairness of AI applications we propose to derive the Lorenz Zonoids of the predictions obtained separately for each population group, similarly to what done for the requirement of sustainability.

The paper is organised as follows: the next section illustrates the proposed methodology and, in particular, the Lorenz Zonoid tool and the proposed Lorenz Zonoid comparison tests; Section 3 discusses the empirical findings obtained applying

our proposal to the available data; finally, Section 4 contains some concluding remarks.

2. Methodology

Lorenz Zonoids were originally proposed by Koshevoy and Mosler (1996) as a generalisation of the ROC curve in a multidimensional setting. When referred to the one-dimensional case, the Lorenz Zonoid coincides with the Gini coefficient, a measure typically used for representing the income inequality or the wealth inequality within a nation or a social group (see, e.g Gini (1936)). Both the Gini coefficient and the Lorenz Zonoid measure statistical dispersion in terms of the mutual variability among the observations, a metric that is more robust to extreme data than the standard variability from the mean.

Given a variable Y and n observations, the Lorenz Zonoid can be defined from the Lorenz and the dual Lorenz curves (see Lorenz (1905)).

The Lorenz curve for a variable Y, denoted with LY, and displayed, from a graphical view point, as the red curve in Fig. 1(a), is obtained by re-ordering the Y values in a non-decreasing sense. It is built joining the set of points with coordinates $(i/n, \sum_{j=1}^{i} y_{rj}/(n\bar{y}))$, for $i=1,\ldots,n$, where r and $\bar{y}$ indicate the (non-decreasing) ranks of Y and the Y mean value, respectively. Similarly, the dual Lorenz curve of Y, pointed out as LY′ and represented by the blue curve in Fig. 1(b), is obtained by re-ordering the Y values in a non-increasing sense. Its coordinates are specified as $(i/n, \sum_{j=1}^{i} y_{dj}/(n\bar{y}))$, for $i=1,\ldots,n$, where d indicates the (non-increasing) ranks of Y. The area lying between the LY and LY′ curves is the Lorenz Zonoid.

The Lorenz Zonoid fulfils some attractive properties. An important one is the "inclusion" of the Lorenz Zonoid of any set of predicted values Yˆ into the Lorenz Zonoid of the observed response variable Y, graphically depicted in Fig. 1(b). The "inclusion property" allows to interpret the ratio between the Lorenz Zonoid of a particular predictor set Yˆ and the Lorenz Zonoid of Y as the mutual variability of the response "explained" by the predictor variables that give rise to Yˆ, similarly to what occurs in the well known variance decomposition that gives rise to the R2 measure.

Fig. 1. [(a)] The Lorenz curve (LY) and the dual Lorenz curve (LY′); [(b)] The inclusion property LZ(Yˆ)⊂LZ(Y).

A second important property concerns the practical implementation of the Lorenz Zonoid calculation. It can be shown that the Lorenz Zonoid-value of a generic variable · (such as the response variable, or the predicted response variable) is calculated as(1)LZ(·)=2Cov(· ,r(·))nE(·),where r(·) are the rank-scores associated with the · variable and E(·) is its expected value.

Eq. (1) provides an easily implementable manner to calculate a Lorenz Zonoid and, consequently, the share of Lorenz Zonoid response explained by a model's predictors. The properties of the Lorenz Zonoids can be leveraged to provide metrics to assess the SAFEty of AI applications, as in the following.

Explainability. In Giudici and Raffinetti (2021), the Lorenz Zonoid approach has been combined with the Shapley framework, to obtain a metric of explainability that measures the additional contribution of each explanatory variable to the Lorenz Zonoid of the predictions.

Given K predictors, the Shapley–Lorenz contribution associated with the additional variable Xk is:(2)LZXk(Yˆ)=∑X′⊆C(X)\ Xk|X′|!(K−|X′|−1)!K!· [LZ(YˆX′∪Xk)−LZ(YˆX′)],where: C(X)\ Xk is the set of all the possible model configurations which can be obtained excluding variable Xk; |X′| denotes the number of variables included in each possible model; LZ(YˆX′∪Xk) and LZ(YˆX′) describe the (mutual) variability of the response variable Y explained by the models which, respectively, include the X′∪Xk predictors and only the X′ predictors.

The application of formula (2) leads to the Shapley–Lorenz values, a measure of the response variable mutual variability explained by each predictor, normalised in the interval [0,1]. Normalisation is an important advantage of the Shapley–Lorenz measure, with respect to the standard Shapley values. Another important advantage is that the Shapley–Lorenz measure can be calculated for any ordered response variable in the same manner, following (1), differently from measures based on the variance decomposition. And, finally, being based on the mutual variability, it is highly robust

to extreme observations.

Given a ML model with K predictors, we can thus measure its explainability score as in the following definition.

Definition 1 Explainability Score

The score for explainability can be calculated on the whole sample as:(3)Ex − Score=∑k=1KSLkLZ(Y),where LZ(Y) corresponds to the response variable Y Lorenz Zonoid-value, and SLk denotes the Shapley–Lorenz values associated with the kth predictor.

Accuracy. The accuracy of the predictions generated by a ML model is crucial for ensuring trustworthiness of AI applications. The statistical learning literature provides a large set of accuracy metrics (for a review see, e.g. Hand et al. (2001)): the most commonly employed are the Root Mean Squared Error (when the response variable is on a continuous scale) and the Area Under the ROC curve (when the response variable is on a binary scale). Both are calculated on a test sample of the data, assuming the model being calculated on the remaining training sample. A more robust measure is the Lorenz Zonoid, which can be calculated on the test set in the same way for binary, ordered categorical and continuous responses. This generality is a clear further advantage of the Lorenz Zonoid.

Given a ML model with k≤K predictors, and a test sample from the dataset, we can measure its accuracy score as in the following definition.

Definition 2 Accuracy Score

The score for accuracy can be defined as:(4)Ac −

Score=LZ(Yˆ X1,…,Xk)LZ(Ytest),where LZ(Yˆ X1,…,Xk) is the Lorenz Zonoid of the predicted response variable, obtained using K predictors on the test set, and LZ(Ytest) is the Y response variable Lorenz Zonoid value computed on the same test set.

Note that, while the explainability score is calculated on the whole dataset, in line with its nature, the accuracy score is calculated on the test data set, using the ML model learned on the train data set.

In this respect, a significance test for the difference in Lorenz Zonoids, which can extend Diebold and Mariano (1995) for continuous responses and DeLong et al. (1988) for binary response into a unifying criterion would provide the basis for a stepwise model comparison algorithm which may lead to a parsimonious model, with $k \leq K$ predictors that, while not significantly losing accuracy, simplifies the computational effort necessary to measure explainability, which can be applied only to k rather than K variables. Additionally, a more parsimonious model will likely be more sustainable: less dependent on data variations.

According to the mentioned saving of computational effort, we suggest a forward stepwise procedure, which starts with the construction of K models, each one depending on only one predictor. The application of formula (1) to all such univariate models will provide a ranking of the candidate predictors, in terms of their (marginal) importance, which can be used to determine insertion into the model. The first explanatory variable to be considered is that with the highest Lorenz Zonoid value. At the second step, a model with also the second ranked variable is fitted and a predictive

gain, measured as the additional contribution to predictive accuracy determined by the second variable can be calculated as:(5)pay − off(Xk)=LZ(Y^X′∪Xk)−LZ(Y^X′),where LZ(Y^X′∪Xk) and LZ(Y^X′) describe the (mutual) variability of the response variable Y explained by the models which, respectively, include X′∪Xk predictors or only X′ predictors.

The procedure can continue until the predictive gain defined in (5) is found not significant. To test for significance, a statistical test can be developed. The formalisation of the test is reported in Appendix.

Fairness. Fairness is a property that essentially requires that AI applications do not present biases among different population groups.

To measure fairness we propose to extend the Gini coefficient, originally developed to measure the concentration of income in a population, to the measurement of the concentration of the explanatory variables which may be affected by bias, in terms of the Shapley–Lorenz values.

Our proposal can be illustrated as follows. Let m=1,…,M be the considered population groups and let K the number of the available predictors. We denote with vmXkSL the Shapley–Lorenz value associated with the kth predictor in the mth population.

Suppose that the stepwise procedure based on the application of the Lorenz-Zonoid test leads to choose only a subset of all the available explanatory variables as the most contributing to the predictive accuracy of the model. Specifically, we denote with k⋆, where k⋆=1,…,k and such that k⋆<K, the number of predictors which compose the

selected model.

With the purpose of measuring the explainability and accuracy provided by each explanatory variable included into the final model, we consider the vector VMSL* defined as VMSL*={v1SL*,…,vmSL*,…,vMSL*}, where vmSL*=vmX1SL+⋯+vmXk*SL represents the sum of the Shapley–Lorenz values related to the predictors X1,…,Xk*.

The Gini coefficient can be applied to the vector VMSL*, obtaining a measure of concentration of the variables' importance among different population groups. For a given set of selected explanatory variables, Shapley–Lorenz values which are similar in the M populations lead to a Gini coefficient close to 0, indicating that the effect of these variables is fair across the different population groups. On the other hand, a Gini coefficient close to 1 indicates that the variables' effect largely depend on some groups, highlighting biasness.

Given a ML model with k* and M population groups, we can measure its fairness score as in the following definition.

Definition 3 Fairness Score

The score for fairness can be defined as:(6)Fair − Score=1−LZ(VMSL*),where LZ(VMSL*) denotes the Lorenz Zonoid (Gini coefficient) computed on the vector VMSL* whose elements correspond to the sum of the selected predictors' Shapley–Lorenz values in each population.

Sustainability. The results from a ML model, especially when a large number of explanatory variables is considered, may be altered by the presence of "extreme" data

points, deriving from anomalous events, or from cyber data manipulation.

We propose to verify sustainability by comparing predictive accuracy, as measured by Shapley–Lorenz values, in different ordered subset of the data, possibly altered artificially by anomalous or cyber manipulated ones.

To this aim, conditionally on a ML model, we can order the predicted response values (in the test set) in terms of their predictive accuracy, from the most accurate to the lowest. We can then divide the ordered predictions in g=1,…,G equal size groups (such as the deciles of the distribution). We can then proceed in analogy with the fairness case and build a vector including the sum of the Shapley–Lorenz values of the predictors composing the final model, i.e. $VGSL^{\star}=\{v1SL^{\star},\ldots,vgSL^{\star},\ldots,vGSL^{\star}\}$, where $vgSL^{\star}=vgX1SL+\cdots+vgXk^{\star}SL$ represents the sum of the Shapley–Lorenz values related to the predictors $X1,\ldots,Xk^{\star}$.

Definition 4 Sustainability Score

The score for sustainability can then be defined as:(7)$Sust-Score=1-LZ(VGSL^{\star})$,where $LZ(VGSL^{\star})$ indicates the Lorenz Zonoid (Gini coefficient) calculated on the vector $VGSL^{\star}$, whose elements correspond to the sum of the selected predictors' Shapley–Lorenz values in each group.

In the next Section we will apply our proposed methodology in the context of bitcoin price prediction.

3. Application to bitcoin price prediction

As an illustrative example of how to apply our proposal, we consider a set of cryptocurrency time series, for the time period between May 18th, 2016 and April

30th, 2018.

3.1. Data description

The considered data are the same described in Giudici and Abu-Hashish (2019) and in Giudici and Lorenz (2020) to explain bitcoin price variation as a function of the available financial explanatory variables.

A further investigation of the data was provided in a work by Giudici and Raffinetti (2021), who introduced a new AI approach resulting in the formalisation of a normalised measure for the assessment of the contribution of each additional predictor to the explanation of the bitcoin prices.

For coherence with the previous cited works, here we choose the same time series observations, with the bitcoin prices from the Coinbase exchange as the target variable to be predicted. As suggested by Giudici and Lorenz (2020) and Giudici and Raffinetti (2021), the time series for Oil, Gold and SP500 prices are taken into account as candidate financial explanatory variables. In line with Giudici and Abu-Hashish (2019), the exchange rates USD/Yuan and USD/Eur are also included as possible further explanatory variables.

Our aim is to exploit the Lorenz Zonoid tool as a unified criterion for measuring the SAFEty of AI methodologies.

3.2. Explorative analysis

We start our explorative analysis of the available data by plotting the time evolution of bitcoin prices, together with that of the Gold, Oil and SP500 prices and the exchange rates, in the considered time period. The trends are displayed in Fig. 2, Fig.

3, Fig. 4, Fig. 5, Fig. 6, Fig. 7, respectively.

Specifically, from Fig. 2 the bitcoin price appears quite stable until the beginning of 2017. But, since the first six months of the 2017 year, bitcoin prices begin to progressively increase reaching the maximum at the end of the same year. This dynamics is followed by a downtrend, which starts in January 2018.

While the trend of the SP500 increases overtime (Fig. 3), the prices of Gold and Oil (Fig. 4, Fig. 5) are characterised by uptrend and downtrend. The former is more evident at the end of the 2016 year for Gold, while for Oil it occurs some months before the end of the 2016.

On the other hand, the behaviour of the exchange rates USD/Eur and USD/Yuan is quite similar overtime, as shown in Fig. 6, Fig. 7.

Fig. 2. Bitcoin prices.

Download: Download high-res image (200KB)

Download: Download full-size image

Fig. 3. SP500 prices.

Download: Download high-res image (248KB)

Download: Download full-size image

Fig. 4. Gold prices.

Download: Download high-res image (276KB)

Download: Download full-size image

Fig. 5. Oil prices.

Download: Download high-res image (221KB)

Download: Download full-size image

Fig. 6. USD/EUR exchange rate.

Download: Download high-res image (209KB)

Download: Download full-size image

Fig. 7. USD/YUAN exchange rate.

To better understand the dynamics reported in Fig. 2, Fig. 3, Fig. 4, Fig. 5, Fig. 6, Fig. 7, some summary statistics are reported in Table 1.

The results in Table 1 highlight that the bitcoin price mean value, as well as the standard deviation and the minimum and maximum values, are largely different with respect to those of the classical assets and exchange rates. To better appreciate the volatility magnitude of the prices, the coefficient of variation (cv) is computed and displayed in Table 1. The findings show that the exchange rates are much less volatile than the bitcoin, SP500 and Oil prices. Indeed, for USD/Eur and USD/Yuan, the standard deviations are only 5% and 3% the size of the mean, respectively. A similar result in terms of volatility is achieved by Gold, whose standard deviation corresponds to 4% the size of the mean, while for Oil and SP500 the standard deviations slightly increase reaching values which are less than 10% of the mean.

Table 1. Summary statistics for Coinbase bitcoin, classic asset prices, SP500 index

and exchange rates (mean, standard deviations (sd), coefficient of variation (cv), minimum and maximum values).

Prices	Mean	sd	cv	Min
Coinbase bitcoin	3919.05	4318.98	1.10	438.38
SP500	2399.17	212.31	0.09	2000.54
Gold	1275.58	52.34	0.04	1128.42
Oil	49.36	3.37	0.07	39.51
USD/Eur	0.88	0.04	0.05	0.80
USD/Yuan	6.68	0.19	0.03	6.27

3.3. Results

The aim of the data analysis is to build an explainable ML model that can predict bitcoin prices. Before proceeding, we transform all price series into their percentage returns. This because returns are scale free and the corresponding series are stationary (see, e.g. Tsay (2005)).

As a ML model we apply, without loss of generality, a neural network with five hidden layers. We consider as training data the time series until December 31st, 2017; and as test data the 2018 time series. Fig. 2, Fig. 3, Fig. 4, Fig. 5, Fig. 6, Fig. 7 show that it will be difficult to obtain a high predictive accuracy, as the time series trends in 2018 change patterns with respect to the training data series.

In any case, the application of our proposed approach leads to a series of predictions

for the 2018 return prices that can be compared with the actual returns, to obtain measures of trustworthiness (S.A.F.E.ty) of the neural network. Fig. 8 shows the results of such assessment, in graphical format.

Fig. 8(a) shows that the score of explainability of the full model, measured as the sum of all Shapley–Lorenz values (on all data), is equal to 0.5714, with the Gold price returns as the highest contributor.

Download: Download high-res image (581KB)

Download: Download full-size image

Fig. 8. Continuous scenario - [(a)] Explainability; [(b)] Accuracy; [(c)] Sustainability; [(d)] Fairness.

To simplify the model, we have then applied our proposed forward stepwise feature selection, following the order of the variables, in terms of their Lorenz Zonoid marginal contribution. The procedure inserts Gold returns, then SP500 returns and then it stops, as no additions lead to a significantly superior model. Our selected model, therefore, contains Gold and SP500 returns as predictors of bitcoin prices.

Fig. 8(b) shows the accuracy score of the selected model, in terms of Lorenz Zonoid. The Lorenz Zonoid gives an accuracy score of 0.3280, which correspond to the percentage of bitcoin price variability explained by the model (on the test data).

We have then assessed the sustainability score of the selected model. To this aim, we have ordered the test data response according to how well is predicted by the model (from the best to the worst predictions) and, accordingly, subdivided it into ten deciles. We have then calculated the Lorenz Zonoid of the model, separately in each cumulative decile. The result is shown in Fig. 8(c).

Fig. 8(c) shows that, as expected, the predictions worsen, although not monotonically, as we increase deciles. Monotonicity does not hold as both the predictions and the values to be predicted vary along deciles. For example, the model goes relatively well in the tenth decile because not only the predictions but also the observations are less variable.

According to our proposal, we can calculate, as a sustainability score, the complement

of the Gini coefficient of the Lorenz Zonoid. It results to be equal to 0.8314, indicating a high sustainability.

With the aim of assessing fairness, we have considered, as a potential biasing variable, the amount traded in each day, and evaluate whether price returns are fair with respect to it. If not, it will mean that bitcoin returns depend on the trading volumes.

To measure fairness we have ordered the test data response in terms of the corresponding trading volumes (from the lowest to the highest) and, accordingly, subdivided it into ten deciles. We have then calculated the Lorenz Zonoid of the model, separately in each cumulative decile. The result is shown in Fig. 8(d).

Fig. 8(d) indicates that the model has the best performance in correspondence to the lowest and highest volumes of trading but also that, overall, the variation is limited.

According to our proposal, we have computed as a fairness score, the complement of the Gini coefficient of the Lorenz Zonoid. It results to be equal to 0.8617, indicating a high fairness.

To show the universality of our proposal, we have binarised the response variable, with Y=1 indicating positive returns and Y=0 indicating negative returns, and applied the same neural network model as before, but to predict a binary, rather than a continuous response. Fig. 9 shows the results of our S.A.F.E.ty assessment, in graphical format.

From Fig. 9(a), note that the model presents a lower overall explainability than before: the overall explainability score is equal to 0.3160. As before, the Gold price return is the most explainable series.

Download: Download high-res image (555KB)

Download: Download full-size image

Fig. 9. Binarised scenario - [(a)] Explainability; [(b)] Accuracy; [(c)] Sustainability; [(d)] Fairness.

Our proposed model selection procedure is then carried out exactly as for the continuous case. The selected model contains SP500 and Gold returns, as in the continuous scenario. The accuracy score of the model (see Fig. 9(b)) is equal to 0.4088, higher than before, as expected, since the response variable now varies on a binary, rather than on a continuous scale.

We have finally applied the sustainability and fairness assessments, in the same manner as for the continuous case. The results are in Figs. 9 (c) and 9 (d), corresponding to scores of, respectively, 0.8184 and 0.7165. While the sustainability of the model is similar to that corresponding to the continuous response case, fairness is lower, indicating that the sign of the returns depend on trading volumes more than the actual returns do.

To better evaluate our proposal, we now compare it with the most employed alternative metrics. Specifically: to measure explainability, we compare our Shapley Lorenz proposal with Shapley values; to measure accuracy, we compare our Lorenz Zonoid proposal with the AUROC (for the binary response case) and the RMSE (for the continuous response case); to measure sustainability (and, similarly, fairness) we employ the Gini index to measure the variability of model accuracy in different percentiles of the response variable.

In Table 2, Table 3, we report the findings for explainability comparing, for each

candidate predictor, the Shapley Lorenz values with the Global Shapley values, obtained summing Shapley values over all statistical units.

From Table 2, Table 3 note that the order of importance of the predictors is the same using either Shapley Lorenz or Shapley values. However, Shapley Lorenz values have the advantage of being normalised and, therefore, of being easily interpretable. For example, in the continuous response case, Shapley Lorenz values lead to the conclusion that Gold explains about 35% of the predictive accuracy of the model; about three times more than Oil. Whereas, when Shapley values are considered, the values are not normalised, and it is much more difficult to interpret ratios between predictors' explanations.

Table 2. Shapley–Lorenz values vs Global Shapley values (continuous case).

Predictor	Shapley–Lorenz values	Global Shapley values
Gold	0.3500	−4.17e−05
SP500	0.1036	−2.04e−05
Oil	0.0123	1.51e−05
USD/Eur	0.0759	6.10e−06
USD/Yuan	0.0237	1.52e−06

Table 3. Shapley–Lorenz values vs Global Shapley values (binarised case).

Predictor	Shapley–Lorenz values	Global Shapley values

Predictor	Shapley–Lorenz values	Global Shapley values
Gold	0.1722	0.0323
SP500	0.0270	0.0023
Oil	0.0591	0.0269
USD/Eur	0.026	−0.0119
USD/Yuan	0.0309	−0.0015

Table 4 reports the findings for explainability comparing our accuracy scores, based on the Lorenz Zonoid of the selected model, against the Root Mean Squared Error of the same model (for a continuous response) and the Area Under the ROC curve (for a binary response).

From Table 4 note that, in the continuous case, the relatively low value of the accuracy score (about 33% of the total variability) is matched by the high value of RMSE (about 6% on the return scale). Clearly, the accuracy score is much easier to interpret. The accuracy score improves (but remains low) when the target response is binarised, reaching about 41% of the total variability, consistently with an AUROC equal to about 55%. In the binary case, both measures are normalised, and the interpretation is quite clear in both cases. We remark that an important advantage of our accuracy score is its universality: it can be applied regardless the type of the underlying target variable or model. In our application, this allows to compare the continuous and the binary case and conclude that the selected model better predicts

the binarised rather than the continuous bitcoin returns. The traditional metrics, including RMSE and AUROC, are response specific, are defined on different measurement scales and do not allow such comparison.

Table 4. Lorenz Zonoid values vs RMSE and AUROC.

Ac-Scoregold,sp500 (continuous case)	RMSEgold,sp5
0.3280	0.0586
Ac-Scoresp500,gold (binary case)	AUROCsp55,g
0.4088	0.5493

Fig. 10 and Table 5 report the findings for sustainability, comparing the Gini coefficient as a measure of variability of the accuracy scores, whether calculated with the Lorenz Zonoid (for both the binary and continuous response case) the RMSE (for the continuous case) or the AUROC (for the binary case), all applied to the cumulative deciles of all observed values of the response variable. The graphical behaviour of the RMSE and AUROC are displayed in Fig. 10(a) and (b), and should be compared with Figs. 8 (c) and 9 (c), respectively.

Comparing Fig. 10(a) with Fig. 8(c), it is clear that the RMSE metric is much more unstable and, therefore, less sustainable, than our proposed Lorenz Zonoid metric. Whereas, comparing Fig. 10(b) with Fig. 9(c), the AUROC metric seems rather stable, similarly to our proposed metric.

Download: Download high-res image (269KB)

Download: Download full-size image

Fig. 10. Sustainability comparison [(a)] RMSE (continuous case); [(b)] AUROC (binary case).

For a more accurate comparison, Table 5 reports the variability scores, all calculated as the complement of the Gini coefficient, for our Lorenz Zonoid case, the RMSE and the AUROC.

Table 5 confirms that, for a continuous response, the use of the RMSE metric leads to an increase of the Gini coefficient and, therefore, a reduction in sustainability. Whereas, for a binary response, the AUROC metric has a high sustainability, similarly to our proposed metrics. This result is consistent with the well known relationship between the AUROC and the Gini coefficient (see e.g. Hand et al. (2001)). We remark that the universality of our proposal allows to directly compare

the binary and continuous case, leading to similar values of the Gini coefficient. Such a comparison is not possible with the other metrics, which are response specific, and expressed in different scales.

Table 5. Sust-Score based on: Shapley–Lorenz values; RMSE; AUROC.

Continuous case	Binary case
Sust-Score (Shapley–Lorenz values)	Sust-Score (Shapley–Lorenz values)
0.8314	0.8184
Sust-Score (based on RMSE)	Sust-Score (based on AUROC)
0.6597	0.9307

Note that what shown for the sustainability metrics can be replicated, in a similar way, for our proposed fairness metrics. This in line with the construction of our proposed fairness score, similar to the sustainability score.

4. Conclusions

The aim of the paper was to provide an integrated set of metrics able to assess the trustworthiness of AI applications.

The suitability of such metrics can be evaluated relying on the meta-concepts expressed by the human communities and, in particular, by those contained in the proposed regulations of Artificial Intelligence, such as the European AI Act. To provide a set of metrics that satisfy the proposed regulatory principles, we have extended the application of Lorenz Zonoids to obtain measurement tools for the Sustainability, Accuracy, Fairness and Explainability, as key S.A.F.E. trustworthiness

criteria.

By means of an easily downloadable dataset of bitcoin prices, and related candidate predictors, we have provided a practical demonstration of how to implement and interpret the proposed metrics.

The application of our proposal, and its comparison with alternative metrics, in both binary and continuous scenarios, and for all ordered deciles of the response variable, shows that our S.A.F.E. framework is a more suitable metric to assess trustworthy AI than the available metrics, such as Shapley Values, RMSE and AUROC. This result derives from the nature of the underlying statistical tool, the Lorenz Zonoid, which allows to obtain a metric that is independent of the considered response variable, and which is more robust under data variations.

Our proposed metrics can be easily embedded in a scorecard that can be beneficial to: asset management companies that need reliable predictions to make investment decisions; financial authorities and supervisors that need to evaluate AI methods implemented by the institutions under their supervision; researchers that need to understand the functioning of financial markets.

CRediT authorship contribution statement

Paolo Giudici: Supervision, Modelling, Interpretation. Emanuela Raffinetti: Modelling, Elaboration, Interpretation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The Authors acknowledge support from the European Horizon2020 PERISCOPE programme, contract number n. 101016233 and 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.

Appendix.

To determine if the predictive gain provided by the predictor additionally included into the model is significant, a statistical test has to be formalised. To do that Eq. (5) has to be rewritten in terms of covariance operators as follows:(8)LZ(YˆX′∪Xk)−LZ(YˆX′)=2Cov(YˆX′∪Xk,r(YˆX′∪Xk))nE(YˆX′∪Xk) −2Cov(YˆX′,r(YˆX′))nE(YˆX′).

As r(·)/n is the empirical transformation of the cumulative distribution function F(·) (see, e.g. Lerman and S. (1984)), the terms in Eq. (8) can be re-expressed as:(9)LZ(YˆX′∪Xk)−LZ(YˆX′)=2Cov(YˆX′∪Xk,F(YˆX′∪Xk))E(YˆX′∪Xk)−2Cov (YˆX′,F(YˆX′))E(YˆX′),where F(YˆX′∪Xk) and F(YˆX′) are the cumulative distribution functions of YˆX′∪Xk and YˆX′, respectively.

In the case of linear regression, the mean of the predicted response values is always equal to the mean of the original target values, implying that E(Y)=E(Yˆ). For more general models, the aforementioned condition does not fully hold, implying that E(YˆX′∪Xk)=E(YˆX′)=μ becomes a reasonable approximation. Assuming such

approximation, Eq. (9), which describes the marginal contribution (MC) provided by Xk, can be simplified as follows:(10)MC=2Cov(Y^X′∪Xk,F(Y^X′∪Xk))μ−2Cov(YX′,F(Y^X′))μ.

In line with the previous mathematical derivations, we propose γ as an adjusted version of Eq. (10), i.e.(11)γ=μ2· MC=Cov(Y^X′∪Xk,F(Y^X′∪Xk))−Cov(Y^X′,F(Y^X′)).

By denoting the covariances Cov(Y^X′∪Xk,F(Y^X′∪Xk))=ξ(Y^X′∪Xk) and Cov(Y^X′,F(Y^X′))=ξ(Y^X′), γ in (11) can be re-written as:(12)γ=ξ(π^X′∪Xk)−ξ(π^X′).

A test for the equality of the two Lorenz Zonoids, can thus be developed by setting the following hypothesesH0:ξ(Y^X′∪Xk)=ξ(Y^X′)vsH1:ξ(Y^X′∪Xk)≠ξ(Y^X′).

To proceed with the test, ξ(Y^X′∪Xk) can be derived in terms of a U-statistic, U1, which estimates Cov(Y^X′∪Xk,

F(Y^X′∪Xk)). The estimator is defined as:ξ^(Y^X′∪Xk)=U1=14n2∑i=1n(2i−1−n)Y^X′∪Xk(i),where Y^X′∪Xk(i) is the ith order statistic of Y^X′∪Xk1,…,

Y^X′∪Xkn.

Similarly, the estimator of ξ(Y^X′) is U2, specified as:ξ^(Y^X′)=U2=14n2∑i=1n(2i−1−n)Y^X(i)′,where Y^X(i)′ is the ith order statistic of Y^X1′,…,

Y^Xn′.

An estimator of γ=ξ(Y^X′∪Xk)−ξ(Y^X′) can then be provided as a function of two

dependent U-statistics:(13)$\hat{\gamma}=\hat{\xi}(\hat{Y}\hat{X}'\cup Xk)-\hat{\xi}(\hat{Y}\hat{X}')=U1-U2$.

Based on Hoeffding (1948), a function of several dependent U-statistics has, after appropriate normalisation, an asymptotically normal distribution. As suggested by Schechtman et al. (2008), a way to estimate the variance is to resort to the jackknife method. Specifically, the n values of $\hat{\gamma}$, pointed out with $\hat{\gamma}(-i)$ (where i=1,…,n), are calculated by omitting one pair $(\hat{Y}\hat{X}'\cup Xk,\hat{Y}\hat{X}')$ at a time and the estimated variance is

Download: Download high-res image (84KB)

Download: Download full-size image

where $\bar{\gamma}$ is the average of $\hat{\gamma}(-i)$, for i=1,…,n.

Following the previous derivations, the null hypothesis H0:$\xi(\hat{Y}\hat{X}'\cup Xk)=\xi(\hat{\pi}\hat{X}')$ can be tested by the test statistic:(14) and, for a given selected significance level α, a rejection region for the null hypothesis H0 can be defined as $|Z|\geq z\alpha 2$.