SAFE Artificial Intelligence in Finance

Paolo Giudici
Department of Economics and Management, University of Pavia, Via San Felice 5, 27100 Pavia (Italy)
email: paolo.giudici@unipv.it

Emanuela Raffinetti
Department of Economics and Management, University of Pavia, Via San Felice 5, 27100 Pavia (Italy)
email: emanuela.raffinetti@unipv.it

Abstract

Financial technologies, boosted by the availability of machine learning models, are expanding in all areas of finance: from payments (peer to peer lending) to asset management (robot advisors) to payments (blockchain coins). Machine learning models typically achieve a high accuracy at the expense of an insufficient explainability. Moreover, according to the proposed regulations, high-risk AI applications based on machine learning must be "trustworthy", and comply with a set of mandatory requirements, such as Sustainability and Fairness. To date there are no standardised metrics that can ensure an overall assessment of the trustworthiness of AI applications in finance. To fill the gap, we propose a set of integrated statistical methods, based on the Lorenz Zonoid tool, that can be used to assess and monitor over time whether an AI application is trustworthy. Specifically, the methods will measure Sustainability (in terms of robustness with respect to anomalous data), Accuracy (in terms of predictive accuracy), Fairness (in terms of prediction bias across different population groups) and Explainability (in terms of human understanding and oversight). We apply our proposal to an easily downloadable dataset, that concerns financial prices, to make our proposal easily reproducible.

Keywords: Lorenz Zonoids; Accuracy; Explainability; Fairness; Sustainability; Bitcoin price prediction.

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. [12] and [1]). 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 [5], 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 those about sustainability, accuracy, fairness and explainability, which need a set of integrated 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.

In this paper, we propose to fill the gap building a framework based on a set of four main metrics, aimed at measuring: Sustainability, Accuracy, Fairness and Explainability (S.A.F.E. in brief). We show how to build such framework using statistical methods based on the unifying notion of Lorenz Zonoid. Doing so, we will extend the recent work of [9], who has showed how to jointly measure Accuracy and Explainability.

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 ([17]), have led to promising applications of explainable AI methods in finance, among which [1] and [2].

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, [9] 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 [9] and employ Lorenz Zonoids to build methods useful to measure not only Accuracy and Explianability, 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 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 requirement 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 organized as follows: the next section illustrates the proposed methodology and, in particular, the Lorenz Zonoid tool and 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 [13] 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 [6]). 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 [15]).

The Lorenz curve for a variable Y, denoted with L_Y, and displayed, from a graphical view point, as a red curve in Figure 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 $(i/n, \sum_{j=1}^{i} y_{r_j}/(n\bar{y}))$, 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 $L_Y^{'}$ and represented by the blue curve in Figure 1 (b), is obtained by re-ordering the Y values in a non-increasing sense. Its coordinates are specified as $(i/n, \sum_{j=1}^{i} y_{d_j}/(n\bar{y}))$, for $i = 1, \ldots, n$, where d indicates the (non-increasing) ranks of Y. The area lying between the L_Y and $L_Y^{'}$ curves is the Lorenz Zonoid.

The Lorenz Zonoid fulfills some attractive properties. An important one is the "inclusion" of the Lorenz Zonoid of any set of predicted values $\hat{Y}$ into the Lorenz Zonoid of the observed response variable Y, graphically depicted in Figure 1 (b). The "inclusion property" allows to interpret the ratio between the Lorenz Zonoid of a particular predictor set $\hat{Y}$ and the Lorenz Zonoid of Y as the mutual variability of the response "explained" by the predictor variables that give rise to $\hat{Y}$, similarly to what occurs in the well known variance decomposition that gives rise to the R^2 measure.

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 $\cdot$ (such as the response variable, or the predicted response variable) is calculated as

$$LZ(\cdot) = \frac{2Cov(\cdot, r(\cdot))}{nE(\cdot)}, \tag{1}$$

where $r(\cdot)$ are the rank-scores associated with the $\cdot$ variable and $E(\cdot)$ is its

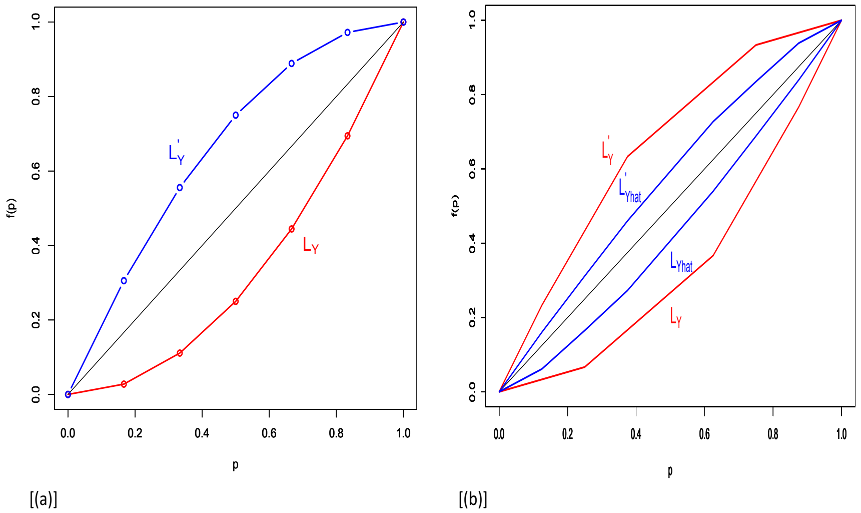

Figure 1: [(a)] The Lorenz curve (L_Y) and the dual Lorenz curve (L'_Y); [(b)] The inclusion property $LZ(\hat{Y}) \subset LZ(Y)$

expected value.

Equation (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 [9], 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 X_k is:

$$LZ^{X_k}(\hat{Y}) = \sum_{X^{'} \subseteq \mathcal{C}(X) \setminus X_k} \frac{|X^{'}|!(K - |X^{'}| - 1)!}{K!} \cdot \\ [LZ(\hat{Y}_{X^{'} \cup X_k}) - LZ(\hat{Y}_{X^{'}})], \tag{2}$$

where: $\mathcal{C}(X) \setminus X_k$ is the set of all the possible model configurations which can be obtained excluding variable X_k; $|X^{'}|$ denotes the number of variables included in each possible model; $LZ(\hat{Y}_{X^{'} \cup X_k})$ and $LZ(\hat{Y}_{X^{'}})$ describe the (mutual) variability of the response variable Y explained by the models which, respectively, include the $X^{'} \cup X_k$ 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:*

$$Ex\text{-}Score = \frac{\sum_{k=1}^{K} SL_k}{LZ(Y)}, \tag{3}$$

where $LZ(Y)$ corresponds to the response variable Y Lorenz Zonoid-value, and SL_k denotes the Shapley-Lorenz values associated with the k-th 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. [10]): 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 $k \leq K$ 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:*

$$Ac\text{-}Score = \frac{LZ(\hat{Y}_{X_1,\ldots,X_k})}{LZ(Y_{test})}, \tag{4}$$

where $LZ(\hat{Y}_{X_1,\ldots,X_k})$ is the Lorenz Zonoid of the predicted response variable, obtained using k predictors on the test set, and $LZ(Y_{test})$ 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 [4] for continuous responses and [3] 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:

$$\textit{pay-off}\,(X_k) = LZ(\hat{Y}_{X' \cup X_k}) - LZ(\hat{Y}_{X'}), \tag{5}$$

where $LZ(\hat{Y}_{X' \cup X_k})$ and $LZ(\hat{Y}_{X'})$ describe the (mutual) variability of the response variable Y explained by the models which, respectively, include $X^{'} \cup X_k$ 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 obtained rewriting equation (5) in terms of covariance operators as follows:

$$\begin{gathered} LZ(\hat{Y}_{X' \cup X_k}) - LZ(\hat{Y}_{X'}) = \\ \frac{2Cov(\hat{Y}_{X' \cup X_k}, r(\hat{Y}_{X' \cup X_k}))}{nE(\hat{Y}_{X' \cup X_k})} - \frac{2Cov(\hat{Y}_{X'}, r(\hat{Y}_{X'}))}{nE(\hat{Y}_{X'})}. \end{gathered} \tag{6}$$

As $r(\cdot)/n$ is the empirical transformation of the cumulative distribution function $F(\cdot)$ (see, e.g. [14]), the terms in equation (6) can be re-expressed as:

$$\begin{gathered} LZ(\hat{Y}_{X' \cup X_k}) - LZ(\hat{Y}_{X'}) = \\ \frac{2Cov(\hat{Y}_{X' \cup X_k}, F(\hat{Y}_{X' \cup X_k}))}{E(\hat{Y}_{X' \cup X_k})} - \frac{2Cov(\hat{Y}_{X'}, F(\hat{Y}_{X'}))}{E(\hat{Y}_{X'})}, \end{gathered} \tag{7}$$

where $F(\hat{Y}_{X' \cup X_k})$ and $F(\hat{Y}_{X'})$ are the cumulative distribution functions of $\hat{Y}_{X' \cup X_k}$ and $\hat{Y}_{X'}$, 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 $E(Y) = E(\hat{Y})$. For more general models, the aforementioned condition does not fully hold, implying that $E(\hat{Y}_{X' \cup X_k}) = E(\hat{Y}_{X'}) = \mu$ becomes a reasonable approximation. Assuming such approximation, equation (7), which describes the marginal contribution (MC) provided by X_k, can be simplified as follows:

$$MC = \frac{2Cov(\hat{Y}_{X' \cup X_k}, F(\hat{Y}_{X' \cup X_k}))}{\mu} - \frac{2Cov(Y_{X'}, F(\hat{Y}_{X'}))}{\mu}. \tag{8}$$

In line with the previous mathematical derivations, we propose γ as an adjusted version of equation (8), i.e.

$$\gamma = \frac{\mu}{2} \cdot MC = Cov(\hat{Y}_{X' \cup X_k}, F(\hat{Y}_{X' \cup X_k})) - Cov(\hat{Y}_{X'}, F(\hat{Y}_{X'})). \tag{9}$$

By denoting the covariances $Cov(\hat{Y}_{X' \cup X_k}, F(\hat{Y}_{X' \cup X_k})) = \xi(\hat{Y}_{X' \cup X_k})$ and $Cov(\hat{Y}_{X'}, F(\hat{Y}_{X'})) = \xi(\hat{Y}_{X'})$, γ in (9) can be re-written as:

$$\gamma = \xi(\hat{Y}_{X' \cup X_k}) - \xi(\hat{Y}_{X'}). \tag{10}$$

A test for the equality of the two Lorenz Zonoids, can thus be developed by setting the following hypotheses

$$H_0 : \xi(\hat{Y}_{X' \cup X_k}) = \xi(\hat{Y}_{X'}) \quad \text{vs} \quad H_1 : \xi(\hat{Y}_{X' \cup X_k}) \neq \xi(\hat{Y}_{X'}).$$

To proceed with the test, $\xi(\hat{Y}_{X' \cup X_k})$ can be derived in terms of a U-statistic, U_1, which estimates $Cov(\hat{Y}_{X' \cup X_k}, F(\hat{Y}_{X' \cup X_k}))$. The estimator is defined as:

$$\hat{\xi}(\hat{Y}_{X' \cup X_k}) = U_1 = \frac{1}{4\binom{n}{2}} \sum_{i=1}^{n} (2i - 1 - n) \hat{Y}_{X' \cup X_{k(i)}},$$

where $\hat{Y}_{X' \cup X_{k(i)}}$ is the i-th order statistic of $\hat{Y}_{X' \cup X_{k1}}, \ldots, \hat{Y}_{X' \cup X_{kn}}$.

Similarly, the estimator of $\xi(\hat{Y}_{X'})$ is U_2, specified as:

$$\hat{\xi}(\hat{Y}_{X'}) = U_2 = \frac{1}{4\binom{n}{2}} \sum_{i=1}^{n} (2i - 1 - n) \hat{Y}_{X'_{(i)}},$$

where $\hat{Y}_{X'_{(i)}}$ is the i-th order statistic of $\hat{Y}_{X'_1}, \ldots, \hat{Y}_{X'_n}$.

An estimator of $\gamma = \xi(\hat{Y}_{X' \cup X_k}) - \xi(\hat{Y}_{X'})$ can then be provided as a function of two dependent U-statistics:

$$\hat{\gamma} = \hat{\xi}(\hat{Y}_{X' \cup X_k}) - \hat{\xi}(\hat{Y}_{X'}) = U_1 - U_2. \tag{11}$$

Based on [11], a function of several dependent U-statistics has, after appropriate normalisation, an asymptotically normal distribution. As suggested by [16], 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 $i = 1, \ldots, n$), are calculated by omitting one pair $(\hat{Y}_{X' \cup X_k}, \hat{Y}_{X'})$ at a time and the estimated variance is

$$\widehat{Var(\hat{\gamma})} = \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\gamma}_{(-i)} - \bar{\gamma})^2,$$

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

Following the previous derivations, the null hypothesis $H_0 : \xi(\hat{Y}_{X' \cup X_k}) = \xi(\hat{\pi}_{X'})$ can be tested by the test statistic:

$$Z = \frac{\hat{\gamma}}{\sqrt{\widehat{Var(\hat{\gamma})}}} \rightarrow N(0,1) \tag{12}$$

and, for a given selected significance level α, a rejection region for the null hypothesis H_0 can be defined as $|Z| \geq z_{\frac{\alpha}{2}}$.

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 $m = 1, \ldots, M$ be the considered population groups and let K the number of the available predictors. We denote with $v^{SL}_{mX_k}$ the Shapley-Lorenz value associated with the k-th predictor in the m-th 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 $k^* = 1, \ldots, k$ 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 V^{SL*}_{M} defined as $V^{SL*}_{M} = \left\{v^{SL*}_{1}, \ldots, v^{SL*}_{m}, \ldots, v^{SL*}_{M}\right\}$, where $v^{SL*}_{m} = v^{SL}_{mX_1} + \ldots + v^{SL}_{mX_{k^*}}$ represents the sum of the Shapley-Lorenz values related to the predictors $X_1, \ldots, X_{k^*}$.

The Gini coefficient can be applied to the vector V^{SL*}_{M}, 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:*

$$\textit{Fair-Score} = 1 - LZ(V_M^{SL*}), \tag{13}$$

where $LZ(V_M^{SL*})$ denotes the Lorenz Zonoid (Gini coefficient) computed on the vector V_M^{SL*} 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 $g = 1, \ldots, G$ 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. $V_G^{SL*} = \left\{v_1^{SL*}, \ldots, v_g^{SL*}, \ldots, v_G^{SL*}\right\}$, where $v_g^{SL*} = v_{gX_1}^{SL} + \ldots + v_{gX_{k^*}}^{SL}$ represents the sum of the Shapley-Lorenz values related to the predictors $X_1, \ldots, X_{k^*}$.

Definition 4 *Sustainability score. The score for sustainability can then be defined as:*

$$\textit{Sust-Score} = 1 - LZ(V_G^{SL*}), \tag{14}$$

where $LZ(V_G^{SL*})$ indicates the Lorenz Zonoid (Gini coefficient) calculated on the vector V_G^{SL*} 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 [7] and in [8] 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 [9], 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 [8] and [9], the time series for Oil, Gold and SP500 prices are taken into account as candidate financial explanatory variables. In line with [7], 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 Figures 2-7, respectively.

Specifically, from Figure 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 (Figure 3), the prices of Gold and Oil (Figures 4 and 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.

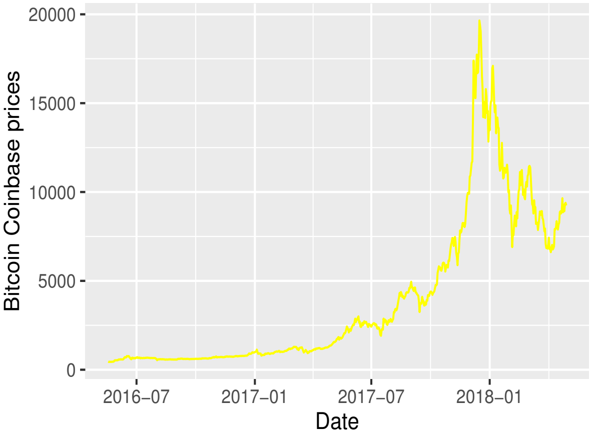

Figure 2: Bitcoin prices

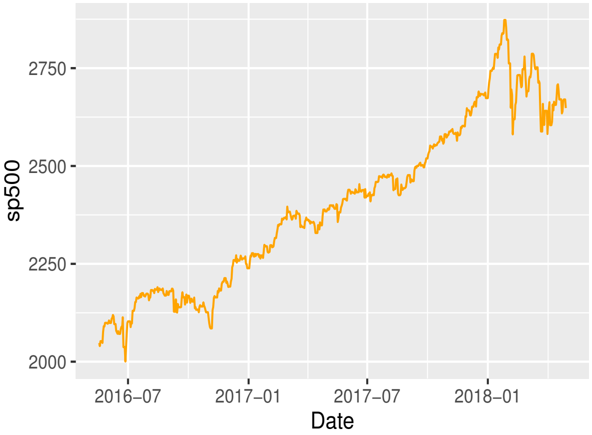

Figure 3: SP500 prices

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

To better understand the dynamics reported in Figures 2-7, some summary statistics are reported in Table 1.

The results in Table 1 highlight that the mean values, as well as the standard deviations 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

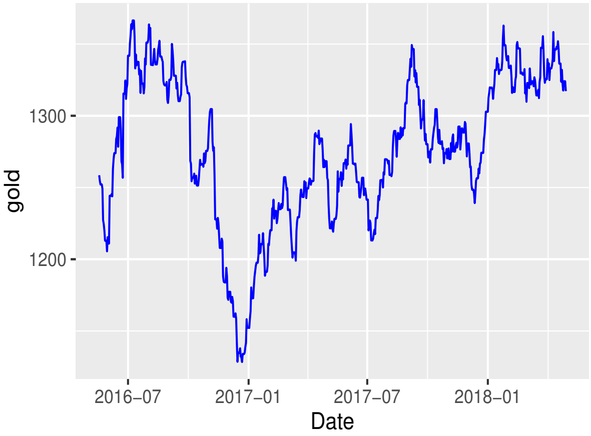

Figure 4: Gold prices

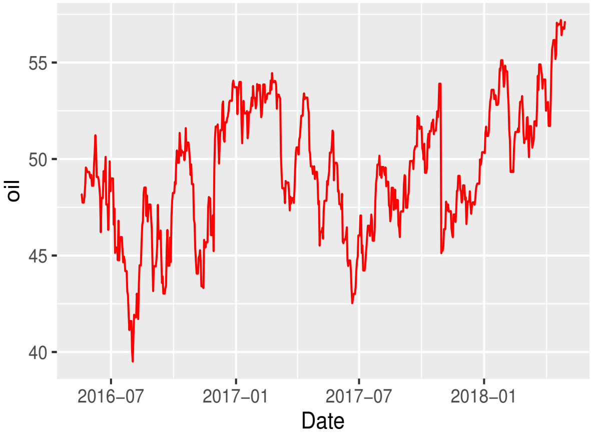

Figure 5: Oil prices

variation (cv) is computed and displayed in Table 1. The findings show that the exchange rates are much less volatile than the bitcoin and classical asset 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.

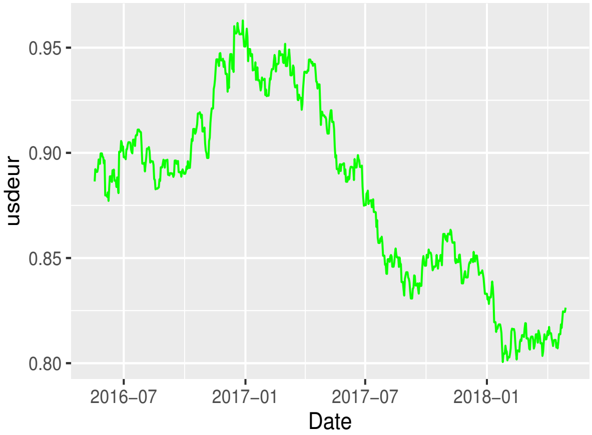

Figure 6: USD/EUR exchange rate

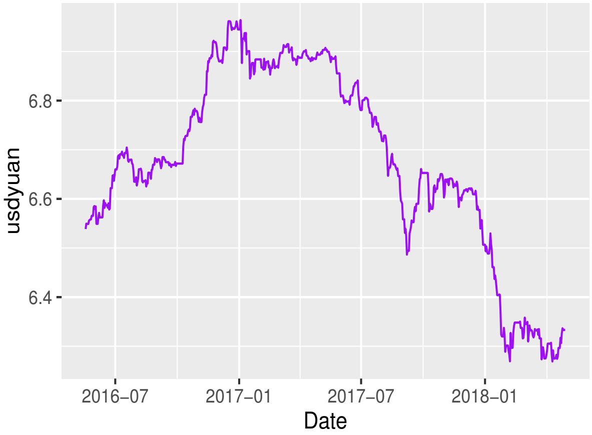

Figure 7: USD/YUAN exchange rate

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. [18]).

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. Figures 2-7 show

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**	**Max**
Coinbase bitcoin	3919.05	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
USD/Eur	0.88	0.04	0.05	0.80	0.96
USD/Yuan	6.68	0.19	0.03	6.27	6.96

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. Figure 8 shows the results of such assessment, in graphical format.

Figure 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.

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.

Figure 8 (b) shows the accuracy score of the selected model, in terms of Lorenz Zonoid. The 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 decile. The result is shown in Figure 8 (c).

Figure 8 (c) shows that, as expected, the predictions worsen, although

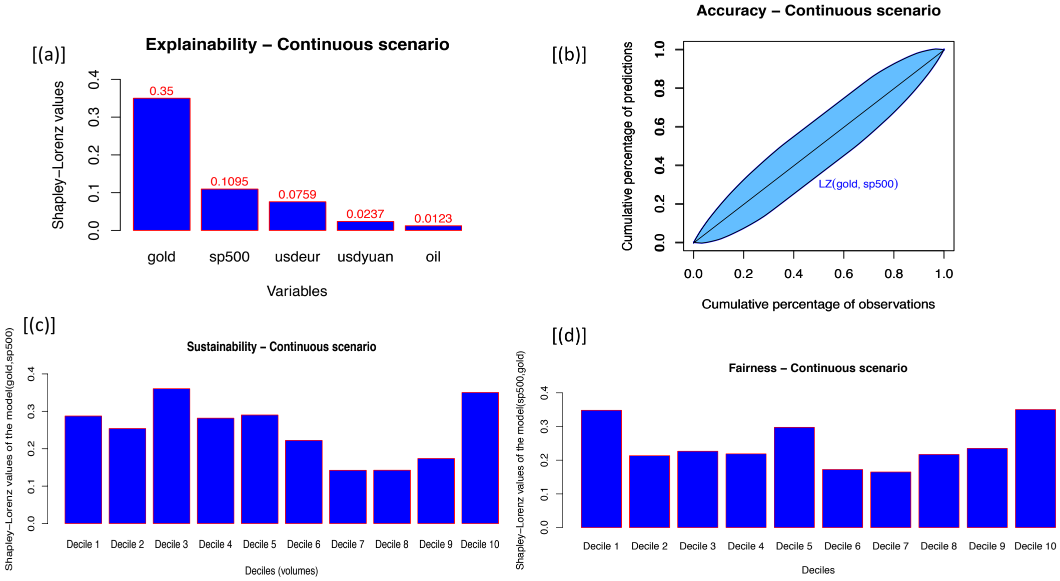

Figure 8: S.A.F.E.ty assessment of the neural network model for bitcoin price returns.

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 depends 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 decile. The result is shown in Figure 8 (d).

Figure 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. Figure 9 shows the results of our S.A.F.E.ty assessment, in graphical format.

From Figure 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.

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 Figure 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 Figures 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.

4 Conclusions

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

To this aim, we have extended the application of Lorenz Zonoids to obtain measurement tools for the Sustainability, Accuracy, Fairness and Explainability, as key trustworthiness criteria.

By means of an easily downloadable datset of bitcoin prices, and related

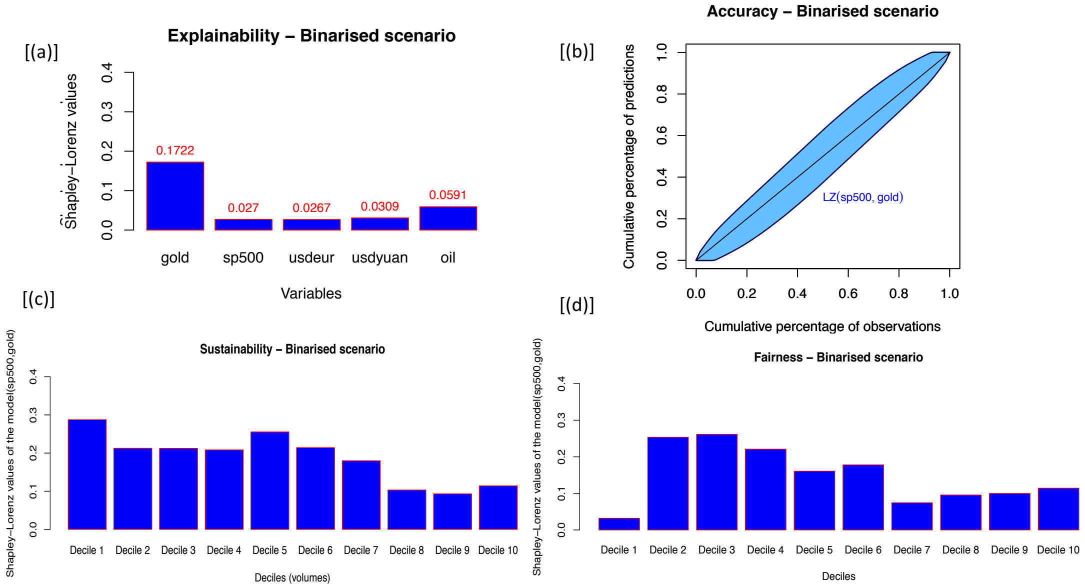

Figure 9: S.A.F.E.ty assessment of the neural network model for bitcoin price returns.

candidate predictors, we have provided a practical demonstration of how to implement and interpret the proposed metrics.

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.

References

[1] Bracke, P., Datta, A., Jung, C., & Shayak, S. (2019). Machine learning explainability in finance: an application to default risk analysis.

Staff Working Paper No. 816, Bank of England.

[2] Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020). Explainable AI in Credit Risk Management. Front Artif Intell, 3(26), 1-5. doi: 10.3389/frai.2020.00026.

[3] DeLong, E.R. and DeLong, D.M., & Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics, 44(3), 837-845.

[4] F. Diebold and R. Mariano, "*Comparing Predictive Accuracy*", J. Bus. Econ. Stat. 13 (1995), no 3, 253-263.

[5] European Commission (2020). On Artificial Intelligence - A European approach to excellence and trust. White Paper, European Commission, Brussels, 19-02-2020.

[6] Gini C. (1936). On the Measure of Concentration with Special Reference to Income and Statistics, Colorado College Publication, General Series No. 208, 73?79.

[7] Giudici P, Abu-Hashish I. What determines bitcoin exchange prices? A network VAR approach, Financ Res Lett 2019; 28:309-318. doi: 10.1016/j.frl.2018.05.013.

[8] Giudici P, Raffinetti E. Lorenz Model Selection. J Classif 2020; 37:754-768. https://doi.org/10.1007/s00357-019-09358-w.

[9] Giudici P, Raffinetti E. Shapley-Lorenz eXplainable Artificial Intelligence. Expert Syst Appl 2021;167:1-9. https://doi.org/10.1016/j.eswa.2020.114104.

[10] Hand D, Mannila H, Smyth P (2001) Principles of data mining. Adaptive Computation and Machine Learning Series. MIT Press.

[11] Hoeffding W. A class of statistics with asymptotically normal distribution. Ann Math Stat 1948;19:293-325. https://doi.org/10.1214/aoms/1177730196.

[12] Joseph A. Parametric inference with universal function approximators, https://www.bankofengland.co.uk/working-paper/2019/shapley-regressions-a-framework-for-statistical-inference-on-machine-learning-models; 2019 [accessed 31 October 2022].

[13] Koshevoy G, Mosler K. The Lorenz Zonoid of a Multivariate Distribution. J Am Stat Assoc 1996;91:873-882. https://doi.org/10.1080/01621459.1996.10476955.

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

[15] Lorenz MO. Methods of measuring the concentration of wealth. Publications of the American Statistical Association 1905;70:209-219. https://doi.org/10.1080/15225437.1905.10503443.

[16] Schechtman E, Yitzhaki S, Artsev Y. The similarity between mean-variance and mean Gini: Testing for equality of Gini correlations. In: Lee CF, Lee AC, editors. Airiti Press; 2008, p. 97-122. Advances in Investment Analysis and Portfolio Management. https://doi.org/10.3390/e22040447.

[17] Shapley LS. A value for n-person games. In: Kuhn H, Tucker A, editors. Contributions to the Theory of Games II. Princeton University Press: Princeton; 1953, p. 307-317.

[18] Tsay RS. Analysis of Financial Time Series, 2nd edition. Wiley; 2005.