Expert Systems With Applications xxx (xxxx) xxx

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Shapley-Lorenz eXplainable Artificial Intelligence

Paolo Giudici [a,*], Emanuela Raffinetti [b]

[a] Department of Economics and Management, University of Pavia, Via San Felice 5, 27100 Pavia, Italy
[b] Department of Economics, Management and Quantitative Methods, University of Milan, Via Conservatorio 7, 20122 Milano, Italy

ARTICLE INFO

Keywords:
Shapley values
Lorenz Zonoids
Predictive accuracy

ABSTRACT

Explainability of artificial intelligence methods has become a crucial issue, especially in the most regulated fields, such as health and finance. In this paper, we provide a global explainable AI method which is based on Lorenz decompositions, thus extending previous contributions based on variance decompositions. This allows the resulting Shapley-Lorenz decomposition to be more generally applicable, and provides a unifying variable importance criterion that combines predictive accuracy with explainability, using a normalised and easy to interpret metric. The proposed decomposition is illustrated within the context of a real financial problem: the prediction of bitcoin prices.

1. Introduction

The growing availability of data and computational power allows to develop machine learning models that are highly predictive. On the other hand, the consideration of the possible adverse consequences on activities that have a high societal impact has led policy makers and regulators to a degree of suspicion towards AI applications. To foster innovations while protecting the society, consensus is emerging on the development of eXplainable AI (XAI) methods, that is, methodologies able to make machine learning models interpretable and, therefore, understood, particularly in terms of causal discovery.

Indeed, in the recent years, the increasing diffusion of artificial intelligence applications and products has led policy makers and regulators to demand the underlying machine learning models to be explainable, so that human users could understand them: see, for example, the recent paper by European Commission (2020). This requirement is particularly evident in highly regulated economic sectors, such as health and finance.

In line with the policy requirements, researchers have recently addressed the issue of how a machine learning model can be made explainable. Existing papers address the contents to different explanation classes. A detailed review of these methods can be found in Guidotti et al. (2018). In this paper, the focus is only on two approaches: global explanations and local explanations. This because our proposal is the result of the combination of local and global explanations. While global explanations describe the model as a whole, in terms of which explanatory variables most determine its predictions, for all the statistical units, local explanations aim at interpreting individual predictions, at the single statistical unit level (for a recent review and comparison, see e.g. Aas, Jullum, & Loland, 2020; Joseph, 2019; Molnar, 2020). 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 a remarkable relevance due to its attractive characteristics. 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 possible sets of features. Note that to obtain reliable explanations, the Shapley value method resorts to all the features. The advantage of Shapley values, over alternative XAI methods, is that they can be used to measure the contribution of each explanatory variable for each point prediction of a machine learning model, regardless of the underlying model itself (see e.g. Lundberg & Lee, 2017; Strumbelj & Kononenko, 2010). In other words, Shapley based XAI are model agnostic so that, differently from the model specific approaches, their interpretation tools are not limited to their respective model classes or data, allowing generality of application and personalisation of their results (they can explain any single point prediction) to be achieved.

Our purpose is to combine the interpretability power of the local Shapley value approach with a more robust global approach, as in Owen and Prieur (2017) and Song, Nelson, and Staum (2016). To this aim, we apply the Shapley value game theoretic approach to Lorenz Zonoid model accuracy tool, recently proposed by Giudici and Raffinetti (2020). In such a way, the advantages associated with the local approach based on the Shapley values are exploited together with the properties of the

* Corresponding author.
E-mail addresses: paolo.giudici@unipv.it (P. Giudici), emanuela.raffinetti@unimi.it (E. Raffinetti).

https://doi.org/10.1016/j.eswa.2020.114104
Received 10 May 2020; Received in revised form 29 July 2020; Accepted 6 October 2020
Available online 16 October 2020

Lorenz Zonoids, giving rise to a global approach which fulfills the interpretability requirement.

On the graphical view point, the Lorenz Zonoids can be seen as a generalisation of the ROC curve in a multidimensional setting. Moreover, in one-dimensional setting, the Lorenz Zonoid is related to the AUROC (Area Under the ROC curve) measure. Therefore, our proposal has the advantage of combining predictive accuracy and explainability performance into one single diagnostics, as highlighted in Giudici and Raffinetti (2020). Furthermore, the nature of Lorenz Zonoids allows them to be easily replicated to any subset of the available units, allowing the diagnostics to be easily applied at any desired local level.

The main contributions of our 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 goodness of fit.

Our proposal lies within the field of explainable AI methods. It extends the global decompositions of Owen and Prieur (2017) and Song et al. (2016), based on the (euclidean) variance decomposition, to a decomposition based on Lorenz Zonoids. The Lorenz Zonoid decomposition presents similarities with the classical variance decomposition. Both the approaches aim to detect the variables which mainly impact the phenomenon of interest. Nevertheless, differently from the classical variance decomposition, the Lorenz Zonoid decomposition is based on the mutual distance between all observations, rather than deviations from the mean and, therefore, is more robust to outlying observations. These features make the Lorenz Zonoid decomposition a promising tool for further extensions in the AI framework, addressed to the assessment of the contribution related to each explanatory variable in terms of the explained mutual variability. As discussed by Giudici and Raffinetti (2020), this methodology appears more generally applicable and directly interpretable within a predictive accuracy context, differently from the approach of Joseph (2019), based on a linear regression approximation.

The expression of our obtained Shapley-Lorenz decomposition also shows that it can be considered as a natural extension of the standard Shapley approach, as it can be calculated not only at the global but also at the local level, providing, in both cases, a normalised measure that can be interpreted within the ROC framework.

The paper is organised as follows. In Section 2 we provide some background on Shapley values. In Section 3 we present our proposal. In Section 4 we exemplify our proposal in the context of a real application that concerns the prediction of bitcoin prices. Section 5 concludes with some final remarks.

2. Background

Shapley values were originally proposed as a pay-off concept from cooperative game theory (Shapley, 1953). Note that the concept of "pay-off" in XAI corresponds to the model prediction, as well described in the papers by Joseph (2019) and Lundberg and Lee (2017).

Shapley values represent the average of the marginal contributions of the players associated with all their possible orders, where, for "order", we intend all the possible orders of players' arrivals to the coalition. The orders are equally likely and, in each order, each player gets his marginal contribution from the coalition he joins to. As discussed by Joseph (2019), Shapley values play a crucial role in improving machine learning model explainability. They allow to evaluate the learned functional forms of a model without having to specify them ex ante.

More generally, Shapley values fulfill a number of useful properties that allow to better understand how the model uses its features to provide a reliable response in a complex decision making process. For example, the sum of the Shapley values is the model accuracy; they are equal for features with the same importance; in a linear model, the Shapley value of a feature is expressed as the linear combination of its Shapley values across the model.

Formally, let $i = 1, \ldots, n$ be a statistical unit, whose (multivariate) characteristics Y_i are to be predicted (on a "test set") with a machine learning model (educated on a "training set"), so that an automated action (say, $a(Y_i)$) is taken.

Let $\widehat{Y}_i^l = \widehat{f}^l(X_i)$ indicate the predicted value for the response vector Y_i, based on an explanatory vector of characteristics X_i, obtained with the machine learning model l. For ease of notation, we drop the suffix l henceforth.

As discussed by Bussmann, Giudici, Marinelli, and Papenbrock (2020), the Shapley value based approach can be developed by using the SHAP (SHapley Additive exPlanations) computational framework (see, e.g. Lundberg & Lee, 2017). This approach differs from the GAM (Generalized Additive Models) approach described by Lou, Caruana, and Gehrke (2012). While the GAM method explicitly decomposes the model into linear combinations of simple models trained by a single explanatory variable, the Shapley value approach decompose the overall model into linear combinations of all the model configurations trained by all the possible combination of the available explanatory variables.

A machine learning model can be decomposed into functions of the additional individual components of x_i (the feature variables) according to a function ϕ as follows:

$$\phi\left(\widehat{f}\left(X_i\right)\right) \equiv \phi_0 + \sum_{k=1}^{K} \phi_k\left(X_i\right), \quad \forall i = 1, \ldots, n, \tag{1}$$

where: k indicates a single feature variable; K denotes the total number of available explanatory variables; n is the total number of units to be predicted; $\phi \in \mathbb{R}^K$; $\phi_k \in \mathbb{R}$. The local functions $\phi_k(X_i)$ are the Shapley values.

Note that linear machine learning models (such as regression models) fulfill this requirement. As shown by Joseph (2019), a linear model satisfies the following:

$$\phi\left(\widehat{f}\left(X_i\right)\right) \equiv \widehat{\beta}_0 + \sum_{k=1}^{K} \widehat{\beta}_k X_{ik}, \tag{2}$$

in which $\phi_0 = \widehat{\beta}_0$ and $\sum_{k=1}^{K} \phi_k(X_i) = \sum_{k=1}^{K} \widehat{\beta}_k X_{ik}$.

Starting from the previous observation, Joseph (2019) proposed to regress the response values on the individual Shapley values to obtain a linear approximation to a machine learning model. While this proposal is tempting, as it provides local explanations which can be statistically tested, it may lead to a highly parameterised model, driven by a computationally expensive procedure. This because the expression in (2) has to be considered for possible subsets of the K available variables, as in a regular model selection procedure.

When referring to a machine learning model, the players of a cooperative game, aimed at generating a pay-off, are the K explanatory variables that can be included in the model and each model is a combination of several variables, which thus "cooperate" towards the predictions $\widehat{f}(x_i)$. Following Lundberg and Lee (2017) and Strumbelj and Kononenko (2010), and using a notation coherent with that considered for the construction of our proposal, the marginal contribution of a variable X_k, ($k = 1, \ldots, K$) can be expressed in the form of Shapley values as

$$\phi\left(\widehat{f}(X_i)\right) = \sum_{X' \subseteq \mathscr{C}(X) \setminus X_k} \frac{|X'|!(K - |X'| - 1)!}{K!} \left[\widehat{f}(X' \cup X_k)_i - \widehat{f}(X')_i\right]. \tag{3}$$

In Eq. (3): $\mathscr{C}(X) \setminus X_k$ is the set of all the possible model configurations which can be obtained with $K-1$ variables, excluding variable X_k; $|X'|$ denotes the number of variables included in each possible model; $\widehat{f}(X' \cup X_k)_i$ and $\widehat{f}(X')_i$ are the predictions associated with all the possible

model configurations including variable X_k and excluding variable X_k, both calculated for the unit i. The quantity within the squared parentheses defines the contribution of variable X_k to the model prediction, for any single unit.

Given the challenging computational efforts needed to calculate the marginal contribution of each variable, especially when K is large, Lundberg and Lee (2017) and Strumbelj and Kononenko (2010) have proposed computational methods to approximate Shapley values with similarly additive feature methods which possess a specified set of properties, such as local accuracy, missingness and consistency.

A remarkable characteristic of the obtained Shapley values approach is that they provide the explanation of the additional importance of each variable for each individual unit. This helps to explain the nature of the contribution of each variable but, on the other hand, it does not explain whether the same variable should be maintained in the model, in a more parsimonious version which, according to Occam's razor principle, improves goodness of fit and interpretation.

Indeed, the drawback of Shapley based XAI methods lies in their very power: being designed to understand point predictions, they may be highly unstable, in the presence of data anomalies, such as fake data, missing data or outliers. In relation with this, they are not suited to understand which variables are important, at the overall level. Although Shapley values can be summed over point predictions, to give an "overall" measure of importance of a single variable, this simple measure leads to compensation, excessive leverage of single observations and, above all, the lack of a normalised measure to assess the relative importance of each variable contribution.

This explains why the tasks of establishing model predictive accuracy for explainable machine learning models based on Shapley values are left to more classic model comparison tools, such as pairwise statistical tests, when possible or, in the more general machine learning context, to cross-validation tools, such as the Receiver Operating Characteristics (ROC) Curve and the corresponding AUROC or Gini value (see e.g., Guégan & Hassani, 2018).

It may be the case that a variable which is highly explainable for most individual predictions is not included into the "best" model that corresponds to the highest Area Under the ROC Curve (AUROC). Conversely, a model selected in terms of best AUROC may contain variables that do not differentiate between individual predictions and, therefore, are not explainable at the local level.

To reconcile the two views (predictive accuracy and local explainability) we propose to develop a Shapley based framework that decomposes predictive accuracy, rather than individual predictions. And that could, possibly, be localised. This is our main contribution.

3. Proposal

To achieve our aim we exploit a model selection measure, recently introduced by Giudici and Raffinetti (2020), which is based on the employment of Lorenz Zonoids and on a mutual notion of variability. The Lorenz Zonoid-based measure fulfills some attractive properties: it is akin to the well known Receiver Operating Curve (ROC), robust to the presence of outlying observations and independent on the nature of the response variable.

Lorenz Zonoids were introduced by Koshevoy and Mosler (1996) as a generalization of the Lorenz curve in d dimensions. The same authors showed that, when $d = 1$, the Lorenz Zonoid corresponds with the well known Gini coefficient which, in turn, is related to the Area Under the ROC Curve.

Suppose to consider a response variable Y and a set of explanatory variables $X_1, \ldots, X_j, \ldots, X_h$, with $j = 1, \ldots, h$. To evaluate the relationships between Y and the $X_1, \ldots, X_h$ explanatory variables, a machine learning model can be applied, and the associated predicted values, denoted with $\widehat{Y}_{X_1,\ldots,X_h}$, are obtained. The Lorenz Zonoid of Y and $\widehat{Y}_{X_1,\ldots,X_h}$ can be defined by (see, e.g. Giudici and Raffinetti, 2020):

$$LZ_{d=1}\left(Y\right) = \frac{2Cov(Y, r(Y))}{n\mu} \text{ and } LZ_{d=1}\left(\widehat{Y}_{X_1,\ldots,X_h}\right) = \frac{2Cov\left(\widehat{Y}_{X_1,\ldots,X_h}, r\left(\widehat{Y}_{X_1,\ldots,X_h}\right)\right)}{n\mu}, \tag{4}$$

where n is the total number of observations, μ is the response variable Y mean value, $r(Y)$ and $r(\widehat{Y}_{X_1,\ldots,X_h})$ are the rank scores corresponding to the Y and $\widehat{Y}_{X_1,\ldots,X_h}$ variables. Given a sample data of size n, formulas in (4) can be reformulated as:

$$LZ_{d=1}\left(y\right) = \frac{2Cov(y, r(y))}{n\bar{y}} \text{ and } LZ_{d=1}\left(\widehat{y}_{y_1,\ldots,y_h}\right) = \frac{2Cov\left(\widehat{y}_{x_1,\ldots,x_h}, r\left(\widehat{y}_{x_1,\ldots,x_h}\right)\right)}{n\bar{y}}, \tag{5}$$

where y and $\widehat{y}_{x_1,\ldots,x_h}$ are the vectors of the observed and predicted values, $r(y)$ and $r(\widehat{y}_{x_1,\ldots,x_h})$ are the ranks of the observed and predicted values, and $\bar{y}$ is the sample mean.

In Giudici and Raffinetti (2020), the Lorenz Zonoids were exploited giving rise to new dependence measures suitable in assessing the contribution of each explanatory variable to the predictive power of a model. Specifically, a Marginal Gini Contribution (*MGC*) measure, allowing to measure the absolute explanatory power of any single covariate,[1] and a Partial Gini Contribution measure (*PGC*), allowing to measure the additional contribution of a new covariate to an existing model, were developed as follows.

Let X_j be one of the h explanatory variables ($j = 1, \ldots, h$). The marginal contribution provided by a single covariate X_j is given by:

$$MGC_{Y|X_j} = \frac{LZ_{d=1}\left(\widehat{Y}_{X_j}\right)}{LZ_{d=1}(Y)} = \frac{Cov\left(\widehat{Y}_{X_j}, r\left(\widehat{Y}_{X_j}\right)\right)}{Cov(Y, r(Y))}. \tag{6}$$

Let $\widehat{Y}_{X_1,\ldots,X_h}$ and $\widehat{Y}_{X_1,\ldots,X_{h-1}}$ be the predicted values provided by a full model, including all the covariates, and a reduced model, excluding covariate X_h. The additional contribution related to the inclusion of covariate X_h can be determined as

$$PGC_{Y,X_h|X_1,\ldots,X_{h-1}} = \frac{LZ_{d=1}\left(\widehat{Y}_{X_1,\ldots,X_h}\right) - LZ_{d=1}\left(\widehat{Y}_{X_1,\ldots,X_{h-1}}\right)}{LZ_{d=1}\left(Y\right) - LZ_{d=1}\left(\widehat{Y}_{X_1,\ldots,X_{h-1}}\right)}. \tag{7}$$

We remark that, when the Y response variable is continuous, and the machine learning model is linear, the marginal contribution provided by a single covariate X to an existing model, in Eq. (7), simplifies to the well known variance decomposition of the multiple correlation coefficient, R^2:

$$R^2_{(Y|X_1,\ldots,X_h)} = \sum_{j=1}^{h} R^2_{Y,X_j|X_{i<j}}\left(1 - R^2_{Y|X_1,\ldots,X_{j-1}}\right), \tag{8}$$

where: $R^2_{(Y|X_1,\ldots,X_h)}$ denotes the multiple correlation coefficient (the Y variability explained by all the involved h covariates); $R^2_{Y,X_j|X_{i<j}}$ denotes the partial correlation coefficient (the variability of Y, additionally explained by the *j-th* explanatory variable, after the previous $i < j$ variables, whose contribution is given by $R^2_{(Y|X_1,\ldots,X_{j-1})}$).

When the covariates are independent, we obtain, as a further special case, that:

[1] Note that to the term "covariate" is employed as a synonym of the term "explanatory variable". The two words will be then used interchangeably.

$$R^2_{(Y|X_1,\ldots,X_h)} = \sum_{j=1}^{h} R^2_{Y,X_j}. \quad (9)$$

The previous remark suggests that, using a Lorenz Zonoid decomposition, we can extend the global explanations of machine learning models proposed by Owen and Prieur (2017) and Song et al. (2016). While the latter Authors employed a variance decomposition approach to explainable machine learning, dealing with the issue of dependent covariates, we extend the approach from variance decomposition to Lorenz Zonoid decomposition, obtaining a simpler and more versatile approach.

In line with the need of diagnosing both predictive accuracy and explainability, we now combine the Lorenz Zonoid, aimed at evaluating predictive accuracy in a rather general context, with the Shapley value approach, aimed at obtaining individual unit explanations.

The main intuition of our proposal is the following. Shapley proposed to employ game theory with pay-offs that are given by:

$$p_{off}\left(X_i^k\right) = \widehat{f}(X \cup X_k)_i - \widehat{f}(X)_i, \quad (10)$$

for any statistical unit i.

We propose to apply game theory with pay-offs that are given by the numerator of the *PGC* measure:

$$p_{off}\left(X^k\right) = LZ_{d=1}\left(\widehat{Y}_{X_1,\ldots,X_k}\right) - LZ_{d=1}\left(\widehat{Y}_{X_1,\ldots,X_{k-1}}\right), \quad (11)$$

for a set of statistical units $(i = 1,\ldots,n)$.

The resulting expression, that we call Shapley-Lorenz decomposition, allows to identify the contribution of each explanatory variable, not in terms of the differential contribution to the locally predicted values (as with standard Shapley values), but in terms of the differential contribution to the global predictive accuracy.

We now proceed with the mathematical derivation of the Shapley-Lorenz decomposition.

First, we replace $LZ_{d=1}(\cdot)$ in place of $\widehat{f}(\cdot)$ in the Shapley expression in (3), and obtain that the marginal contribution associated with the additional variable X^k is equal to

$$LZ^{X_k}_{d=1}\left(\widehat{Y}\right) = \sum_{X' \subseteq \mathscr{C}(X)\setminus X_k} \frac{|X'|!(K-|X'|-1)!}{K!}\left[LZ_{d=1}\left(\widehat{Y}_{X'\cup X_k}\right) - LZ_{d=1}\left(\widehat{Y}_{X'}\right)\right], \quad (12)$$

where $LZ_{d=1}(\widehat{Y}_{X'\cup X_k})$ and $LZ_{d=1}(\widehat{Y}_{X'})$ describe the (mutual) variability explained by the models including the $X' \cup X_k$ variables and the X' variables, respectively. Note that $LZ_{d=1}(\widehat{Y}_{X'\cup X_k})$ and $LZ_{d=1}(\widehat{Y}_{X'})$ in Eq. (12) can be expressed as function of the covariance operator, as reported in Appendix Section A1.

As what we observe is indeed a sample of n observations, we need to estimate the population mean μ, with the sample mean, $\bar{y}$. Then, denoting with $\widehat{y}_{X'\cup X_k}$ and $\widehat{y}_{X'}$ the predicted values provided by the model including and excluding the X_k covariate, ordered in non-decreasing sense, the formula in Eq. (12) becomes

$$LZ^{X_k}_{d=1}\left(\widehat{y}\right) = \sum_{X' \subseteq \mathscr{C}(X)\setminus X_k} \frac{|X'|!(K-|X'|-1)!}{K!}\left[LZ_{d=1}\left(\widehat{y}_{X'\cup X_k}\right) - LZ_{d=1}\left(\widehat{y}_{X'}\right)\right]. \quad (13)$$

Through some mathematical manipulations, whose details are contained in Appendix Section A2, Eq. (13) can be re-written as

$$LZ^{X_k}_{d=1}\left(\widehat{y}\right) = \sum_{X' \subseteq \mathscr{C}(X)\setminus X_k} \frac{|X'|!(K-|X'|-1)!}{K!}\left\{\frac{2}{n^2\bar{y}}\left[\sum_{i=1}^{n} i\left(\widehat{y}_{X'\cup X_k}(i) - \widehat{y}_{X'}(i)\right)\right]\right\}, \quad (14)$$

where $\widehat{y}_{X'\cup X_k}(i)$ and $\widehat{y}_{X'}(i)$ are the predicted values for the i-th statistical unit obtained by the model including and excluding the X_k covariate. Comparing Eq. (14) with (3) note, part from the different notation, the similarities between the two expressions. While the standard Shapley decomposition "explains" the covariate contributions at the individual level, the Shapley-Lorenz decomposition "explains" the same contributions at the global level. Indeed, through Eq. (14), a description of the model as a whole, in terms of the explanatory variables mostly determining its prediction, is provided.

Shapley-Lorenz decomposition, differently from standard ones, allows to detect which variables could be eliminated, as unnecessary for model predictions, leading to a more parsimonious structure. Indeed standard Shapley values can be summed across units, leading to "global" variable importance measures which, however, are not normalised within a model accuracy context, as Shapley-Lorenz ones.

In addition, looking at expression (14), note that Shapley-Lorenz decomposition can always be calculated, without loss of generality, to subsets of the n units to be predicted. This leads to a natural "localisation" of the measure, without altering its predictive meaning.

4. Application

In line with our initial discussion, to illustrate our proposal we consider the application of machine learning models in the highly regulated field of finance.

In finance, the notion of XAI is increasingly discussed by public and private institutions, to provide transparent and effective machine learning methods (see, e.g. Arras, Horn, Montavon, Müller, & Samek, 2017; Arrieta et al., 2019). The idea is to introduce a suite of techniques that allows to improve the interpretability of the models while preserving an adequate level of prediction accuracy. This idea has recently led some scholars to promote XAI methods aimed at making both the financial technology risk measurement models interpretable and transparent, and the risks of financial innovations, enabled by the application of AI, sustainable (see, e.g. Bracke, Datta, Jung, & Shayak, 2019; Bussmann et al., 2020). In particular, in Bussmann et al. (2020) an explainable AI model based on similarity networks (Mantegna & Stanley, 1999) and Shapley values is proposed to measure the credit risks associated to the use of AI based credit scoring platforms.

To exemplify our proposal, we apply it to a dataset that has been used to predict bitcoin prices, and their up or downtrends. As illustrated in Giudici and Raffinetti (2020), the available data provide information on the daily bitcoin prices in eight different crypto exchanges, from 18 May, 2016 to 30 April, 2018. For the sake of brevity, we refer to the time series observations on Coinbase prices, which represent the response variable to be predicted by the available financial explanatory variables. Specifically, as candidate financial explanatory variables the time series for Oil, Gold and SP500 prices are taken into account. The choice of such set of variables is related with their economic importance, and with the need to explain our proposal with a model simple enough so that calculations can be clearly understood.

In this application, we will select, as our candidate machine learning model, a linear regression model and we will calculate the Shapley-Lorenz marginal contributions, associated with the inclusion of SP500, Gold and Oil, according to the formula (13). When considering SP500, Gold and Oil as additional explanatory variable, the corresponding marginal contributions can be written in full as follows:

$$\begin{aligned} LZ^{SP500}_{d=1}\left(\widehat{Coinbase}\right) &= \left(1/3\right)\left(LZ\left(\widehat{y}_{SP500,Gold,Oil}\right) - LZ\left(\widehat{y}_{Gold,Oil}\right)\right) \\ &+ \left(1/6\right)\left(LZ\left(\widehat{y}_{SP500,Gold}\right) - LZ\left(\widehat{y}_{Gold}\right)\right) + \left(1/6\right)\left(LZ\left(\widehat{y}_{SP500,Oil}\right) - LZ\left(\widehat{y}_{Oil}\right)\right) \\ &+ (1/3)\left(LZ\left(\widehat{y}_{SP500}\right)\right) \end{aligned}$$

$$LZ^{Gold}_{d=1}\left(\widehat{Coinbase}\right) = \left(1/3\right)\left(LZ\left(\hat{y}_{Gold,SP500,Oil}\right) - LZ\left(\hat{y}_{SP500,Oil}\right)\right) + (1/6)\left(LZ(\hat{y}_{Gold,SP500}) - LZ(\hat{y}_{SP500})\right) + (1/6)\left(LZ(\hat{y}_{Gold,Oil}) - LZ(\hat{y}_{Oil})\right) + (1/3)(LZ(\hat{y}_{Gold}))$$

$$LZ^{Oil}_{d=1}\left(\widehat{Coinbase}\right) = \left(1/3\right)\left(LZ\left(\hat{y}_{Oil,SP500,Gold}\right) - LZ\left(\hat{y}_{SP500,Gold}\right)\right) + (1/6)\left(LZ(\hat{y}_{Oil,SP500}) - LZ(\hat{y}_{SP500})\right) + (1/6)\left(LZ(\hat{y}_{Oil,Gold}) - LZ(\hat{y}_{Gold})\right) + (1/3)(LZ(\hat{y}_{Oil})).$$

For the sake of comparison, we will also consider the variance decomposition associated with the same variables, which holds under the assumption of a linear model. Applying the Shapley formula as before, but replacing Lorenz Zonoid with Partial correlation coefficients, we obtain the following marginal contributions:

$$R^2_{SP500} = \left(1/3\right)\left(R^2_{SP500,Gold,Oil} - R^2_{Gold,Oil}\right) + \left(1/6\right)\left(R^2_{SP500,Gold} - R^2_{Gold}\right) + \left(1/6\right)\left(R^2_{SP500,Oil} - R^2_{Oil}\right) + \left(1/3\right)R^2_{SP500}$$

$$R^2_{Gold} = \left(1/3\right)\left(R^2_{Gold,SP500,Oil} - R^2_{SP500,Oil}\right) + \left(1/6\right)\left(R^2_{Gold,SP500} - R^2_{SP500}\right) + \left(1/6\right)\left(R^2_{Gold,Oil} - R^2_{Oil}\right) + \left(1/3\right)R^2_{Gold}$$

$$R^2_{Gold} = \left(1/3\right)\left(R^2_{Oil,SP500,Gold} - R^2_{SP500,Gold}\right) + \left(1/6\right)\left(R^2_{Oil,SP500} - R^2_{SP500}\right) + \left(1/6\right)\left(R^2_{Oil,Gold} - R^2_{Gold}\right) + \left(1/3\right)R^2_{Oil}.$$

We can also calculate the "standard" global Shapley value for each variable summing, for each variable, its contribution to any single unit prediction. We remark that the result is a measure that, differently from before, is not normalised and, therefore, not easily interpretable.

The results of all the previous calculations are displayed in Table 1. Table 1 shows that, employing the Lorenz-Shapley approach, variable SP500 provides the highest marginal contribution in the prediction of the Coinbase prices (as in Giudici & Abu-Hashish, 2019), while the other two give a minimal contribution. This conclusion is in line with the economic literature, which shows that the bitcoin has reached the status of a speculative asset, that is used to diversify portfolios, being significantly negatively correlated with classic assets such as stock prices summarised by the SP500 index.

The conclusions from the Shapley-Lorenz approach are also quite similar to those obtained with the linear R^2-based Shapley approach. This shows that a non linear machine learning model does not lead to a substantial change of the interpretability that could be drawn from a linear model, applied to the same data. Note that, in general, the Shapley-Lorenz approach has to be preferred to the linear R^2-based Shapley approach, especially in the presence of outlying observations.

Finally, the Global Shapley values, obtained summing the Shapley variable contributions across all units are, as expected, non normalised, and with a sign. While the Global Shapley value fails to tell which variable is most important in terms of the explained variability, the sign is consistent with the previously commented economic finding: the SP500 index is negatively correlated with the Coinbase prices.

The Shapley value approach appears as more intuitive than further typically used AI methods, as shown by the use case about explainability of risk management models described and developed by Bussmann et al. (2020) within the European FINTECH Project (https://www.fintech-h o2020.eu). The use case was indeed presented, in a "human-centric" study, to the regulators of most European countries, and one of the main feedback was that the approach is nice and promising but should be compared to what obtained with "classical" model assessment methods, which is what the Shapley-Lorenz approach provides.

Table 1
Marginal contribution of each explanatory variable in terms of the linear Shapley-Lorenz approach, in terms of the R^2 coefficient and the standard Shapley approach.

Additional covariate (X_k)	$LZ^{X_k}_{d=1}(\widehat{Coinbase})$	$R^2_{X_k}$	Global Shapley
SP500	0.336	0.631	−96377.28
Gold	0.097	0.072	59811.19
Oil	0.075	0.049	−43428.39

5. Conclusions

In this paper we have introduced a novel global explainable AI model, based on the application of the Shapley approach to Lorenz Zonoid.

The proposed decomposition extends those recently proposed in terms of variance decomposition, leading to a variable contribution measure that is more generally applicable, and easier to interpret. In addition, the expression of the marginal contribution shows how global explanations can be mapped to local ones and viceversa.

We believe that our proposal could be quite useful, as it provides a unified criterion to assess both predictive accuracy and explainability of the explanatory variables contained in a machine learning model. In addition, the metric in which the measure is expressed is a normalised one, related to the AUROC and Gini index and, therefore, easier to interpret.

The application of the measure to a financial problem that concerns bitcoin price prediction shows its ease of application, consistency and versatility.

The potential users of our model are, besides academic researchers, AI developers, for compliance and regtech purposes; and policy makers and regulators, for AI certification, monitoring, and suptech purposes.

Future extensions of the research concern, on one hand, the development of a statistical testing procedure, which could add to variable contributions a significance measure. On the other hand, the extensive application to several other application fields.

CRediT authorship contribution statement

Paolo Giudici: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing - review & editing. **Emanuela Raffinetti:** Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft.

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.

Acknowledgments

This research has received funding from the European Union's Horizon 2020 research and innovation program "FIN-TECH: A Financial supervision and Technology compliance training programme" under the grant agreement No 825215 (Topic: ICT-35-2018, Type of action: CSA).

Acknowledges go to the three anonymous reviewers for their valuable comments and suggestions which allowed to improve the paper.

Appendix A

A1. Covariance formulation of Lorenz Zonoids

As shown in Eq. (4), $LZ_{d=1}(\widehat{Y}_{X'\cup X_k})$ and $LZ_{d=1}(\widehat{Y}_{X'})$ in Eq. (12) can be written through the covariance formulation leading to

$$LZ_{d=1}\left(\widehat{Y}_{X'\cup X_k}\right) = \frac{2}{n\mu} Cov\left(\widehat{Y}_{X'\cup X_k}, r\left(\widehat{Y}_{X'\cup X_k}\right)\right) \tag{15}$$

and

$$LZ_{d=1}\left(\widehat{Y}_{X'}\right) = \frac{2}{n\mu} Cov\left(\widehat{Y}_{X'}, r\left(\widehat{Y}_{X'}\right)\right). \tag{16}$$

A2. Derivation of equation in (14)

Given a sample of n observations, formulas in Eqs. (15) and (16) become

$$LZ_{d=1}\left(\widehat{y}_{X'\cup X_k}\right) = \frac{2}{n\overline{y}} Cov\left(\widehat{y}_{X'\cup X_k}, r\left(\widehat{y}_{X'\cup X_k}\right)\right) = \frac{2}{n\overline{y}}\left[\frac{1}{n}\sum_{i=1}^{n} i\widehat{y}_{X'\cup X_k}\left(i\right) - \frac{n(n+1)}{2n}\right] \tag{17}$$

and

$$LZ_{d=1}\left(\widehat{y}_{X'}\right) = \frac{2}{n\overline{y}} Cov\left(\widehat{y}_{X'}, r\left(\widehat{y}_{X'}\right)\right) = \frac{2}{n\overline{y}}\left[\frac{1}{n}\sum_{i=1}^{n} i\widehat{y}_{X'}\left(i\right) - \frac{n(n+1)}{2n}\right]. \tag{18}$$

Inserting the expressions (15) and (16) into (12), we obtain that the marginal contribution of an explanatory variable X_k is a function of:

$$\begin{aligned} LZ_{d=1}\left(\widehat{Y}_{X'\cup X_k}\right) - LZ_{d=1}\left(\widehat{Y}_{X'}\right) &= \frac{2}{n\mu} Cov\left(\widehat{Y}_{X'\cup X_k}, r\left(\widehat{Y}_{X'\cup X_k}\right)\right) - \frac{2}{n\mu} Cov\left(\widehat{Y}_{X'}, r\left(\widehat{Y}_{X'}\right)\right) \\ &= \frac{2}{n\mu}\left[Cov\left(\widehat{Y}_{X'\cup X_k}, r\left(\widehat{Y}_{X'\cup X_k}\right)\right) - Cov\left(\widehat{Y}_{X'}, r\left(\widehat{Y}_{X'}\right)\right)\right], \end{aligned} \tag{19}$$

whose sample version, by resorting to Eqs. (17) and (18), can be obtained as:

$$\begin{aligned} &LZ_{d=1}\left(\widehat{y}_{X'\cup X_k}\right) - LZ_{d=1}\left(\widehat{y}_{X'}\right) = \frac{2}{n\overline{y}}\left[Cov\left(\widehat{y}_{X'\cup X_k}, r\left(\widehat{y}_{X'\cup X_k}\right)\right) - Cov\left(\widehat{y}_{X'}, r\left(\widehat{y}_{X'}\right)\right)\right] \\ &= \frac{2}{n\overline{y}}\left[\frac{1}{n}\sum_{i=1}^{n} i\widehat{y}_{X'\cup X_k}(i) - \frac{1}{n}\sum_{i=1}^{n} i\widehat{y}_{X'}(i)\right] = \frac{2}{n\overline{y}}\left[\frac{1}{n}\left(\sum_{i=1}^{n} i\widehat{y}_{X'\cup X_k}(i) - \sum_{i=1}^{n} i\widehat{y}_{X'}(i)\right)\right] \\ &= \frac{2}{n^2\overline{y}}\left[\sum_{i=1}^{n} i\left(\widehat{y}_{X'\cup X_k}(i) - \widehat{y}_{X'}(i)\right)\right]. \end{aligned} \tag{20}$$

The previous quantity defines the contribution of variable X_k to a particular model configuration, with X' the considered explanatory variables. It is the analog of the quantity in squared parentheses in the Shapley Eq. (3). Comparing the two quantities, note that the Shapley-Lorenz decomposition is indeed a function of the individual Shapley differences. A function that, differently from the pure sum of the individual Shapley values, considers a normalised sum of their cumulative intensities.

Completing (20) with the remaining part of Eq. (13), that takes into account all possible model configurations, the Shapley-Lorenz marginal contribution of a covariate X_k is finally obtained as:

$$LZ^{X_k}_{d=1}\left(\widehat{y}\right) = \sum_{X'\subseteq \mathscr{C}(X)\setminus X_k} \frac{|X'|!(K-|X'|-1)!}{K!}\left\{\frac{2}{n^2\overline{y}}\left[\sum_{i=1}^{n} i\left(\widehat{y}_{X'\cup X_k}(i) - \widehat{y}_{X'}(i)\right)\right]\right\},$$

which corresponds to expression in (14).

References

Aas, K., Jullum, M., & Loland, A. (2020). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. arXiv preprint arXiv:1903.10464.

Arras, L., Horn, F., Montavon, G., Müller, K.-R., & Samek, W. (2017). "What is relevant in a text document?": An interpretable machine learning approach. *PLoS One, 12*(8), 1–23. https://doi.org/10.1371/journal.pone.0181142

Arrieta, A. B., Dríaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2019). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. arXiv preprint arXiv:1910.10045.

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.

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020). Explainable AI in credit risk management. *Frontiers in Artificial Intelligence, 3*(26), 1–5. https://doi.org/10.3389/frai.2020.00026

European Commission. (2020). *On artificial intelligence – A European approach to excellence and trust*. White Paper, European Commission, Brussels, 19-02-2020.

Giudici, P., & Abu-Hashish, I. (2019). What determines bitcoin exchange prices? A network VAR approach. *Finance Research Letters, 28*, 309–318. https://doi.org/10.1016/j.frl.2018.05.013

Giudici, P., & Raffinetti, E. (2020). Lorenz model selection. *Journal of Classification*. https://doi.org/10.1007/s00357-019-09358-w

Guégan, D., & Hassani, B. (2018). Regulatory learning: How to supervise machine learning models? An application to credit scoring. *The Journal of Finance and Data Science, 4*(3), 157–171. https://doi.org/10.1016/j.jfds.2018.04.001

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black-box models. *ACM Computing Surveys (CSUR), 51*(5), 1–42. https://doi.org/10.1145/3236009

Joseph, A. (2019). *Shapley regressions: A framework for statistical inference in machine learning models*. Staff Working Paper No. 784, Bank of England.

Koshevoy, G., & Mosler, K. (1996). The Lorenz Zonoid of a multivariate distribution. *Journal of the American Statistical Association, 91*(434), 873–882. https://doi.org/10.2307/2291682

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

Lundberg, S. M., & Lee, S. (2017). *A unified approach to interpreting model predictions*. arXiv preprint arXiv:1705.07874.

Mantegna, R. N., & Stanley, H. E. (1999). *Introduction to econophysics: Correlations and complexity in finance*. Cambridge University Press.

Molnar, C. (2020). *Interpretable machine learning – A guide for making black box models explainable*. Available at URL: https://cristophm.github.io/interpretable-ml-book.

Owen, A. B., & Prieur, C. (2017). On Shapley value for measuring importance of dependent inputs. *SIAM/ASA Journal of Uncertainty Quantification, 5*, 986–1002. https://doi.org/10.1137/16M1097717

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, 307–317.

Song, E., Nelson, B., & Staum, J. (2016). Shapley effects for global sensitivity analysis: Theory and computation. *SIAM/ASA Journal of Uncertainty Quantification, 4*, 1060–1083. https://doi.org/10.1137/15M1048070

Strumbelj, E., & Kononenko, I. (2010). An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research, 11*, 1–18. https://doi.org/10.1145/1756006.1756007