Sustainable, Accurate, Fair and Explainable Machine Learning Models

Paolo Giudici[a] and Emanuela Raffinetti[a]

[a]Department of Economics and Management, University of Pavia, Via San Felice al Monastero, 27100 Pavia (Italy); paolo.giudici@unipv.it, emanuela.raffinetti@unipv.it

Abstract

Machine learning models are currently favouring Artificial Intelligence applications in several fields, such as for instance, in finance. Through the employment of machine learning models, high predictive accuracy is achieved but at the expense of interpretability. The loss of explainability represents a crucial issue, especially in regulated industries, as authorities may not validate Artificial Intelligence methods if they are unable to monitor and limit the risks deriving from them. For this reason and according to the proposed regulations, high-risk Artificial Intelligence applications based on machine learning must be "trustworthy" and fulfill a set of basic requirements. In this paper, we propose a methodology based on Lorenz Zonoids to assess whether a machine learning model is S.A.F.E.: Sustainable, Accurate, Fair and Explainable.

Keywords: Artificial Intelligence, Lorenz Zonoids, S.A.F.E. requirement

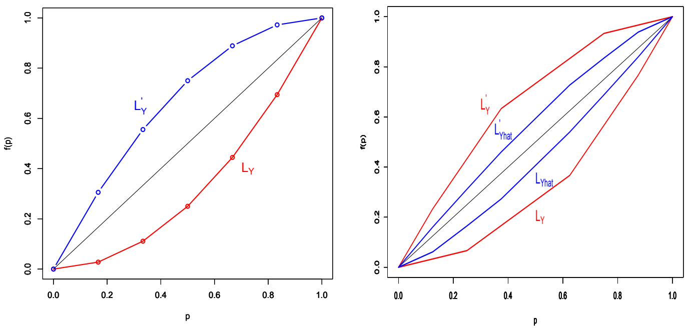
1. Introduction

Data driven Artificial Intelligence (AI), boosted by the availability of Machine Learning (ML) models, is rapidly expanding and changing financial services. ML models typically achieve a high accuracy, at the expense of an insufficient explainability (see e.g. [2], [1]). Moreover, according to the proposed regulations, high-risk AI applications based on machine learning must be "trustworthy", and comply with a set of further 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, the multidimensional extension of the Gini coefficient, 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 openly downloadable dataset, that concerns financial prices, to make our proposal easily reproducible.

2. Methodology

Lorenz Zonoids were originally proposed by [6] as a generalisation of the Lorenz 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



[(a)] [(b)]

Figure 1: [a] The Lorenz Zonoid; [b] The inclusion property: $LZ(\hat{Y}) \subset LZ(Y)$

within a nation or a social group (see, e.g [3] and [7]). 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 (L_Y) and the dual Lorenz curves ($L_Y^{'}$) (see [7]), whose graphical representations are provided in Fig. 1 [a].

The Lorenz curve for a variable Y (L_Y), obtained by re-ordering the Y values in non-decreasing sense, has points whose coordinates can be specified as $(i/n, \sum_{j=1}^{i} y_{r_j}/(n\bar{y}))$, 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 ($L_Y^{'}$), obtained by re-ordering the Y values in a non-increasing sense, has points with coordinates $(i/n, \sum_{j=1}^{i} y_{d_j}/(n\bar{y}))$, 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 corresponds to the Lorenz Zonoid, which coincides with the Gini coefficient in the one dimensional case.

From a practical view point, given n observations, the Lorenz Zonoid of a generic variable $\cdot$ is computed through the covariance operator as

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

where $r(\cdot)$ and $E(\cdot)$ are the corresponding rank score and mean value, respectively.

The Lorenz Zonoid fulfills some attractive properties. An important one is the "inclusion" of the Lorenz Zonoid of any set of predicted values $\hat{Y}$ ($LZ(\hat{Y})$) into the Lorenz Zonoid of the observed response variable Y ($LZ(Y)$). The "inclusion property", whose graphical representation is displayed in Fig. 1 [b], 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.

In this paper, we leverage the inclusion property to derive a machine learning feature selection method that, while maintaining a high predictive accuracy, increases explainabiity via parsimony and can also improve both sustainability and fairness. More precisely, we present novel scores for assessing both explainability and accuracy.

Given K predictors, a score for evaluating explainability can be defined as:

$$\textit{Ex-Score} = \frac{\sum_{k=1}^{K} SL_k}{LZ(Y)}, \tag{2}$$

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. It is worth noting that, as illustrated in [5], the Shapley-Lorenz contribution associated with the additional included variable X_k equals to:

$$LZ^{X_k}(\hat{Y}) = \sum_{X' \subseteq \mathcal{C}(X) \setminus X_k} \frac{|X'|!(K - |X'| - 1)!}{K!} [LZ(\hat{Y}_{X' \cup X_k}) - LZ(\hat{Y}_{X'})], \tag{3}$$

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 the $X' \cup X_k$ predictors and only the X' predictors.

In a similar way, and following a cross-validation procedure consisting in splitting the whole dataset into a train and a test set, the accuracy of the predictions generated by a ML model can be derived as:

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

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.

By exploiting the Shapley-Lorenz values and the set of the predictors which allow to ensure a suitable degree of predictive accuracy, appropriate scores for measuring both fairness and sustainability can be formalised.

3. Data

The considered data are described in [4] and are aimed to understand whether and how bitcoin price returns vary as a function of a set of classical financial explanatory variables.

A further investigation of the data was carried out in a work by [5], who introduced a normalised Shapley measure for the assessment of the contribution of each additional predictor, in terms of Lorenz Zonoids.

The data include a time series of daily bitcoin price returns in the Coinbase exchange, as the target variable to be predicted, and the time series of the Oil, Gold and SP500 return prices, along with those of the exchange rates USD/Yuan and USD/Eur, as candidate explanatory variables.

The aim of the data analysis is to employ the proposed S.A.F.E. metrics derived from the Lorenz Zonoid tool as criteria for measuring the SAFEty of a collection of machine learning models, based on the application of neural networks.

For lack of space, we present only the results that concern explainability, in Fig. 2; and accuracy, in Fig. 3. Both are calculated on the predictions obtained from the application of a neural network model to the data.

Fig. 2 shows the Shapley-Lorenz measure of explainability ([5]), which is a normalised extension of the classic Shapley values, for all considered explanatory variables of the daily bitcoin price returns. Fig. 2 clearly highlights that the price returns of Gold is the most important variable that explains bitcoin price return variations, followed by the others.

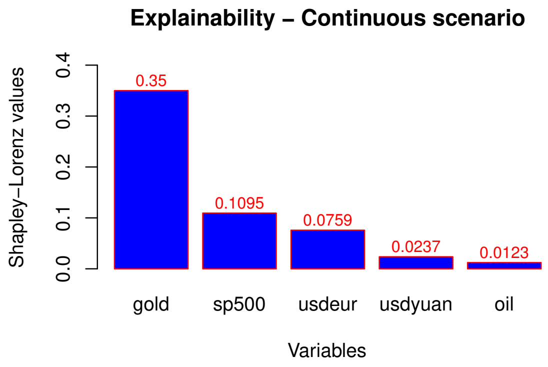


Figure 2: Explainability of the considered explanatory variables, in terms of the Shapley-Lorenz measure, for a continuous response.

Fig. 3 shows the Lorenz Zonoid of the machine learning model selected by our proposed feature selection procedure, based on the comparison between Lorenz Zonoids.

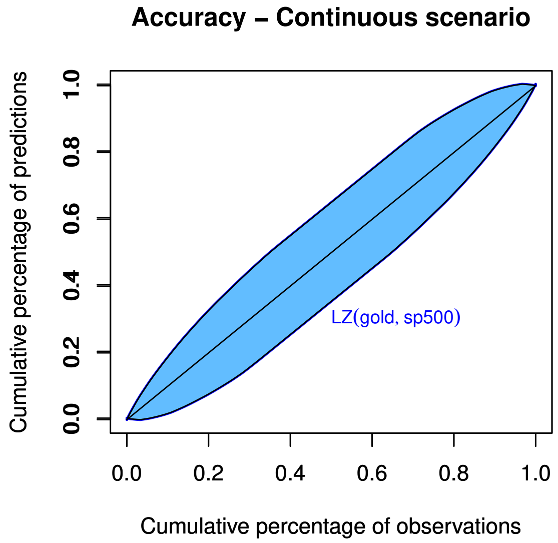


Figure 3: Accuracy of the selected model, in terms of its Lorenz Zonoid, for a continuous response.

Note that the model selected in Fig. 3 contains Gold and SP500 as the relevant predictors.

For robustness, we have repeated the analysis binarising the response variable around the zero value. The advantage of our proposed methodology is that no changes of metrics is requested to repeat the assessment of trustworthy AI, although the nature of the response variable has changed. For lack of space, we present only the results in terms of explainability, in Fig. 4.

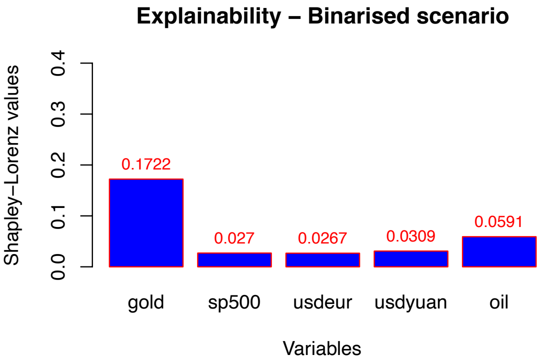


Figure 4: Explainability of the considered explanatory variables, in terms of the Shapley-Lorenz measure for a binary response.

From Fig. 4 note that Gold price returns is confirmed as the most important variable, but the second important variable is the Oil price, rather than SP500.

4. Conclusions

In the paper we propose a set of statistical measures that can ensure an overall assessment of the trustworthiness of AI applications. The application of the proposed scores to a neural network model, used to predict bitcoin price returns in terms of a set of classical financial variables, shows the practical utility of our approach.

References

[1] Bussmann, N., Giudici, P., Marinelli, D., Papenbrock, J.: Explainable Machine Learning in Credit Risk Management. Comput. Econ. **57**, 203–216 (2020) doi: 10.1007/s10614-020-10042-0

[2] Bracke, P., Datta, A., Jung, C., Shayak, S.: Machine learning explainability in finance: an application to default risk analysis (2019). https://www.bankofengland.co.uk/working-paper/2019/machine-learning-explainability-in-finance-an-application-to-default-risk-analysis

[3] Gini, C.: On the measure of concentration with special reference to income and statistics. General Series **208**, pp. 73-79. Colorado College Publication (1936)

[4] Giudici, P., Abu-Hashish, I: What determines bitcoin exchange prices? A network var approach. Financ. Res. Lett. **28**, 309–318 (2019). doi: 10.1016/j.frl.2018.05.013

[5] Giudici, P., Raffinetti, E.: Shapley-Lorenz eXplainable Artificial Intelligence. Expert Syst. Appl. **167**, 1–9 (2021) doi: 10.1016/j.eswa.2020.114104

[6] Koshevoy, G., Mosler, K.: The Lorenz Zonoid of a Multivariate Distribution. J. Am. Stat. Assoc. **91**, 873–882 (1996). doi: 10.2307/2291682

[7] Lorenz, M.O.: Methods of measuring the concentration of wealth. Publications of the American Statistical Association **70**, 209-219 (1905) doi:10.1080/15225437.1905.10503443