Journal of ISOSS
2023 Vol. 9(1), 233-250

STANDARDIZED METRICS FOR THE ASSESSMENT OF THE TRUSTWORTHINESS OF ARTIFICIAL INTELLIGENCE APPLICATIONS IN FINANCE

Maawish Bashir[1] and **Kehkashan Nizam**[2]
[1] Department of Mathematics, JS Public School and College, Rawalpindi, Pakistan.
Email: mahwish.jspublicschoolcollege@gmail.com
[2] Department of Business Administration, Iqra University Karachi, Pakistan. Email: kehkashan.60003@iqra.edu.pk

ABSTRACT

In all areas of finance, financial technologies are advancing thanks to machine learning models: from lending of peer-to-peer (payment) to management of asset (robot advisors) to block chain coins (payment). Models based on machine learning usually have high accuracy but a limited amount of explain ability. Further, high-risk AI applications using machine learning comply with a set of mandatory requirements and must be trustworthy, including Fairness and Sustainability, according to the proposed regulations. AI applications in finance cannot be evaluated for their trustworthiness based on standardized metrics. In order to fill this gap, the study propose integrated statistical methods that is based on the Lorenz Zonoid tool for assessing and monitoring the trustworthiness of AI applications over time. Sustainability was assessed by robustness against anomalous data, Accuracy was assessed through predictive accuracy, Fairness was assessed through prediction bias among different populations, and Explain ability was assessed through understanding of human and oversight. Using a dataset on financial prices that is easily downloadable, we tested our proposals

KEYWORDS

Macroeconomic Factors, PSX, Co-integration and VECM.

INTRODUCTION

Artificial Intelligence (AI) applications are booming in many fields, including health care and finance. In comparison with "classic" statistical models, they have an advantage in terms of predictive accuracy. The black-box nature of machines learning models, however, makes them capable of high predictive performance. In regulated businesses, this is a concern because authorities charged with monitoring the hazards associated with the use of Artificial Intelligence (AI) methods may not be able to validate them (see, for example, Joseph, 2019) and Bracke et al., 2019). For instance, the use of AI in credit lending may result in automated determinations that categorized a company as being at risk of default without providing the underlying reasoning, preventing corrective measures.

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

ORCID(s) using statistical methods based on the unifying notion of Lorenz Zonoid. Doing so, we will extend the recent work of (Giudici & Raffinetti, 2020), who has showed how to jointly measure Accuracy and Explainability. With the help of conventional statistical models like logistic and linear regression, the explainability condition is "by design" satisfied. In contrast to "black-box" ML models like neural networks and random forests, conventional statistical models may only have a modest predicted accuracy in complex data processing issues. This shows that post-modelling tools that can "explain" ML models might be beneficial.

Recent attempts in this direction, based on the cooperative game theory work of Shapley ((Shapley, 1953)), have led to promising applications of explainable AI methods in finance, among which (Bracke et al., 2019) and (Bussmann et al., 2020). Shapley values have the benefit of being agnostic—unaffected by the underlying model used to compute classifications and predictions—but they also have the drawback of not being normalized, making them challenging to understand and compare. To overcome this limitation, (Giudici & Raffinetti, 2020) proposed Shapley-Lorenz values, which combine Shapley values with Lorenz Zonoids, obtaining a measure of the contribution of each explanatory to the predictive accuracy of the response, rather than to the value of the predictions, as is the case for standard Shapley values.

In this paper we extend (Giudici & Raffinetti, 2020) and employ Lorenz Zonoids to build methods useful to measure not only Accuracy and Explianability, but also Sustainability and Fairness. The extension will allow developing an integrated measurement model for Sustainability, Accuracy, Fairness and Explainability, and a unified score of AI SAFEty.

The requirement of sustainability implies the model results are stable under variations in the data and, in particular, when extreme data, resulting from stressed scenarios and/or from cyber data manipulations, are inserted into the observed data. To measure the sustainability of AI applications we propose to extend variable selection methods, available for probabilistic models, to non-probabilistic models, such as random forests and neural network models, using statistical tests based on the comparison between the Lorenz Zonoids of the predictions. The extension provides a model selection criterion for (non-probabilistic) ML models, not available at the moment.

The criterion will lead to the choice of a parsimonious model, more sustainable than a complex one. The extension will also allow comparing 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. Specifically, the Lorenz Zonoid tool and the proposed Lorenz Zonoid comparison tests are used to illustrate the proposed methodology in the following section. Section 3 then discusses the empirical results we obtained by applying our proposal to the available data, and Section 4 concludes with some closing thoughts.

LITERATURE REVIEW

Stock Market Returns

According to Akhtar (2006), Pakistan is emerging as a global economy due to its steady political and macroeconomic stability, which boosts investor confidence and creates an attractive hub for capital, and the growth and revolution in its financial sector. Brealey et al. (2007) found that through historical analysis of stock market prices, investors can estimate their future returns and the risk of investment. Osamwonyi (2012) studied the impact of macroeconomic variables on stock market returns, highlighting the need for investors to understand this link in order to make informed decisions. The profitability of investing in emerging markets can be measured through aggregate returns on equity and dividend yield, with Pakistan's stock market resulting in the highest percent gain in the local index return in 2002 among global equity markets. The role of Pakistan's stock market in economic growth has been established over the years, maintaining its position as the appropriate trading market in South Asia in 2016 (Robert, 2016). Researchers such as Kheradyar, Ibrahim, and Nor (2011) have conducted studies to predict stock market returns in Pakistan based on financial ratios, which is a challenging issue in prominent markets (Fama and French, 2004). Empirical research shows that financial ratios can be used to predict fluctuations in stock market returns. Other studies have analyzed the link among macroeconomic variables and stock exchange returns in countries such as Kenya (Aroni, 2011), Ghana (Owusu Nantwi, 2011), Pakistan (Hussain and Sohail, 2011), China (Masood and Triki, 2012), India (Padhi and Naik, 2012), and Indonesia (Astuti, Nugroho, and Yogaswari, 2012).

Impact of Macroeconomic Variables on Stock Market Returns.

Interest Rate

The performance of stock markets and stock returns are affected by various macroeconomic variables, including interest rates. However, the impact of interest rates on stock returns is not always straightforward, and research findings are mixed. In less developed countries with no established stock markets, research on the link among macroeconomic variables and stock returns has yielded inconclusive results (Adjasi, 2009).

Interest rates, which represent the cost of capital or the price of borrowing money, are a key element in macroeconomic variables. When interest rates increase, the required rate of return on stocks rises, leading to a decrease in stock prices. This can deter investors from investing in the market, resulting in a decline in the stock market and economy. Conversely, a decrease in interest rates can be a positive sign for investors, indicating a growing market with lower borrowing costs (Blancher, 1981).

Research has shown that changes in interest rates can affect the profitability and returns of stocks. High stock prices and low interest rates lead to lower cost of capital, making the market more attractive and leading to growth (Blancher, 1981). However, interest rate increases can increase stock price volatility and impact financial stock returns (Adjasi, 2009).

Studies have found a negative correlation among interest rates and stock returns. This means that when interest rates increase, stock returns decrease, and vice versa. Researchers such as Aurangzed (2012) and Lobo (2000) have shown that interest rates have a negative impact on stock returns. Fama (1989) also found a negative impact of interest rates on stock prices.

Other studies have explored the impact of macroeconomic variables on stock returns in specific regions, such as South Asia (Aurangzed, 2012) and Pakistan (Ishan et al., 2007). These studies suggest that economic and financial variables are significant determinants of stock returns in these regions.

In summary, the link among macroeconomic variables and stock returns, including interest rates, is complex and can vary across regions. While some studies have found a negative impact of interest rates on stock returns, others have yielded inconclusive results. Understanding the impact of macroeconomic variables on stock returns can provide valuable insight for investors.

H1: Interest rates and PSX (PSX) Returns have Significant Relation.

Inflation Rate

Khans et al. (2012) examined the effects of exchange rate, inflation rate, and interest rate on stock returns of the KSE 100 index using data from 2001 to 2010. They did this by using multiple linear regressions. They discovered that the exchange rate has a large impact on stock returns on the PSX, but rates of interest and inflation rates had little effects.

According to Fama's (1981) proxy hypothesis, there is a negative connection among rate of inflation and prices of stock, which is attributed to the positive connection among stock returns and basic determinants of equity values. Inflation rate can be categorized into expected and unexpected inflation, with unexpected inflation having a greater impact on stock returns, particularly during economic contractions.

Ozlen & Ergun (2012) found exchange rate and interest rate to be the significant variables in stock price fluctuation using Autoregressive distributed lag technique with data from February 2005 to May 2012, whereas Sohail & Hussain (2009) found long and short link among economics variables and stock returns in Lahore stock market from Dec 2002 to June 2008.

The rising or falling of stock prices creates uncertainties for potential investors, which can affect the demand and supply forces at stocks. Price increases can also impact the investment decision of potential investors, which has a negative effect on the overall stock returns. Despite the Pakistani economy's poor performance, its capacity for development of major sectors cannot be doubted.

H2: Inflation and PSX (PSX) Returns have Significant Relation.

Exchange Rates

According to Wongbongo and Sharma (2020), the effect of macroeconomic variables on stock returns can be positive/negative depend on the foreign exchange rate, which is influenced by various internal and external factors. Sohail and Hussain found a positive and significant link among exchange rates and stock returns, while Robert Johnson emphasized the importance of exchange rates in stock market macroeconomics. Exchange rates not only affect policy makers and economists, but also investors who rely on them for returns on stocks. Currency depreciation can have both short-term and long-term negative effects on market returns, and abrupt changes in exchange rates can negatively impact a country's import and export. Fraser and Groenewold discovered a substantial influence of exchange rate variation on stock returns, while Aurangzeb found a positive link among foreign direct investment and exchange rates in South Asian countries.

Tsoukalas examined the strong link among macroeconomic variables and stock returns, with exchange rates influencing inflation and output. Beirne and Kumar found exchange rates to be a risk factor in financial stock returns, with Kumar also identifying a bidirectional linear and nonlinear causality among stock returns and exchange rates. Moysami applied the vectors error correction model to study the link among macroeconomic variables and stock returns, while Huge explored the sensitivity of stock returns to interest rates, exchange rates, and market risk. Adjasi analyzed the effect of exchange rate fluctuations on the stock exchange and recommended measures to ensure a stable macroeconomic environment for better investment decisions.

According to Adam et al., exchange rates have a positive impact on stock prices and are a main factor in macroeconomic variables, directly affecting money supply, interest rates, and inflation. Husung and Young noted that stock returns and currency devaluation can affect each other, with devaluation being a deliberate downward adjustment of a country's value against another currency. Johnson investigated the short-term and long-term link among exchange rates and stock returns, finding a positive and significant link that determines company performance and suggests higher rates of inflation in the future. Finally, Ibrahim found a bivariate link among exchange rates and stock returns, while Zahid studied the negative impact of macroeconomic variables on stock returns, but also found a positive link with stocks. Abrupt changes in exchange rates can adversely affect a country's exports and imports.

H3: Exchange Rates and PSX (PSX) Returns have Significant Relation.

Money Supply

Different industries may be impacted differently by macroeconomic variables, with some variables having a positive effect on one industry and a negative effect on another. Sohail and Hussain (2009) found that in the long-run, money supply had a significant positive effect on PSX returns. Conversely, the study by Humpe and Macmillan (2007) on the link among macroeconomic variables and PSX returns found that the effect of inflation rate and money supply on stock market prices was negative. Gunsel and Cukur (2007) concluded that various macroeconomic variables, including supply of money, rate of interest, rate of inflation, rate of exchange, and Pakistan stock market returns, had an important impact on the PSX market returns.

Similarly, the study by Gan et al. (2006) on the impact of macroeconomic variables on the stock prices of New Zealand Stock Exchange found that inflation rate and money supply had a negative link with stock market prices. They also noted that the percentage of capitalization in terms of GDP was low, resulting in a lower effect on the capital market.

Brahmasrene and Jiranyakul (2007) analyzed the link among stock market returns and macroeconomic variables in Thailand using cointegration, Granger causality, and unit root tests. They found a positive link among money supply and stock market returns, while the exchange rate, industrial production, and oil prices had a negative impact.

Tripathi and Seth (2014) noted that the equity market plays a significant role in establishing the speed of policy changes in a country and is highly sensitive to changes in monetary policy tools that control macroeconomic variables such as exchange rate, interest rate, and inflation rate. The study by Nizam, Liaqat, and Saghir (2022) emphasized the importance of observing macroeconomic variables, including inflation, interest rate, exchange rate, money supply, and industrial production, for a healthy and secure investment. Overall, these studies provide empirical evidence of the impact of macroeconomic variables on stock market returns and prices.

H1: Money Supply and PSX (PSX) Returns have Significant Relation.

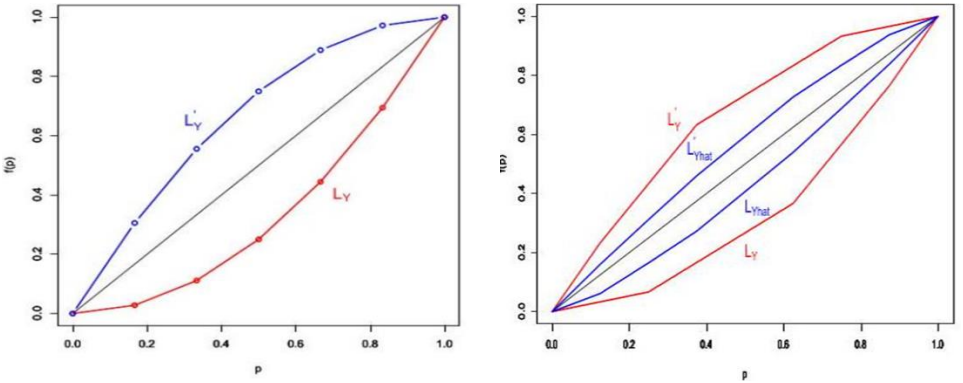
RESEARCH METHODOLOGY

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

Given a variable Y and n observations, the Lorenz Zonoid can be defined from the Lorenz and the dual Lorenz curves (see (Lorenz, 1905)). The Lorenz curve for a variable$Y$, denoted withL_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 $\left(i/n, \sum_{j=1}^{i} y_{r_j}/(n\bar{y})\right)$, 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 $\left(i/n, \sum_{j=1}^{i} y_{d_j}/(n\bar{y})\right)$, 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) (b)

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

A second important property concerns the practical implementation of the Lorenz Zonoid calculation. It can be shown that the Lorenz Zonoid-value of a generic variable - (such as the response variable, or the predicted response variable) is calculated as

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

where $r(\cdot)$ are the rank-scores associated with the $\cdot$ variable and $E(\cdot)$ is its 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 (Giudici & Raffinetti, 2020), 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:

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

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; $LZ\left(\hat{Y}_{X' \cup X_k}\right)$ and $LZ\left(\hat{Y}_{X'}\right)$ 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 ShapleyLorenz 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:

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

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. (Hand, Mannila & Smyth (2001)): the most commonly employed are the Root Mean Squared Error (when the response variable is on a continuous scale) and the Area Under the ROC curve (when the response variable is on a binary scale). Both are calculated on a test sample of the data, assuming the model being calculated on the remaining training sample. A more robust measure is the Lorenz Zonoid, which can be calculated on the test set in the same way for binary, ordered categorical and continuous responses. This generality is a clear further advantage of the Lorenz Zonoid.

Given a ML model with $k \leq K$ 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:

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

where $LZ\left(\hat{Y}_{X_1,\ldots,X_k}\right)$ is the Lorenz Zonoid of the predicted response variable, obtained using K predictors on the test set, and $LZ(Y_{\text{test}})$ 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 (Diebold and Mariano, 2002) for continuous responses and (DeLong et al., 1998) 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:

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

where $LZ\left(\hat{Y}_{X' \cup X_k}\right)$ and $LZ\left(\hat{Y}_{X'}\right)$ 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:

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

As $r(\cdot)/n$ is the empirical transformation of the cumulative distribution function $F(\cdot)$ (see, e.g. (Lerman & Yitzhaki, 1984)), the terms in equation (6) can be re-expressed as:

$$LZ\left(\hat{Y}_{X'\cup X_k}\right) - LZ\left(\hat{Y}_{X'}\right) = \frac{2\mathrm{Cov}\left(\hat{Y}_{X'\cup X_k}, F\left(\hat{Y}_{X'\cup X_k}\right)\right)}{E\left(\hat{Y}_{X'\cup X_k}\right)} - \frac{2\mathrm{Cov}\left(\hat{Y}_{X'}, F\left(\hat{Y}_{X'}\right)\right)}{E\left(\hat{Y}_{X'}\right)}, \quad (7)$$

where $F\left(\hat{Y}_{X'\cup X_k}\right)$ and $F\left(\hat{Y}_{X'}\right)$ 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 $E\left(\hat{Y}_{X'\cup X_k}\right) = E\left(\hat{Y}_{X'}\right) = \mu$ becomes a reasonable approximation. Assuming such approximation, equation (7), which describes the marginal contribution (MC) provided by X_k, can be simplified as follows:

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

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

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

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

$$\gamma = \xi\left(\hat{\pi}_{X'\cup X_k}\right) - \xi\left(\hat{\pi}_{X'}\right) \quad (10)$$

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

$$H_0: \xi\left(\hat{Y}_{X'\cup X_k}\right) = \xi\left(\hat{Y}_{X'}\right) \text{ vs } H_1: \xi\left(\hat{Y}_{X'\cup X_k}\right) \neq \xi\left(\hat{Y}_{X'}\right) \quad (11)$$

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

$$\hat{\xi}\left(\hat{Y}_{X'\cup X_k}\right) = U_1 = \frac{1}{4\binom{n}{2}} \sum_{i=1}^{n} (2i - 1 - n)\hat{Y}_{X'\cup X_{k(i)}}, \quad (12)$$

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

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

$$\hat{\xi}\left(\hat{Y}_{X'}\right) = U_2 = \frac{1}{4\binom{n}{2}} \sum_{i=1}^{n} (2i - 1 - n)\hat{Y}_{X'_{(i)}}, \quad (13)$$

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

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

$$\hat{\gamma} = \hat{\xi}\left(\hat{Y}_{X' \cup X_k}\right) - \hat{\xi}\left(\hat{Y}_{X'}\right) = U_1 - U_2 \tag{14}$$

Based on (Hand, Mannila & Smyth (2001), a function of several dependent U statistics has, after appropriate normalisation, an asymptotically normal distribution. As suggested by (Schechtman et al., 2008), a way to estimate the variance is to resort to the jackknife method. Specifically, the n values of $\hat{\gamma}$, pointed out with $\hat{\gamma}_{(-i)}$ (where $i = 1, \dots, n$), are calculated by omitting one pair $\left(\hat{Y}_{X' \cup X_k}, \hat{Y}_{X'}\right)$ at a time and the estimated variance is

$$\widehat{\mathrm{Var}(\hat{\gamma})} = \frac{n-1}{n} \sum_{i=1}^{n} \left(\hat{\gamma}_{(-i)} - \bar{\gamma}\right)^2 \tag{15}$$

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

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

$$Z = \frac{\hat{\gamma}}{\sqrt{\widehat{\mathrm{Var}(\hat{\gamma})}}} \to N(0,1) \tag{16}$$

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 $m = 1, \dots, M$ be the considered population groups and let K the number of the available predictors. We denote with $v_{mX_k}^{SL}$ 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 $k^* = 1, \dots, k$ 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_M^{SL*} defined as $V_M^{SL*} = \{v_1^{SL*}, \dots, v_m^{SL*}, \dots, v_M^{SL*}\}$, where $v_m^{SL*} = v_{mX_1}^{SL} + \cdots + v_{mX_{k^*}}^{SL}$ represents the sum of the Shapley-Lorenz values related to the predictors $X_1, \dots, X_{k^*}$.

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

$$\text{Fair-Score } = 1 - LZ(V_M^{SL*}) \tag{17}$$

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' ShapleyLorenz 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 $g = 1, \dots, G$ 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. $V_G^{SL*} = \{v_1^{SL*}, \dots, v_g^{SL*}, \dots, v_G^{SL*}\}$, where $v_g^{SL*} = v_{gX_1}^{SL} + \dots + v_{gX_{k^*}}^{SL}$ represents the sum of the Shapley-Lorenz values related to the predictors $X_1, \dots, X_{k^*}$.

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

$$\text{Sust-Score } = 1 - LZ(V_G^{SL*}) \tag{18}$$

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' ShapleyLorenz values in each group.

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

Empirical Analysis

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. The considered data are the same described in (Giudici & Abu-Hashish, 2019) and in (Giudici & Raffinetti, 2020) to explain bitcoin price variation as a function of the available financial explanatory variables. A further investigation of the data was provided

in a work by (Giudici & Raffinetti, 2020), who introduced a new AI approach resulting in the formalisation of a normalised measure for the assessment of the contribution of each additional predictor to the explanation of the bitcoin prices. For coherence with the previous cited works, here we choose the same time series observations, with the bitcoin prices from the Coinbase exchange as the target variable to be predicted. As suggested by (Giudici & Raffinetti, 2020) and (Giudici & Raffinetti, 2020), the time series for Oil, Gold and SP500 prices are taken into account as candidate financial explanatory variables. In line with (Giudici & Abu-Hashish, 2019), the exchange rates USD/Yuan and USD/Eur are also included as possible further explanatory variables. Our aim is to exploit the Lorenz Zonoid tool as a unified criterion for measuring the SAFEty of AI methodologies. 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.

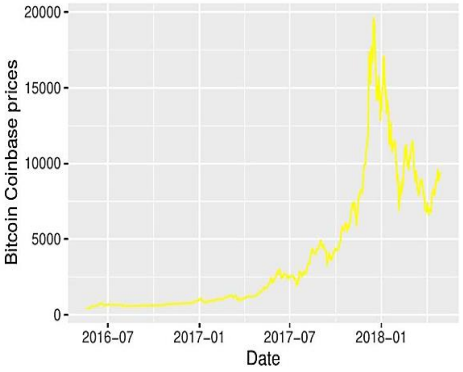


Figure 2: Bitcoin Prices

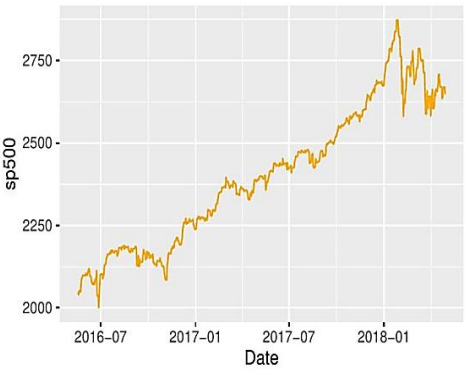


Figure 3: SP500 prices

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.

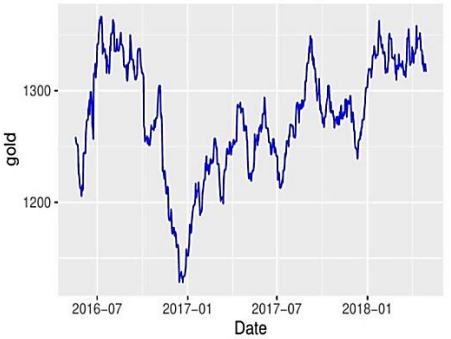


Figure 4: Gold Prices

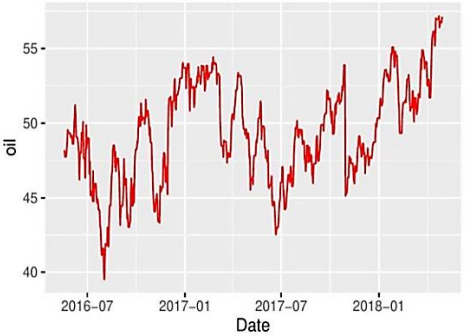


Figure 5: Oil Prices

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.

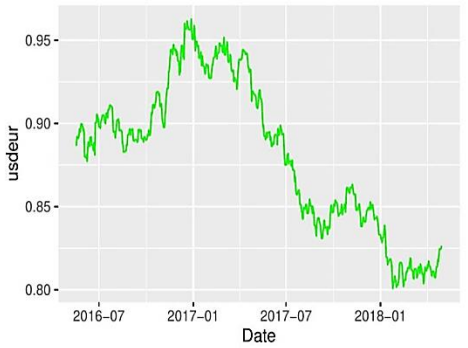


Figure 6: USD/EUR Exchange Rate

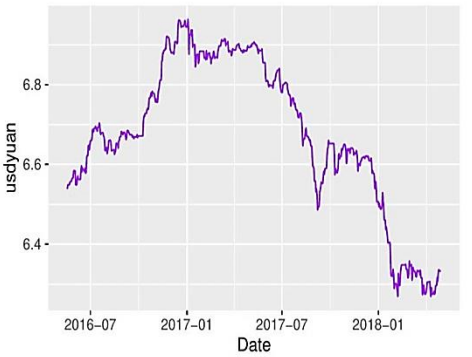


Figure 7: USD/YUAN Exchange Rate

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

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 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.

The aim of the data analysis is to build an explainable ML model that can predict bitcoin prices. Before proceeding, we transform all price series into their percentage returns. This because returns are scale free and the corresponding series are stationary (see, e.g. (Tsay, 2005)). As a ML model we apply, without loss of generality, a neural network with five hidden layers. We consider as training data the time series until December 31st, 2017; and as test data the 2018 time series. Figures 2-7 show that it will be difficult to obtain a high predictive accuracy, as the time series trends in 2018 change patterns with respect to the training data series.

In any case, the application of our proposed approach leads to a series of predictions for the 2018 return prices that can be compared with the actual returns, to obtain measures of trustworthiness (S.A.F.E.ty) of the neural network. 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 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.

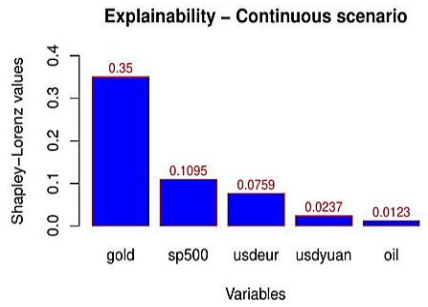
According to our proposal, we can calculate, as a sustainability score, the complement of the Gini coefficient of the Lorenz Zonoid. It results to be equal to 0.8314, indicating a high sustainability. With the aim of assessing fairness, we have considered, as a potential biasing variable, the amount 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.

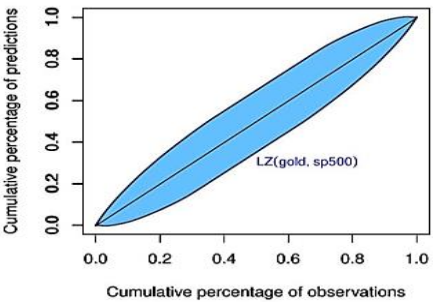
To show the universality of our proposal, we have binarised the response variable, with $Y = 1$ indicating positive returns and $Y = 0$ indicating negative returns, and applied the same neural network model as before, but to predict a binary, rather than a continuous response. Figure 9 shows the results 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.

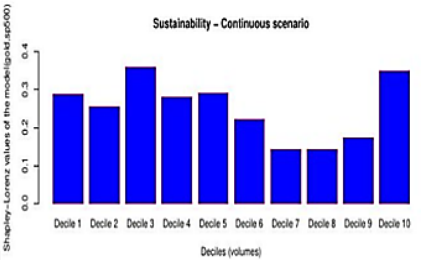
We have finally applied the sustainability and fairness assessments, in the same manner as for the continuous case. The results are in Figures 9(c) and 9(d), corresponding to scores of, respectively, 0.8184 and 0.716, While the sustainability of the model is similar to that corresponding to the continuous response case, fairness is lower, indicating that the sign of the returns depend on trading volumes more than the actual returns do 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.



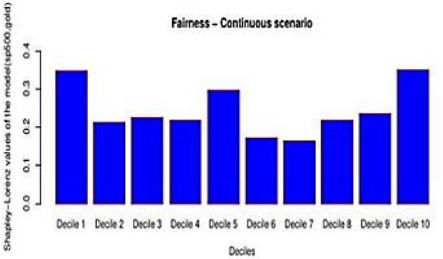
(a) **Explainability**



(b) **Accuracy**

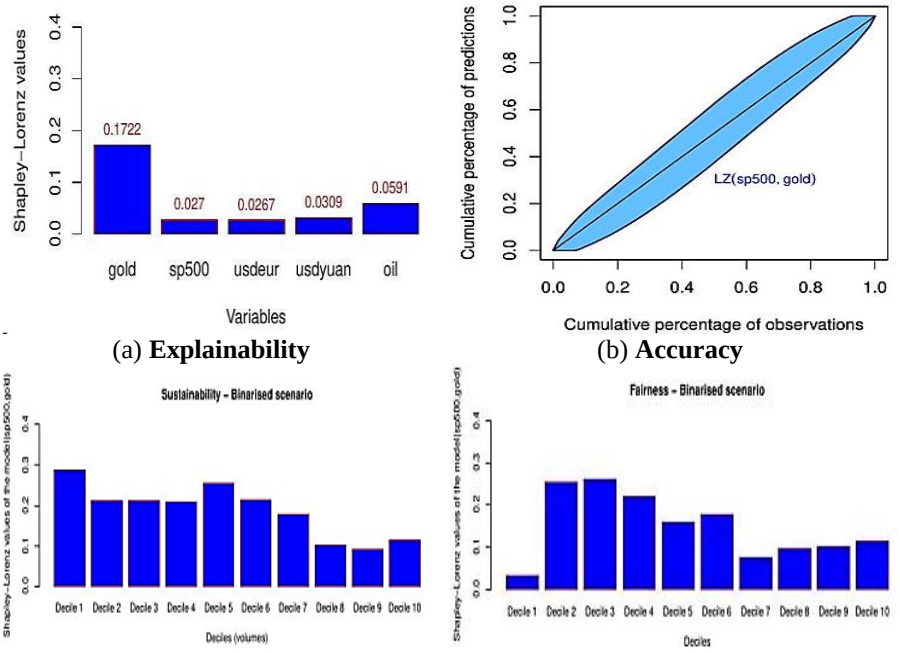


(c) **Sustainability**



(d) **Fairness**

Figure 8: Continuous Scenario



(a) **Explainability** (b) **Accuracy**

(c) **Sustainability** (d) **Fairness**

Figure 9: S.A.F.E.ty Assessment of the Neural Network Model for Bitcoin Price Returns

CONCLUSION

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 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 Explainability - Binarised scenario.

REFERENCES

1. Bracke, P., Datta, A., Jung, C. and 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. and 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. and Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3), 837845

4. Diebold, F.X. and Mariano, R.S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 20(1), 134-144.
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*, 208(1), 73-79.
7. Giudici, P. and Abu-Hashish, I. (2019). What determines bitcoin exchange prices? A network VAR approach. *Finance Research Letters*, 28, 309-318.
8. Giudici, P. and Raffinetti, E. (2020). Lorenz model selection. *Journal of Classification*, 37(3), 754-768.
9. Hand, D., Mannila, H. and Smyth, P. (2001). *Principles of Data Mining, Adaptive Computation and Machine Learning Series*. A Bradford Book.
10. Joseph, A. (2019). *Shapley regressions: A framework for statistical inference on machine learning models*. Working paper No. 2019/7
11. Rossini, A.J. and Tsiatis, A.A. (1996). A semiparametric proportional odds regression model for the analysis of current status data. *Journal of the American Statistical Association*, 91(434), 713-721.
12. Lerman, R.I. and Yitzhaki, S. (1984). A note on the calculation and interpretation of the Gini index. *Economics Letters*, 15(3-4), 363-368.
13. Liaqat, O., Nizam, K. and Haris, M. (2022). Link among Oil Price, Exchange Rate and Stock Market. *International Journal of Business Diplomacy and Economy*, 1(2), 6-17.
14. Lorenz, M.O. (1905). Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, 9(70), 209-219.
15. Nizam, K. (2022). Link Among Oil Price, Exchange Rate, And Stock Market. *International Journal of Economics Social and Technology*, 1(3), 85-96.
16. Nizam, K., Liaqat, O. and Saghir, W. (2022). The Impact of Commodities Price on Stock Market Price: Evidence from Pakistan. *Competitive Social Science Research Journal*, 3(2), 53-73.
17. Nizam, K. (2022). The Influence of Corruption, Political Stability, Foreign Direct Investment, and Domestic Investment on Economic Growth in BRICS Countries. *Front. Environ. Sci*, 10, 1036105.
18. Parkash, R., Ahmad, R., Qasim, S. and Nizam, K. (2022). Investor Sentiments and Stock Risk and Return: Evidence from Asian Stock Markets. *Competitive Social Science Research Journal*, 3(1), 341-371.
19. Schechtman, E., Yitzhaki, S. and Artsev, Y. (2008). The similarity between mean-variance and mean-Gini: Testing for equality of Gini correlations. *Advances in Investment Analysis and Portfolio Management*, 3, 97-122.
20. Shapley, L.S. (1953). A value for n-person games. 1-13. DOI: https://doi.org/10.7249/P0295
21. Tsay, R.S. (2005). *Analysis of financial time series*. John wiley & sons.