Socio-Economic Planning Sciences 87 (2023) 101560

Contents lists available at ScienceDirect

Socio-Economic Planning Sciences

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

Machine Learning Classification Model Comparison

Paolo Giudici, Alex Gramegna, Emanuela Raffinetti [*]

Department of Economics and Management, University of Pavia, Via San Felice al Monastero 5, Pavia, 27100, Italy

ARTICLE INFO

Keywords:
Lorenz Zonoids
Model selection
Predictive accuracy

ABSTRACT

Machine learning models are boosting Artificial Intelligence applications in many domains, such as automotive, finance and health care. This is mainly due to their advantage, in terms of predictive accuracy, with respect to classic statistical models. However, machine learning models are much less explainable: less transparent, less interpretable. This paper proposes to improve machine learning models, by proposing a model selection methodology, based on Lorenz Zonoids, which allows to compare them in terms of predictive accuracy significant gains, leading to a selected model which maintains accuracy while improving explainability. We illustrate our proposal by means of simulated datasets and of a real credit scoring problem. The analysis of the former shows that the proposal improves alternative methods, based on the AUROC. The analysis of the latter shows that the proposal leads to models made up of two/three relevant variables that measure the profitability and the financial leverage of the companies asking for credit.

1. Introduction

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

This is a problem in regulated industries, as authorities aimed at monitoring the risks arising from the application of Artificial Intelligence (AI) methods may not validate them (see, e.g. [1]). For example, the application of AI to finance may lead to automated decisions that can, for example, classify a company at risk of default, without explaining the underlying rationale and, therefore, impeding remedial actions.

The need to "explain" AI has become very important in recent years, following the increasing application of AI methods that impact the daily life of individuals and societies. At the institutional level, explanations can answer different kinds of questions about a model's operations, depending on the stakeholder they are addressed to (see, e.g. [2]): developers, managers, model checkers, regulators. In general, to be explainable, AI methods have to provide details or reasons clarifying their functioning.

The explainability requirement is fulfilled "by design" when classic statistical models, such as logistic and linear regression, are employed within AI applications. However, in complex data analysis problems, classical statistical models may be improved using "black-box" machine learning models, such as neural networks and random forests.

From the previous discussion, it emerges the need to empower highly predictive machine learning models with statistical tools that can "explain" them.

Recent attempts in this direction are based on the work of Shapley (see [3]) who proposed to assign a score to each candidate explanatory variable based on its additional contribution to each prediction. The application of Shapley's work has led to the development of a very promising research, especially in the field of computer science (see, e.g. [1,4]). One of the first applications of Shapley's work to finance is due by [5], who proposed to apply correlation networks (see, e.g. [6]) to the Shapley scores and, then, cluster them into rating classes.

Shapley values have the advantage of being agnostic: independent on the underlying model with which classifications and predictions are computed; but have the disadvantage of not being normalised and, therefore, difficult to be used in comparisons outside the specific application.

Interpretability and explainability appear more relevant in complex applications, where model comparison is necessary to select a model which, maintaining accuracy, becomes parsimonious and understandable. In the traditional paradigm, a statistical model is chosen through a sequence of pairwise comparisons, based on the ratio of the likelihoods (or of the posterior probabilities) of the models being compared. Unfortunately, these criteria are generally not applicable to machine learning models such as neural networks and random forests, which do not necessarily have an underlying probabilistic model.

[*] Corresponding author.
E-mail addresses: paolo.giudici@unipv.it (P. Giudici), alex.gramegna01@universitadipavia.it (A. Gramegna), emanuela.raffinetti@unipv.it (E. Raffinetti).

https://doi.org/10.1016/j.seps.2023.101560
Received 3 November 2022; Received in revised form 25 January 2023; Accepted 2 March 2023
Available online 8 March 2023

The previous consideration explains why the last few years have witnessed the growing importance of model selection methods based on the comparison between the predicted and the actually observed cases. In these methods, the data is split in two sets, with a "training" set used to fit a model and a "validation" set used to compare the predictions made by the trained model with the actual observed values.

In this paper, we contribute to the literature on model selection for machine learning models with a model comparison criterion based on the extension of Shapley values. Specifically, rather than evaluating the additional contribution of each variable to the point values of the predictions (as in the Shapley's approach), we propose to evaluate the additional contribution of each variable to the predictive accuracy of the predictions. To achieve this aim, we employ the decomposition property of the Lorenz Zonoid tool introduced by [7].

Doing so, we extend the available likelihood comparison procedures, applicable only to machine learning models that have a probabilistic background, to a predictive accuracy comparison framework, applicable to all models. To achieve this goal, we propose a statistical test to assess the significance of the additional contribution to predictive accuracy deriving from the inclusion of an extra explanatory variable in a sequence of models. This allows to overcome the main drawbacks of the BIC and the AIC, which require a probabilistic model specification to derive the likelihood of the data. When this is missing, as in complex machine learning models, model selection needs to be reformulated in terms of descriptive statistics of the distributions of the residuals (see, e.g. [8] for a discussion), for which statistical tests for variable importance can be derived only under specific conditions. This is the case for the Diebold–Mariano test, based on the Mean Squared Error of the residuals (see [9]).

To derive our proposed model comparison procedure, we will adapt to the binary response case the work of [10], who have shown the advantage of combining Lorenz Zonoids with Shapley values to select machine learning models. We will show how to build a model comparison methodology which can be used to order variables in terms of their contribution to predictive accuracy. Doing so, we provide a methodology that is able to simultaneously achieve the goals of predictive accuracy and explainability, rather than one after the other, as done in the explainable AI literature (see, e.g. [5]).

We will test our methodology in two different contexts: in simulated studies, aimed at assessing the comparative properties of our method; and in a real application, that concerns the prediction of financial default by means of a large set of highly correlated company performance variables, taken from balance sheets.

The paper is organised as follows. The next section illustrates the methodology: its background, the notion of Lorenz Zonoid predictive accuracy, and the proposed Lorenz Zonoid model comparison test; Section 3 introduces simulation studies to assess the performance of the methodology in the model selection context; Section 4 discusses the empirical findings obtained applying our proposal to the available financial data; finally, Section 5 contains some concluding remarks.

2. Methodology

In this section we adapt the Lorenz Zonoid decomposition approach illustrated by [7] to the binary classification context. Our proposal derives from the combination of two research streams. The first concerns the development of predictive machine learning methods for classification problems. The second concerns the development of explainable methods to understand the contribution of each explanatory variable to the predictive accuracy of machine learning models. The result is a methodology to select models that are, at the same time, predictively accurate and interpretable.

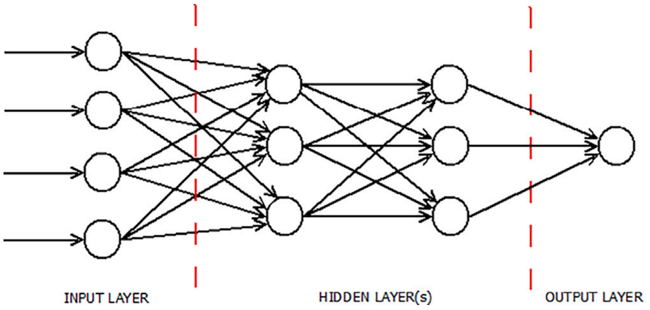


Fig. 1. Structure of a neural network model.

2.1. Background

Let Y be a binary response variable which can, for example and without loss of generality, express whether a company defaults ($Y = 1$) or not ($Y = 0$), as in a typical credit scoring problem. A popular model to predict Y is the logistic regression model (see, e.g. [5]).

Given K explanatory variables $X_1, \dots, X_K$, a logistic regression model for Y can be specified as follows:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{k=1}^{K} \beta_k x_{ki} = \eta_i,$$

where $i = 1, \dots, n$; $\eta_i = \beta_0 + \sum_{k=1}^{K} \beta_k x_{ki}$; π_i represents the probability of the event for the ith observation (company); $\mathbf{x}_i = (x_{1i}, \dots, x_{Ki})$ is the K-dimensional vector reporting the values taken by the K explanatory variables referred to the ith observation; β_0 and β_k $(k = 1, \dots, K)$ are the parameters representing the intercept and the kth regression coefficient, respectively.

By means of the maximum likelihood estimation method, the parameters β_0 and β_k can be estimated leading to derive the predicted probability of default as:

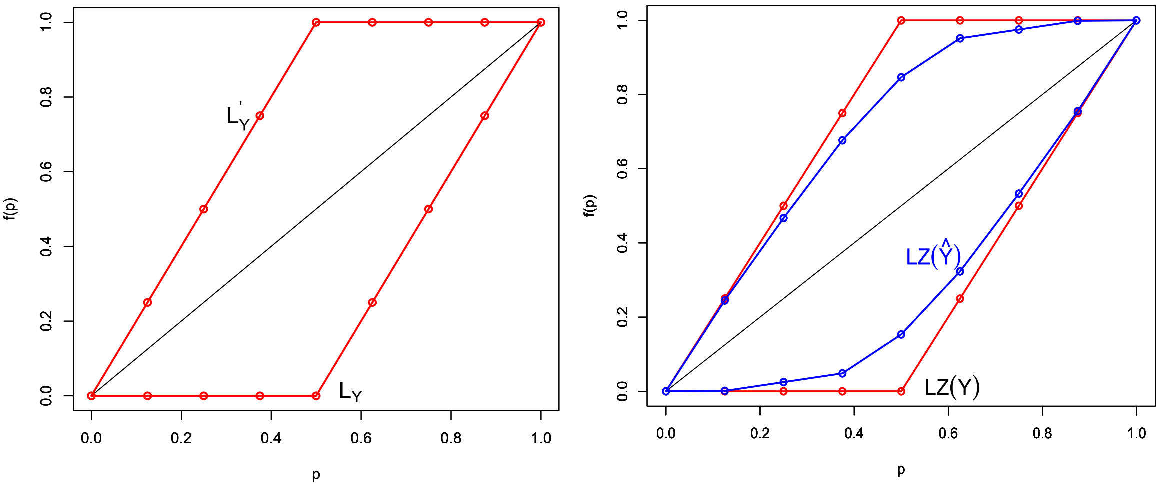
$$\hat{\pi}_i = \frac{e^{\hat{\eta}_i}}{1 + e^{\hat{\eta}_i}},$$

which can be employed to attach to the ith observation a "score": a number between zero and one which can be interpreted to signal, for example, the creditworthiness of a company: the higher the score the lower the trust. A classification of each company as $Y = 1$ or $Y = 0$ can then follow, comparing the score with an appropriate threshold, chosen on the basis of subject matter experience.

Employing logistic regression models for the analysis of credit scoring seems appropriate, as they are interpretable by default. However, they sometimes provide a limited predictive accuracy. To improve predictive accuracy, more complex machine learning models may be considered, such as neural network models and ensemble tree models. A high predictive accuracy is fundamental, particularly in the field of credit risk classification (see, e.g., [11–13], among others). A literature review on the use of machine learning methods in credit risk can be found in [14].

Neural network models were developed to mimic the structure of the human brain. The idea is to treat the brain as made up of highly interconnected elements (neurons) that work together to solve specific problems. Neural network models can be described by a graph organised according to different levels: the input, the hidden and the output layers, as displayed in Fig. 1.

While the input layer receives information from the external environment and each neuron in it usually corresponds to a predictor, the output layer provides the final result to be sent outside of the system. The hidden layers define the complexity of the neural network as they contain intermediate computational neurons, whose role is to increase the model fit. Data allow to learn the weights of the different connections between the neurons of the network.



[(a)] [(b)]

Fig. 2. [(a)] The Lorenz curve (L_Y) and the dual Lorenz curve (L'_Y) in the binary case; [(b)] The inclusion property $LZ(\hat{Y}) \subset LZ(Y)$ in the binary case.

More formally, a generic neuron j receives n input signals $x = [x_1, x_2, \ldots, x_n]$ from the neurons it is connected to in the previous layer. Each signal has an importance weight: $w_j = [w_{1j}, w_{2j}, \ldots, w_{nj}]$. The same neuron elaborates the input signals through a combination function which gives rise to a value, called "potential", computed as:

$$P_j = \sum_{j=1}^{n}(x_i w_{ij} - \theta_j),$$

where θ_j is a threshold which is activated only above a certain value: a cut-off point. The output of the jth neuron, denoted with y_j, derives from the application of a function, called activation function, to the potential P_j:

$$y_j = f(x, w_j) = f(P_j) = f\left(\sum_{j=1}^{n} x_i w_{ij} - \theta_j\right).$$

Ensemble tree models aim to combine the predictions derived from alternative tree models, thereby improving generalisation and robustness (see, e.g. [15] and the references therein).

The eXtreme Gradient Boosting (XGBoost) is one of the most popular ensemble tree models, particularly in the context of credit scoring, as discussed by [5]. The XGBoost is a supervised model based on a Gradient Boosting Machine (GBM), which combines distinct decision trees' predictions to obtain an "average" final prediction. In each decision tree, the nodes are built on a different subset of the features, implying that the trees are all different from each other and can catch distinct information from the data. At each step of the procedure, a new tree is built, learning from the errors generated by the previous trees. The XGBoost method shares the same functioning as the GBM, but it is faster and more advanced, as it provides specific regularisation techniques that reduce under-fitting and over-fitting of the model, increasing its performance. A mathematical formalisation of the XGBoost is illustrated in [16].

2.2. Lorenz Zonoid predictive accuracy

Lorenz Zonoids were introduced in [17] as a generalisation of the ROC curve in a multidimensional setting. They were further developed by [7] who proposed a Lorenz Zonoid decomposition approach that can be employed for model comparison purposes. The Lorenz Zonoid is based on the notion of mutual variability and can be exploited to develop a partial dependence measure that allows to detect the additional contribution of a new predictor to the predictive accuracy of an existing model.

Given a variable Y and n observations, the Lorenz Zonoid is the area between the Lorenz and the dual Lorenz curves (see, e.g. [18]).

The Lorenz curve for a variable Y, denoted with L_Y and obtained 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, indicated as L'_Y and obtained 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, whose graphical representation in the case of a binary response variable $Y = \{0, 1\}$ is displayed in Fig. 2(a).

It is worth mentioning that the Lorenz Zonoid fulfils some relevant properties. An important one is the "inclusion" of the Lorenz Zonoid built on the predicted values $\hat{Y}$ ($LZ(\hat{Y})$) into the Lorenz Zonoid of the response variable Y ($LZ(Y)$). This property is graphically depicted in Fig. 2(b).

As shown in [7], given a set of covariates X', denote with $\hat{Y}_{X' \cup X_k}$ and $\hat{Y}_{X'}$, respectively, the predicted values obtained from a model which includes X' and a further covariate X_k, and those obtained using only X'. The additional contribution of a covariate X_k can then be expressed in terms of a Partial Gini Contribution (PGC) measure as:

$$PGC_{Y,X_k|X'} = \frac{LZ(\hat{Y}_{X' \cup X_k}) - LZ(\hat{Y}_{X'})}{LZ(Y) - LZ(\hat{Y}_{X'})}, \tag{1}$$

where $LZ(\hat{Y}_{X' \cup X_k})$, $LZ(\hat{Y}_{X'})$ and $LZ(Y)$ define: the Lorenz Zonoids computed on the predicted values provided by the model that includes also X_k; the Lorenz Zonoids computed on the predicted values provided by the model that includes only X'; the Lorenz Zonoid computed on the Y target variable values.

The PGC measure can be interpreted in a game theoretical context, defining as pay-off for X_k a function of the numerator of the PGC measure in Eq. (1), as follows:

$$\textit{pay-off}\,(X_k) = LZ(\hat{Y}_{X' \cup X_k}) - LZ(\hat{Y}_{X'}), \tag{2}$$

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.

Remark 1. *When the response variable is binary, $Y = \{0, 1\}$, the terms $\hat{Y}_{X' \cup X_k}$ and $\hat{Y}_{X'}$ in Eqs.* (1) *and* (2) *can be re-written as the predicted probabilities of default $\hat{\pi}_{X' \cup X_k}$ and $\hat{\pi}_{X'}$, using a model that includes also the* *explanatory variable X_k, or a model that does not include the explanatory variable X_k. Thus, equations in* (1) and (2) *become*

$$PGC_{Y,X_k|X'} = \frac{LZ(\hat{\pi}_{X' \cup X_k}) - LZ(\hat{\pi}_{X'})}{LZ(Y) - LZ(\hat{\pi}_{X'})} \tag{3}$$

and

$$pay\text{-}off(X_k) = LZ(\hat{\pi}_{X' \cup X_k}) - LZ(\hat{\pi}_{X'}). \tag{4}$$

Remark 2. *The pay-off in Eqs.* (2) and (4) *measures a predictive gain, that is, the contribution to the explanation of the response variable due to each additional predictor included into the model. This result derives from the decomposition of the Lorenz Zonoid, which can be expressed as the sum of a component related to the explanatory variables X', and of a further component, function of the additional explanatory variable X_k.*

The previously mentioned decomposition specialises what proved by [7], in the case of a continuous response, to the binary case. More precisely, in [7] the Authors prove that the overall contribution provided by K covariates to the explanation of a continuous response variable depends on the single contributions according to the following formula:

$$MGC_{(Y|X_1,\ldots,X_K)} = \sum_{j=1}^{K} PGC_{Y,X_j|X_{i<j}}(1 - MGC_{Y|X_1,\ldots,X_{j-1}}), \tag{5}$$

where $MGC_{(Y|X_1,\ldots,X_K)}$ denotes the overall response variable variability explained by all the explanatory variables (i.e., $LZ(\hat{Y}_{X_1,\ldots,X_K})$); $PGC_{Y,X_j|X_{i<j}}$ is the contribution associated with the jth explanatory variable included into the model and $MGC_{Y|X_1,\ldots,X_{j-1}}$ is the overall contribution provided by the remaining $(j-1)$th explanatory variables (i.e., $LZ(\hat{Y}_{X_1,\ldots,X_{j-1}})$), with $j = 1,\ldots,K$.

Note that the previous decomposition parallels the well known decomposition of the goodness of fit coefficient R^2 for linear models:

$$R^2_{Y,X_1,\ldots,X_K} = \sum_{j=1}^{K} r^2_{Y,X_j|X_{i<j}}(1 - R^2_{Y,X_1,\ldots,X_{j-1}}), \tag{6}$$

where $R^2_{Y,X_1,\ldots,X_K}$ represents the determination coefficient of the linear model built on the K explanatory variables, $R^2_{Y,X_1,\ldots,X_{j-1}}$ denotes the coefficient of multiple correlation between Y and the fitted plane determined by the explanatory variables $X_1,\ldots,X_{j-1}$, and $r_{Y,X_j|X_{i<j}}$ denotes the coefficient of partial correlation between Y and X_j, conditional on the explanatory variables previously included into the model.

The analogy with the R^2 decomposition can be exploited to derive a decomposition of the Lorenz Zonoid for binary response variables. To achieve this goal, we need to define goodness of fit for a binary response variable. A contribution in this direction can be found in [19], which shows that, in the binary case

$$R^2 = \frac{Var(\hat{\pi})}{Var(\hat{\pi}) + \sum_{i=1}^{n} \hat{\pi}_i(1 - \hat{\pi}_i)/n}, \tag{7}$$

where $Var(\hat{\pi})$ is the sample variance (see, e.g. [20]).

Suppose to consider, for the sake of simplicity, only two explanatory variables X_1 and X_2 (i.e., $K = 2$). Eq. (6) can then be expressed as:

$$\begin{aligned} R^2_{X_1,X_2} =& \frac{Var(\hat{\pi}_{X_1})}{Var(\hat{\pi}_{X_1}) + \sum_{i=1}^{n} \hat{\pi}_{X_{1i}}(1 - \hat{\pi}_{X_{1i}})/n} + \frac{Var(\hat{\pi}_{X_1 \cup X_2}) - Var(\hat{\pi}_{X_1})}{\sum_{i=1}^{n} \hat{\pi}_{X_{1i}}(1 - \hat{\pi}_{X_{1i}})/n} \\ &+ \left[1 - \frac{Var(\hat{\pi}_{X_1})}{Var(\hat{\pi}_{X_1}) + \sum_{i=1}^{n} \hat{\pi}_{X_{1i}}(1 - \hat{\pi}_{X_{1i}})/n}\right]. \end{aligned} \tag{8}$$

And the decomposition in Eq. (5) can be expressed as:

$$\begin{aligned} MGC_{(Y|X_1,X_2)} &= MGC_{Y|X_1} + PGC_{Y,X_2|X_1}(1 - MGC_{Y|X_1}) \\ &= \frac{LZ(\hat{\pi}_{X_1})}{LZ(Y)} + \frac{LZ(\hat{\pi}_{X_1 \cup X_2}) - LZ(\hat{\pi}_{X_1})}{LZ(Y) - LZ(\hat{\pi}_{X_1})}\left[1 - \frac{LZ(\hat{\pi}_{X_1})}{LZ(Y)}\right], \end{aligned} \tag{9}$$

where $\frac{LZ(\hat{\pi}_{X_1 \cup X_2})}{LZ(Y)} = MGC_{(Y|X_1,X_2)}$ represents the response variability share explained by the two jointly considered explanatory variables X_1 and X_2; $\frac{LZ(\hat{\pi}_{X_1 \cup X_2}) - LZ(\hat{\pi}_{X_1})}{LZ(Y) - LZ(\hat{\pi}_{X_1})} = PGC_{Y,X_2|X_1}$ measures the partial contribution provided by the inclusion of the explanatory variable X_2 in the model; $\left[1 - \frac{LZ(\hat{\pi}_{X_1})}{LZ(Y)}\right]$ denotes the variability not explained by X_1.

We remark that the relation in Eq. (9) can be derived using the proof of Result 5 in [7]. It can also be shown that, when used in a stepwise model selection procedure, the path selected by the Lorenz Zonoid has a monotonicity property. More precisely, following the inclusion property, The Lorenz Zonoids of the predictions generated by a more complex model is an area which is greater than that associated with a simpler model, implying that the explained variation of Y monotonically increases with the number of predictors included into the model.

Remark 3. *The Lorenz Zonoids $LZ(\hat{\pi}_{X' \cup X_k})$ and $LZ(\hat{\pi}_{X'})$ can also be expressed using ordinary covariance operators (see, e.g. [21]):*

$$LZ(\hat{\pi}_{X' \cup X_k}) = \frac{2Cov(\hat{\pi}_{X' \cup X_k}, r(\hat{\pi}_{X' \cup X_k}))}{nE(\hat{\pi}_{X' \cup X_k})} \quad \textit{and}$$

$$LZ(\hat{\pi}_{X'}) = \frac{2Cov(\hat{\pi}_{X'}, r(\hat{\pi}_{X'}))}{nE(\hat{\pi}_{X'})}, \tag{10}$$

where $r(\hat{\pi}_{X' \cup X_k})$ and $r(\hat{\pi}_{X'})$ are the rank scores of $\hat{\pi}_{X' \cup X_k}$ and $\hat{\pi}_{X'}$; n is the sample size; $E(\hat{\pi}_{X' \cup X_k})$ and $E(\hat{\pi}_{X'})$ are the expected values of $\hat{\pi}_{X' \cup X_k}$ and $\hat{\pi}_{X'}$.

2.3. Lorenz Zonoid model comparison

We now move to model comparison.

A stepwise model comparison procedure can be implemented considering the term $LZ(\hat{\pi}_{X' \cup X_k}) - LZ(\hat{\pi}_{X'})$ in Eq. (4). The procedure starts building K models, each depending on one of the K predictors, and then computing the Lorenz Zonoids of the predicted values derived from any single model.

In a forward stepwise algorithm the predictor providing the highest Lorenz Zonoid value can be chosen as the first variable to be included into the model. Otherwise, in a backward stepwise algorithm, the predictor with the lowest Lorenz Zonoid value can be chosen as the first variable to be removed from the full model.

In the former case, the procedure continues by fitting, at each step, a more complex model that includes the predictor which provides the highest contribution in terms of the difference in Eq. (4). In the latter case, the procedure continues by fitting, at each step, a simpler model obtained deleting the predictor with the lowest contribution in terms of the same difference in Eq. (4).

To evaluate the statistical contribution of a single variable, we need to derive the distribution of the difference $LZ(\hat{\pi}_{X' \cup X_k}) - LZ(\hat{\pi}_{X'})$, where $\hat{\pi}_{X' \cup X_k}$ are the predicted values generated by the most complex model (involving the additional X_k variable) and $\hat{\pi}_{X'}$ are the predicted values generated by the simplest model (without the X_k variable).

To this aim, based on Eq. (10), the difference in Eq. (4) can be expressed as:

$$LZ(\hat{\pi}_{X' \cup X_k}) - LZ(\hat{\pi}_{X'}) = \frac{2Cov(\hat{\pi}_{X' \cup X_k}, r(\hat{\pi}_{X' \cup X_k}))}{nE(\hat{\pi}_{X' \cup X_k})} - \frac{2Cov(\hat{\pi}_{X'}, r(\hat{\pi}_{X'}))}{nE(\hat{\pi}_{X'})}. \tag{11}$$

As $r(\cdot)/n$ is the empirical transformation of the cumulative distribution function $F(\cdot)$, the terms in Eq. (11) can be re-written as:

$$LZ(\hat{\pi}_{X'\cup X_k}) - LZ(\hat{\pi}_{X'}) = \frac{2Cov(\hat{\pi}_{X'\cup X_k}, F(\hat{\pi}_{X'\cup X_k}))}{E(\hat{\pi}_{X'\cup X_k})} - \frac{2Cov(\hat{\pi}_{X'}, F(\hat{\pi}_{X'}))}{E(\hat{\pi}_{X'})}, \tag{12}$$

where $F(\hat{\pi}_{X'\cup X_k})$ and $F(\hat{\pi}_{X'})$ are the cumulative distribution functions of $\hat{\pi}_{X'\cup X_k}$ and $\hat{\pi}_{X'}$, respectively.

In the case of linear regression, the mean of the predicted response values is always equal to the mean of the original target values, implying that $E(Y) = E(\hat{Y})$. For more general models, the aforementioned condition does not fully hold, implying that $E(\hat{\pi}_{X'\cup X_k}) = E(\hat{\pi}_{X'}) = \mu$ becomes a reasonable approximation. Assuming such approximation, Eq. (12), which describes the marginal contribution (MC) provided by X_k, can be simplified as follows:

$$MC = \frac{2Cov(\hat{\pi}_{X'\cup X_k}, F(\hat{\pi}_{X'\cup X_k}))}{\mu} - \frac{2Cov(\hat{\pi}_{X'}, F(\hat{\pi}_{X'}))}{\mu}. \tag{13}$$

In line with the previous mathematical derivations, we propose γ as an adjusted version of Eq. (13), i.e.

$$\gamma = \frac{\mu}{2} \cdot MC = Cov(\hat{\pi}_{X'\cup X_k}, F(\hat{\pi}_{X'\cup X_k})) - Cov(\hat{\pi}_{X'}, F(\hat{\pi}_{X'})). \tag{14}$$

By indicating the covariances $Cov(\hat{\pi}_{X'\cup X_k}, F(\hat{\pi}_{X'\cup X_k})) = \xi(\hat{\pi}_{X'\cup X_k})$ and $Cov(\hat{\pi}_{X'}, F(\hat{\pi}_{X'})) = \xi(\hat{\pi}_{X'})$, γ in (14) can be re-expressed as:

$$\gamma = \xi(\hat{\pi}_{X'\cup X_k}) - \xi(\hat{\pi}_{X'}). \tag{15}$$

A test for the equality of the two Lorenz Zonoids, assuming the continuity of the $\hat{\pi}$ distribution, can thus be developed by setting the following hypotheses:

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

To proceed with the test, $\xi(\hat{\pi}_{X'\cup X_k})$ and $\xi(\hat{\pi}_{X'})$ can be derived in terms of U-statistics, which estimate $Cov(\hat{\pi}_{X'\cup X_k}, F(\hat{\pi}_{X'\cup X_k}))$ and $Cov(\hat{\pi}_{X'}, F(\hat{\pi}_{X'}))$, respectively.

To better clarify this issue, let us first provide the general definition of a U-statistic.

Definition 1. A U-statistic of order p with kernel h is defined as

$$U = \frac{1}{\binom{n}{p}} \sum_{i \subseteq [n]} h(X_{i_1}, \dots, X_{i_p}),$$

where h is a real-valued measurable function symmetric in its arguments (see, e.g. [22]).

For a generic variable X, [23] has shown that $Cov(X, F(X))$ can be expressed in terms of a U-statistic as follows:

$$\begin{aligned} U &= \frac{1}{\binom{n}{2}} \sum_{i<j}\sum h(X_i, X_j) \\ &= \frac{1}{\binom{n}{2}} \sum_{i<j}\sum \frac{1}{4}\left[(X_i - X_j)I_{X_i > X_j} + (X_j - X_i)I_{X_j > X_i}\right] \\ &= \frac{1}{\binom{n}{2}} \sum_{i<j}\sum \frac{1}{4}|X_i - X_j|, \end{aligned} \tag{16}$$

where $I_{X_i > X_j}$ and $I_{X_j > X_i}$ are the indicator functions taking values equal to 1 if $X_i > X_j$ and $X_j > X_i$, respectively, and values equal to 0 otherwise.

In Eq. (16), the U-statistic's kernel $h(X_i, X_j)$ is symmetric of degree 2 for the parameter $Cov(X, F(X))$ and it thus results a consistent estimate of $Cov(X, F(X))$ (see, e.g. [23]). Note that $4U$ corresponds to the Gini's mean difference.

It is also worth remarking that a direct computation of a U-statistic is computationally intensive, with a complexity $O(n^2)$ (see, e.g. [24]). To overcome this drawback, U can be written as a linear combination of order statistics, reducing the computation to $O(n \log n)$. For this reason, as shown by [25] and subsequently by [26] (page 199), U can be re-formulated as:

$$U = \frac{1}{4\binom{n}{2}} \sum_{i=1}^{n} (2i - 1 - n) X_{(i)}. \tag{17}$$

Following the previous arguments, the estimators of $\hat{\xi}(\hat{\pi}_{X'\cup X_k})$ and $\hat{\xi}(\hat{\pi}_{X'})$ in Eq. (15) can be defined as:

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

and

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

where $\hat{\pi}_{X'\cup X_{k(i)}}$ and $\hat{\pi}_{X'_{(i)}}$ are the ith order statistics of $\hat{\pi}_{X'\cup X_{k1}}, \dots, \hat{\pi}_{X'\cup X_{kn}}$ and $\hat{\pi}_{X'_1}, \dots, \hat{\pi}_{X'_n}$, respectively.

Inference on the single parameters $\xi(\hat{\pi}_{X'\cup X_k})$ and $\xi(\hat{\pi}_{X'})$ can then be formalised based on Theorem 10.1, reported in [26], as follows:

Theorem 1. *Let $X_1, \dots, X_n$ be independent random variables having a distribution function F with finite second moment. The U-statistic for the parameter θ, with a symmetric kernel $h(X_1, \dots, X_p)$ of degree p, is an unbiased estimator for θ and the distribution of $\sqrt{n}(U - \theta)$ converges to a normal distribution as $n \to \infty$ under the condition that $E(h^2(X_1, \dots, X_p))$ exists.*

From Theorem 1, it follows that $\sqrt{n}(U_1 - \xi(\hat{\pi}_{X'\cup X_k}))$ and $\sqrt{n}(U_2 - \xi(\hat{\pi}_{X'}))$ converge to a normal distribution as $n \to \infty$.

An estimator of γ in Eq. (15) can then be provided as a function of two dependent U-statistics:

$$\hat{\gamma} = \hat{\xi}(\hat{\pi}_{X'\cup X_k}) - \hat{\xi}(\hat{\pi}_{X'}) = U_1 - U_2. \tag{18}$$

According to Theorem 3.5 proposed by [27] and further reported as Theorem 10.4 in [26], it also results that the asymptotic distribution of a function of several (dependent) U-statistics converges to the normal distribution as $n \to \infty$. Thus, it follows that $\sqrt{n}(\hat{\gamma} - \gamma)$ is gaussian.

To be generalisable, the proposed test should be applicable in a large class of distributions, including the gaussian, and should be essentially "distribution-free" or at least asymptotically distribution free (see, e.g. [26]). In order to draw inference on Lorenz Zonoids, we can exploit the fact that their estimators are U-statistics of degree 2. Then, we can resort to the U-statistic theory (Theorem 1 above) to develop hypotheses tests, based on the asymptotic normality of U_1 and U_2, so to find the asymptotic variance of $\hat{\gamma}$. Following this line of reasoning, the test can be applied to any kind of data distributions (binary, ordinal and continuous).

As proved by [28], consistent estimators of the standard error of U-statistics can be obtained by the jackknife resampling method. Specifically, the n values of $\hat{\gamma}$, $\hat{\gamma}_{(-i)}$ (where $i = 1, \dots, n$), can be calculated by omitting one pair $(\hat{\pi}_{X'\cup X_k}, \hat{\pi}_{X'})$ at a time. The estimated variance equals to:

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

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

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

$$Z = \frac{\hat{\gamma}}{\sqrt{\widehat{Var(\hat{\gamma})}}} \to N(0,1) \tag{19}$$

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

Table 1
Correlation matrix.

	Y	X_1	X_2	X_3	X_4	X_5	X_6
Y	1	0.8	0.5	0.3	0.1	0	0
X_1		1	0.2	0.7	0.3	0	0
X_2			1	0.05	0.1	0	0
X_3				1	0.5	0	0
X_4					1	0	0
X_5						1	0
X_6							1

3. Simulation studies

In this section we present simulation studies aimed at examining the performance of the proposed model comparison procedure based on Lorenz Zonoids.

The simulation designs are illustrated in Section 3.1, whereas the corresponding results are reported and commented in Section 3.2.

3.1. Simulation designs

To validate our proposal, we focus on three different simulation designs:

1. we first consider a vector of seven random variables, including a response variable Y and six explanatory variables $X_1, \dots, X_6$, and we assume it to be distributed as a seven-dimensional normal distribution, with a correlation matrix specified as in Table 1. Table 1 assumes that:
   - Y is highly correlated with X_1: $\rho = 0.8$;
   - Y is correlated with X_2: $\rho = 0.5$;
   - Y has a low correlation with X_3: $\rho = 0.3$;
   - Y has a very low correlation with X_4: $\rho = 0.1$;
   - Y is not correlated with X_5 and X_6: $\rho = 0$.
   - Variables X_5 and X_6 are not correlated with the other four explanatory variables X_1, X_2, X_3 and X_4.

   From the above distribution, we generate two samples with, respectively, 1,000 and 10,000 observations.
   In line with the paper, we then binarise the response variable Y around its average, with a resulting proportion of 1s approximately equal to 50%: a balanced dataset. We also binarise Y around the first quartile, leading to an unbalanced dataset with a proportion of 1s approximately equal to 75%.
   We apply our procedure to the predictions obtained from a logistic regression models, explainable by design and, consequently, simpler to be understood.
   The data is split into a training set, composed of 80% of the observations, and a test set, composed of the remaining 20%. A forward stepwise procedure is implemented by first fitting a logistic regression model on the training set and, then, including the explanatory variables which progressively provide the highest marginal contribution to predictive accuracy on the test set, as measured by the pay-off based on the Lorenz Zonoids. For comparison purposes, we consider a forward stepwise model selection based on the AUROC, as well as a goodness of fit selection procedure based on the AIC. The procedure stops when the additional contribution provided by a new included predictor is not significant, using the proposed test to compare Lorenz Zonoids and the DeLong test to compare ROC curves (see, e.g. [29]);
2. we replicate the previous simulation 1,000 times. To do so, we generate 1,000 samples of size 10,000 from the same seven-dimensional normal distribution, whose correlation matrix is displayed in Table 1 and we binarise the response variable as before, considering, without loss of generality, the case of balanced data. Because of the sample replications, we will get, for any pairwise model comparison, 1,000 pay-offs rather than only one. To carry out model selection, we will derive the sample cumulative distribution function of the considered metrics (Lorenz Zonoid and AUROC measures) for both models under testing and apply the Page's test [30] to verify whether the difference between two cumulative distribution functions is significant;
3. we extend the simulation study to a high-dimensional setting by increasing the number of predictors from six to nineteen, to further investigate the performance of our proposal in comparison with that of the AUROC. We then generate 10,000 observations from a twenty-dimensional normal distribution, with varying levels of correlation between the response variable Y and the 19 predictors, specified as follows:
   - medium correlation (ρ taking values in the range $[0.55; 0.44]$);
   - slightly low correlation: (ρ taking values in the range $[0.35; 0.23]$);
   - low correlation: (ρ taking values in the range $[0.19; 0.14]$);
   - almost no correlation: ($\rho = 0.07$).

   We then binarise the response variable as before (considering the case of balanced data) and follow a forward stepwise procedure similar to what described in the simulation design 1.

3.2. Simulation results

Results from simulation design 1.
The results for model comparison are displayed in Figs. 3 and 4, for the case of balanced data, and Figs. 5 and 6, for the case of unbalanced data. More precisely, Figs. 3 and 5 refer to the generating data process with 1,000 observations, while Figs. 4 and 6 refer to the case of 10,000 observations.

At each step of the stepwise procedure, the significance of the contribution given by an additional explanatory variable is assessed through the Lorenz Zonoid and DeLong tests, whose results are reported in terms of the corresponding p-values in Figs. 3, 4, 5 and 6 (a)–(b).

Figs. 3 and 5 order the six considered explanatory variables in terms of their marginal Lorenz Zonoids, AUROC and AIC. When the marginal Lorenz Zonoid are used (Figs. 3 and 5(a)), the ordering is consistent with the assumed correlations between the X variables and the response variable. When the AUROC is applied, in Figs. 3 and 5(b) the ordering changes with X_6 (not correlated with Y) replacing X_4 (correlated with Y). Moreover, in the case of unbalanced data, variable X_4 is the last to be included, providing the lowest contribution. Finally, the application of the AIC measure (Figs. 3 and 5(c)) reveals a behaviour similar to that of the Lorenz Zonoids.

Figs. 3 and 5(a)–(b) report the p-values that correspond to the progressive tests of variable inclusion. Fig. 3(a), for balanced data, indicates that a stepwise selection based on the Lorenz Zonoid tests stops with a model that contains (X_1, X_2, X_3), the most correlated variables. Fig. 5 (a), for unbalanced data, highlights that the model should include variables X_1 and X_2 and, possibly, variable X_3, if a significance level of $\alpha = 0.10$ is considered. Figs. 3 and Fig. 5(b) indicate similar results when the stepwise selection is based on the DeLong test for the AUROC.

Figs. 4 and 6 replicate the previous analysis using a larger sample of 10,000 observations.

The p-values in Fig. 4(a), for the balanced data, indicate that, with the Lorenz Zonoid procedure, variable X_4 becomes significant. On the other hand, the procedure based on the AUROC (Fig. 4(b)) fails to recognise the correct model, as it selects, besides X_1, X_2, X_3, X_4 also variable X_6. Looking at Fig. 6, in the case of unbalanced data, the Lorenz Zonoid procedure leads to the same model selected in the case

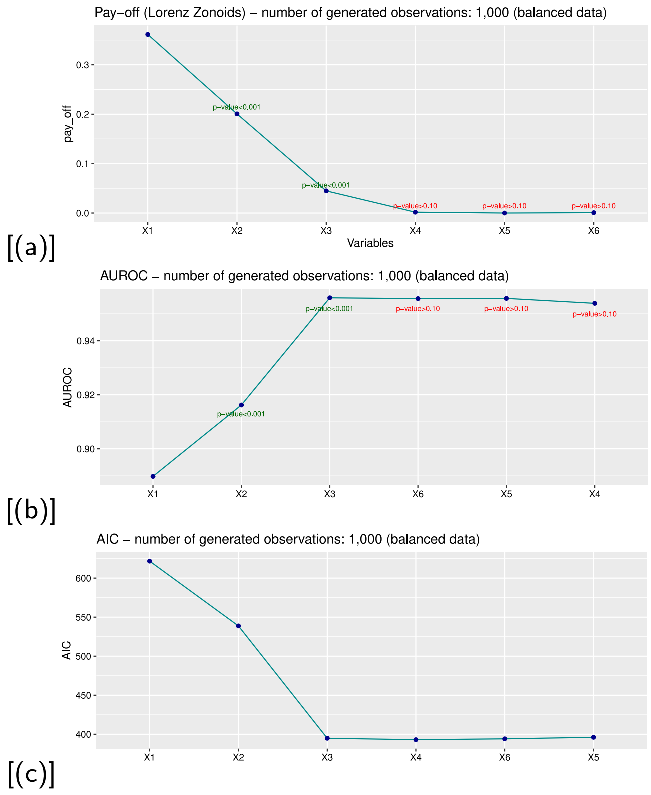


Fig. 3. [(a)] Pay-off (Lorenz Zonoids): 1,000 observations (balanced data) [(b)] AUROC: 1,000 observations (balanced data) [(c)] AIC: 1,000 observations (balanced data).

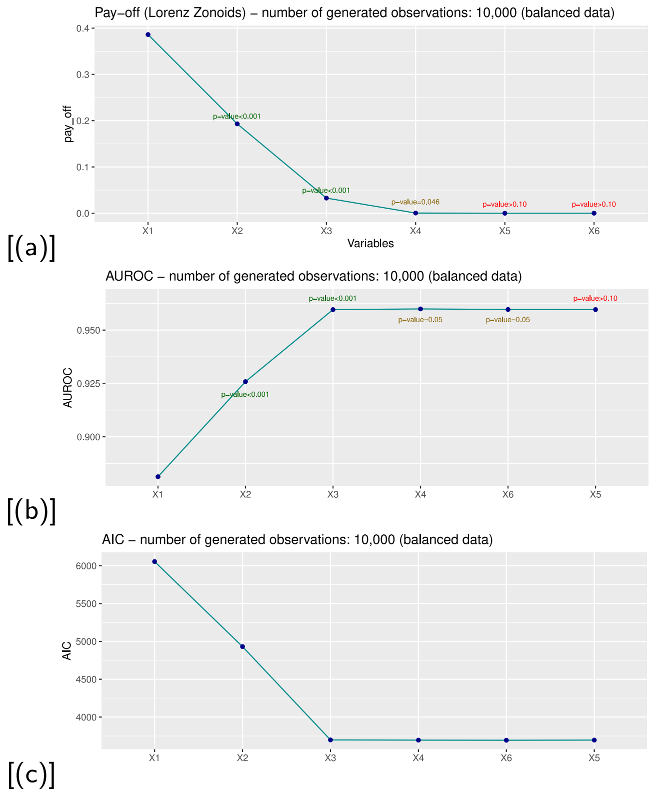


Fig. 4. [(a)] Pay-off (Lorenz Zonoids): 10,000 observations (balanced data) [(b)] AUROC: 10,000 observations (balanced data) [(c)] AIC: 10,000 observations (balanced data).

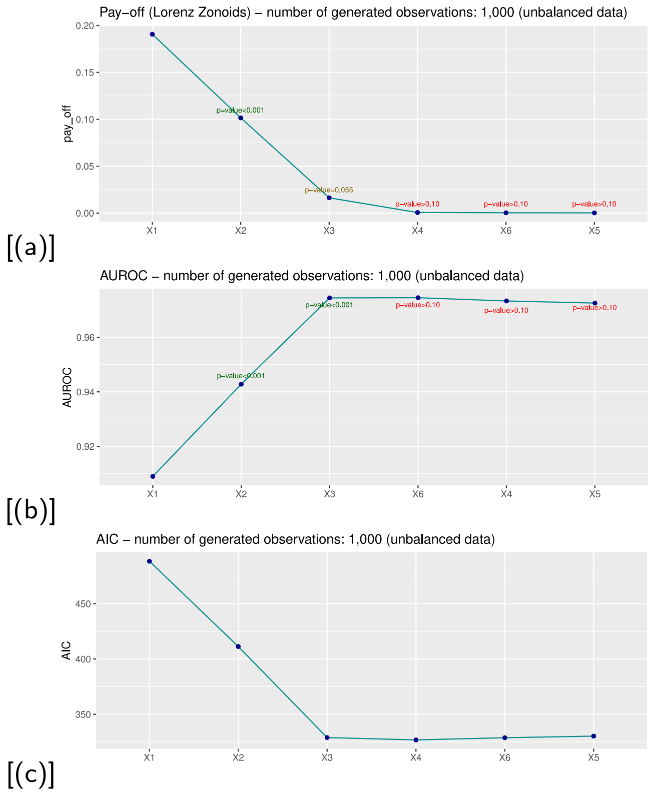


Fig. 5. [(a)] Pay-off (Lorenz Zonoids): 1,000 observations (unbalanced data) [(b)] AUROC: 1,000 observations (unbalanced data) [(c)] AIC: 1,000 observations (unbalanced data).

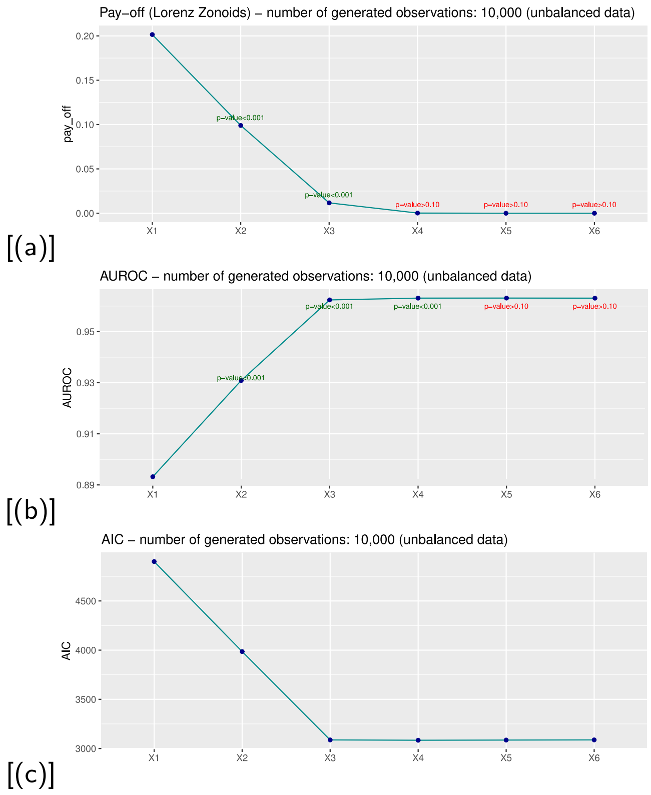


Fig. 6. [(a)] Pay-off (Lorenz Zonoids): 10,000 observations (unbalanced data) [(b)] AUROC: 10,000 observations (unbalanced data) [(c)] AIC: 10,000 observations (unbalanced data).

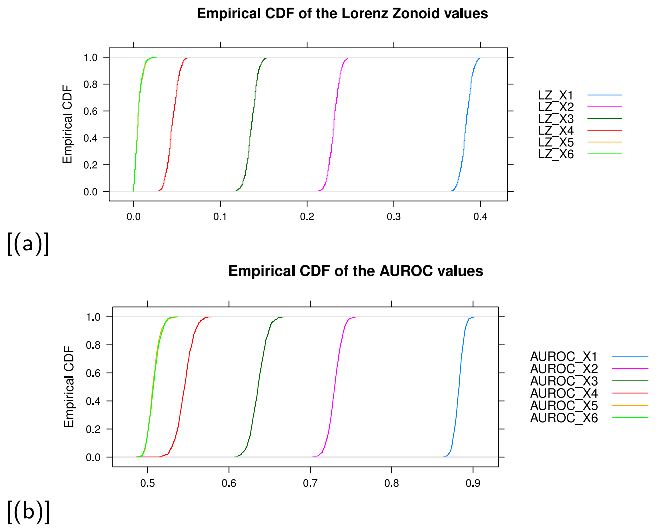


Fig. 7. [(a)] Lorenz Zonoid-based measures for each single predictor [(b)] AUROC measures for each single predictor.

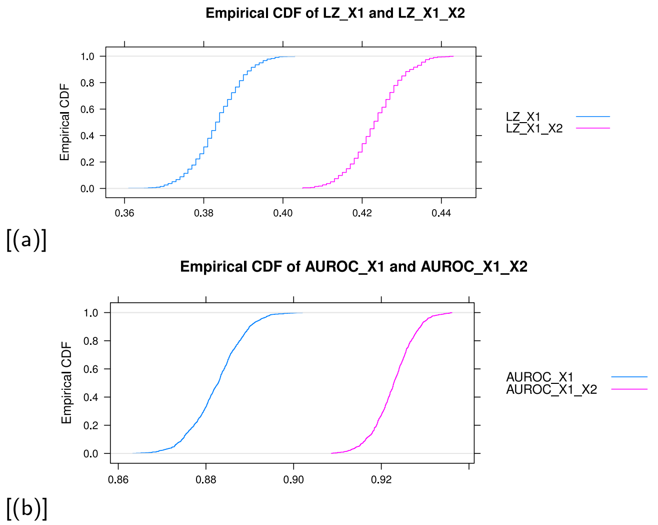


Fig. 8. [(a)] Lorenz Zonoid-based measures when including variable X_2 [(b)] AUROC measures when including variable X_2.

of 1,000 observations, with the only difference being that variable X_3 becomes significant at a significance level smaller than 0.001.

Last, the AIC procedure provides results which are similar to those related to the sample composed of 1,000 observations, both in the case of balanced and unbalanced data.

In summary, it seems that our proposal is the best performer, as it recognises the correct underlying correlation structure in all cases.

Results from simulation design 2.

For each simulated sample, the Lorenz Zonoid and AUROC measures, associated with any single predictor, are computed. The corresponding empirical distribution functions are depicted in Figs. 7 (a) and (b).

We can now compare the empirical cumulative distribution functions of the two measures, in terms of stochastic dominance. By definition, first order stochastic dominance provides an order relationship between cumulative distribution functions (see, e.g. [31]). Given two variables Y and X, and denoting with $F(x)$ and $F(y)$ the corresponding cumulative distribution functions, with $F(x), F(y) : \mathbb{R} \rightarrow [0,1]$, $F(x)$ dominates $F(y)$ if and only if $F(x) < F(y)$, $\forall x, y \in \mathbb{R}$.

From Fig. 7(a) and (b), one can gather that variables X_5 and X_6 present overlapping empirical cumulative distribution functions which are dominated by the empirical distribution functions of each of the remaining variables. Specifically, the variable X_1 empirical cumulative distribution function always dominates the others, meaning that in all the 1,000 extracted samples, it provides the highest contribution. To further strengthen this conclusion, the Page's test, which is based on pointwise comparison of the empirical cumulative distribution functions, is considered, testing the null hypothesis

$$H_0 : F(LZ(\hat{Y}_{X_1})) = F(LZ(\hat{Y}_{X_2})) = F(LZ(\hat{Y}_{X_3})) = F(LZ(\hat{Y}_{X_4})) = F(LZ(\hat{Y}_{X_5})) = F(LZ(\hat{Y}_{X_6})),$$

against the alternative hypothesis

$$H_1 : F(LZ(\hat{Y}_{X_1})) > F(LZ(\hat{Y}_{X_2})) > F(LZ(\hat{Y}_{X_3})) > F(LZ(\hat{Y}_{X_4})) > F(LZ(\hat{Y}_{X_5})) > F(LZ(\hat{Y}_{X_6})),$$

where $F(LZ(\hat{Y}_{X_1})), F(LZ(\hat{Y}_{X_2})), F(LZ(\hat{Y}_{X_3})), F(LZ(\hat{Y}_{X_4})), F(LZ(\hat{Y}_{X_5}))$, $F(LZ(\hat{Y}_{X_6}))$ are the cumulative distribution functions of the Lorenz Zonoid measures calculated on the predicted values derived from the simple logistic regression models. The same framework can be formalised for the AUROC measure, which indeed preserves the same ordering induced by the Lorenz Zonoid measure, except for variable X_6 anticipating variable X_5.

The p-values of the test are smaller than 0.001 for both the Lorenz Zonoid and AUROC measures, allowing to conclude that the ordered relationship specified by the alternative hypothesis is fulfilled and that variable X_1 has the highest impact on the response variable. It follows that such variable is included as the first into the model. The stepwise procedure continues to detect the ordering of inclusion of the remaining predictors. In the following steps, the hypotheses test can be generalised as follows:

$$H_0 : F(LZ(\hat{Y}_{X_1,\ldots,X_k})) = F(LZ(\hat{Y}_{X_1,\ldots,X_{k-1}}))$$

versus

$$H_1 : F(LZ(\hat{Y}_{X_1,\ldots,X_k})) > F(LZ(\hat{Y}_{X_1,\ldots,X_{k-1}})),$$

where $F(LZ(\hat{Y}_{X_1,\ldots,X_{k-1}}))$ and $F(LZ(\hat{Y}_{X_1,\ldots,X_k}))$ are the empirical cumulative distribution functions of the Lorenz Zonoid measures associated with the reduced and complex models, respectively.

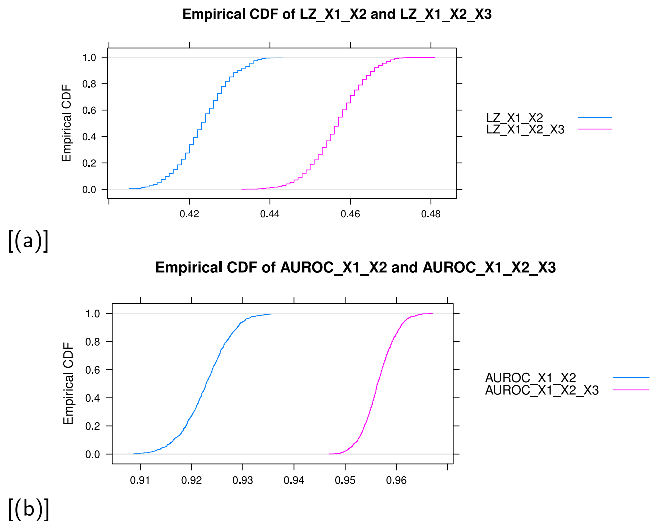
In Figs. 8–12, the empirical cumulative distribution functions of the Lorenz Zonoid and AUROC measures are displayed. In each plot we compare the empirical cumulative distribution function of the Lorenz Zonoid and AUROC measures associated with the complex model (including k variables) with that related to the reduced model (including $k-1$ variables). The stepwise procedure stops when the difference between the empirical cumulative distribution functions of the complex and reduced models is no more significant, according to the Page's test. The results of the procedure are summarised in Table 2.

By looking at Figs. 8 and 9, note that the inclusion of variables X_2 and X_3 has a relevant impact on the response variable. This result is confirmed by the Page's test p-values which are smaller than 0.001. Thus, both variables have to be included in the model.

On the contrary, when considering Fig. 10, it seems that the additional inclusion of variable X_4 does not significantly contribute to the explanation of the response variable. However, from the Page's test results, variable X_4 needs to be included in the model, for both the Lorenz Zonoid and the AUROC measures.

Finally, when adding variables X_5 and X_6, both AUROC and Lorenz Zonoid based stepwise procedures indicate no improvement in model accuracy. Indeed, the empirical distribution functions associated with the complex and reduced models overlap (see Figs. 11 and 12) and the Page's test p-values are greater than 10%.

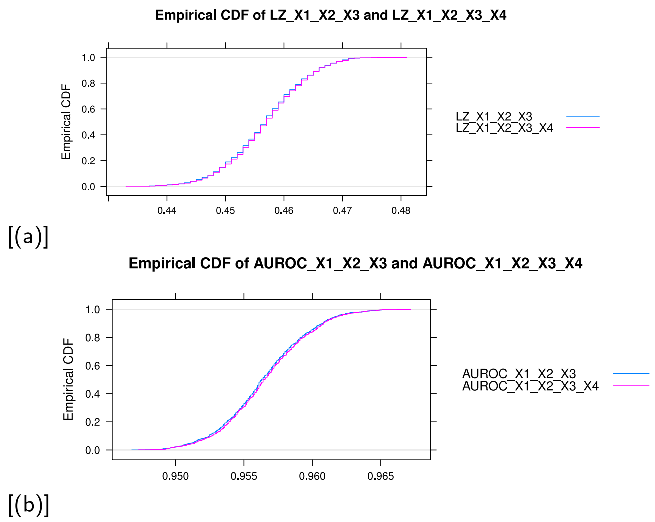
In summary, the simulation design 2, in which the stepwise procedure has been repeated 1,000 times, confirms the conclusions derived



[(a)]

[(b)]

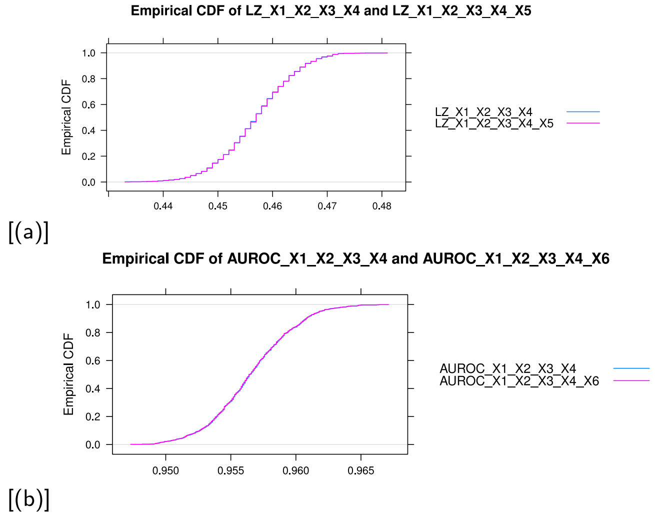
Fig. 9. [(a)] Lorenz Zonoid-based measure when including variable X_3 [(b)] AUROC measure when including variable X_3.



[(a)]

[(b)]

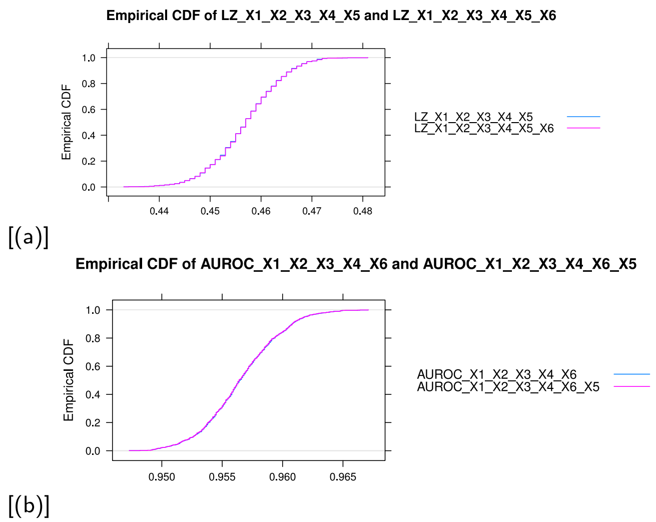
Fig. 10. [(a)] Lorenz Zonoid-based measure when including variable X_4 [(b)] AUROC measure when including variable X_4.



[(a)]

[(b)]

Fig. 11. [(a)] Lorenz Zonoid-based measure when including variable X_5 [(b)] AUROC measure when including variable X_6.



[(a)]

[(b)]

Fig. 12. [(a)] Lorenz Zonoid-based measure when including variable X_6 [(b)] AUROC measure when including variable X_5.

Table 2
Page test *p*-values.

Lorenz Zonoid measure		AUROC	
Additional included variable	*p*-value	Additional included variable	*p*-value
X_2	<0.001	X_2	<0.001
X_3	<0.001	X_3	<0.001
X_4	<0.001	X_4	<0.001
X_5	>0.10	X_6	>0.10
X_6	>0.10	X_5	>0.10

Table 3
Stepwise procedure based on Lorenz Zonoids.

Progressively included variables	*pay-off* (X_k) based on Lorenz Zonoids	*p*-value
X_{19}	–	<0.001
X_8	0.0537	<0.001
X_6	0.0095	<0.001
X_9	0.0324	<0.001
X_{16}	0.0023	<0.001
X_{14}	0.0049	<0.001
X_7	0.0013	<0.001
X_{13}	0.0321	<0.001
X_{11}	0.0026	<0.001
X_{18}	0.0161	<0.001
X_1	0.0377	<0.001
X_{15}	0.0285	<0.001
X_{12}	0.0061	>0.10

from the previous one based on only a single sample. When considering a significance level of $\alpha = 0.05$, both our proposal and the AUROC measure lead to select the model built on the X_1, X_2, X_3 and X_4 variables.

Results from simulation design 3.

The results derived from the simulation involving the high-dimensional setting are illustrated in Tables 3 and 4. Specifically, in Tables 3 and 4, the pay-offs based on Lorenz Zonoids and AUROC are highlighted. The stepwise procedure ends when the contribution provided by an additional predictor results no more significant, according to the Lorenz Zonoid or to the DeLong tests.

At a significance level $\alpha = 0.05$, the Lorenz Zonoid-based model selection procedure leads to include twelve variables out of the nineteen originally considered. Conversely, the AUROC-based model selection procedure selects six variables. From Tables 3 and 4 note also that the first five variables progressively chosen and added into the model

Table 4
Stepwise procedure based on AUROC.

Progressively included variables	*pay-off* (X_k) based on AUROC	*p*-value
X_{19}	–	<0.001
X_8	0.0558	<0.001
X_6	0.0106	<0.001
X_9	0.0224	<0.001
X_{16}	0.0030	0.0152
X_{11}	0.0070	0.0020
X_7	0	>0.10

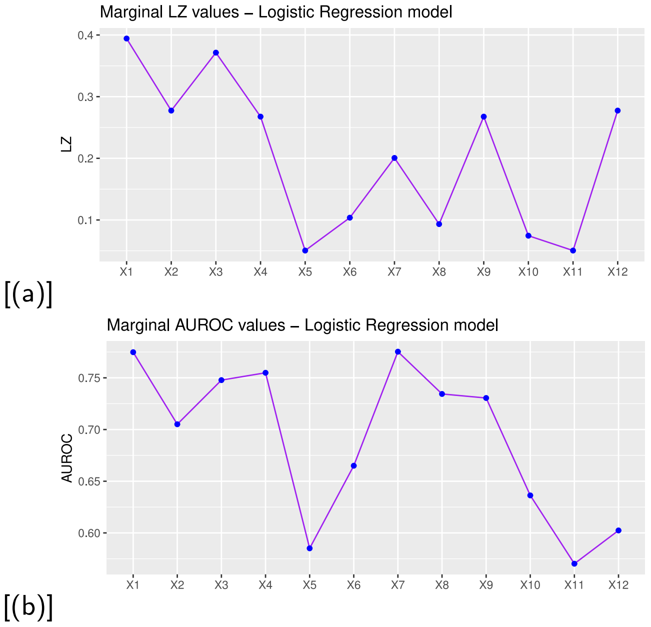


Fig. 13. [(a)] Logistic regression model - Marginal Lorenz Zonoid [(b)] Logistic regression model - Marginal AUROC.

are the same for both Lorenz Zonoid and AUROC. Whereas the sixth selected predictor is different: the Lorenz Zonoid chooses X_{14} whereas AUROC chooses X_{11}.

We remark that, although our proposal leads to a model that is more complex than that chosen with the AUROC measure, it is more in line with the true underlying structure, as the selected variables are those that have a stronger relationship ($\rho \geq 0.23$) with the response variable.

4. Application

4.1. Data

In this section we apply our proposed method to data supplied by Modefinance, a European Credit Assessment Institution (ECAI) that specialises in credit scoring for P2P platforms focused on SME commercial lending. The whole dataset is described by [32] to which we refer for further details. Here we focus on the twelve explanatory variables selected by the Authors: Total Assets/Total Liabilities (X_1); Current Assets/Current Liabilities (X_2); (Profit or Loss before tax+Interest paid)/Total Assets (X_3); Return on Equity (X_4); Operating Revenues/Total Assets (X_5); Interest paid/(Profit before taxes+Interest paid) (X_6); EBITDA/Interest paid (X_7); EBITDA/Operating Revenues (X_8); EBITDA/Sales (X_9); Trade Receivables/Operating Revenues (X_{10}); Inventories/Operating Revenues (X_{11}); Turnover (X_{12}).

The data on the above mentioned explanatory variables are extracted from the balance-sheets of 15,045 SMEs, mostly based in Southern Europe, for the year 2015. The data on the response variable are obtained from information about the status (0 = active, 1 = defaulted) of each SME one year later (2016), as collected from the official registers of bankruptcy. Note that the observed proportion of defaulted companies is equal to 10.9%.

4.2. Results

With the same data, [32] have constructed logistic regression scoring models that aim at estimating the probability of default of each company, using the available explanatory data and, in addition, network centrality measures that are obtained from similarity networks.

To improve the predictive performance of the model, [5] have applied the Gradient Boosting (XGBoost) tree algorithm, and obtained a substantial increase in predictive performance: the Area Under the ROC Curve (AUROC) increases from a value of 0.81, obtained with the application of the logistic regression, to a value of 0.93, obtained with the Gradient Boosting method.

The same Authors identify the variables X_1 and X_3 as the variables that rank highest in terms of the Shapley value explanation of the probability of default, a result that is quite consistent with most credit scoring models, that typically include, among the explanatory variables of credit default, a measure of financial leverage (such as variable X_1) and a measure of profitability (such as variable X_3).

We consider the same data, and the same twelve explanatory variables as in [5], to which we apply a logistic regression model after the data is randomly split in a training set (80%) and a test set (20%). We then calculate, on the test set, the contribution of each of the explanatory variables to the estimate of the probability of default, using our proposed Lorenz Zonoid based approach. Additionally to what Shapley values can do, we provide contributions that are normalised in the $[0, 1]$ interval, and whose additional value can be assessed in terms of its statistical significance. Doing so, we show how a model comparison procedure based on the Lorenz Zonoids can improve the explainability of a machine learning model, choosing a parsimonious set of explanatory variables while maintaining a high predictive accuracy.

The implementation of our proposed model comparison procedure starts by computing the marginal contribution of each single explanatory variable X_k, for $k = 1, \ldots, 12$, to the explanation of the probability of default. The marginal contributions are determined by building twelve simple logistic regression models, each of them involving only one of the twelve predictors, and calculating the Lorenz Zonoid value $LZ(\hat{Y}_{X_k})$ for each of them. This leads to a ranking of the explanatory variables, to be used in the stepwise procedure. In the forward perspective, the variable with the highest $LZ(\hat{Y}_{X_k})$ value is selected as the first variable to be included in the model. Then, progressively, more complex models are implemented by introducing at each step an additional variable, according to the obtained variable ranking. Conversely, in the backward perspective, the variable with the lowest $LZ(\hat{Y}_{X_k})$ value is selected as the first variable to be removed from the full model and, then, progressively, simpler models are implemented by deleting at each step according to the reversed variable ranking.

The marginal contributions of each considered explanatory variable, measured in terms of $LZ(\hat{Y}_{X_k})$, along with the corresponding value of the AUROC, for comparison purposes, are displayed in Fig. 13(a) and (b), respectively.

From Fig. 13(a), the variables that contribute the most are variables X_1 and X_3, as in [5], followed by X_9 and, then, the others. The least important results to be X_{11}. Differently, from Fig. 13(b), the most important variable is X_7, followed by X_1, X_4 and the others. The least important variable results to be X_{11}.

We have then implemented a Lorenz Zonoid and an AUROC forward stepwise procedure starting from X_1 and, then, progressively adding the other variables, up to the full model. At each step, the additional contribution of the new added variable is measured by *pay-off* (X_k). For the sake of completeness, we also report the F_1 accuracy index, a standard practice as the AUROC, in the seventh column of Table 5.

Table 5
Logistic regression model (forward stepwise) - Marginal contributions ($LZ(\hat{Y}_{X_k})$); additional contributions (*pay-off* (X_k)); significance (p-value) of the additional contributions; F_1 metric.

ID	Variable	$LZ(\hat{Y}_{X_k})$	ID of the included variables	*pay-off* (X_k)	p-value	F_1
1	TA/TL	0.3943	1	–	–	–
3	(PLBT+IP)/TA	0.3714	1, 3	0.0544	<0.001	0.3844
9	EBITDA/S	0.3244	1, 3, 9	0.0081	<0.001	0.3865
12	TO	0.3061	1, 3, 9, 12	0.0002	0.2069	0.3865

Legend: TA/TL = Total assets/Total Liabilities; (PLBT+IP)/TA = (Profit or Loss before tax+Interest paid)/Total Assets; EBITDA/S = EBITDA/Sales; TO = Turnover.

Table 6
Logistic regression model (forward stepwise) - Marginal contributions (AUROC); additional contributions (difference of AUROC); significance (p-value) of the additional contributions; F_1 metric.

ID	Variable	$AUROC_{X_k}$	ID of the included variables	*pay-off* (X_k)	p-value	F_1
7	EBITDA/IP	0.7753	7	–	–	–
1	TA/TL	0.4113	7, 1	0.0016	0.9050	0.2942

Legend: EBITDA/IP = EBITDA/Interest paid; TA/TL = Total assets/Total Liabilities.

To decide when to stop the procedure, we apply the statistical test proposed in Section 2.3 and, then, continue the process until the additional contribution is significantly different from zero. In this way the selected model represents a good trade-off between predictive accuracy (which increases with model complexity) and explainability (which decreases with model complexity).

The results of the procedure, based on the Lorenz Zonoid pay-offs, are illustrated in Table 5.

Looking at Table 5 and, in particular, at the p-values of the test, reported in the sixth column, we obtain that the best model includes three explanatory variables: X_1, X_3, as in the reference literature (see, e.g. [5]), and also variable X_9. For comparison purposes, Table 6 highlights the results of the procedure based on the AUROC differences.

In agreement with Fig. 13(b), Table 6 shows that the best model contains variable X_7 (EBITDA/Interest paid). In addition, the DeLong test indicates to stop at that point, leading to a very parsimonious model, with only one variable. We remark that the result of the AUROC based procedure is not in line with the literature, as it includes in the model a measure of profitability but not a measure of financial leverage.

We also remark that, for robustness purposes, we have implemented a backward stepwise procedure, for both the Lorenz Zonoid pay-off and the AUROC. The results have confirmed the significance of the variables contained in the models selected with the forward procedure.

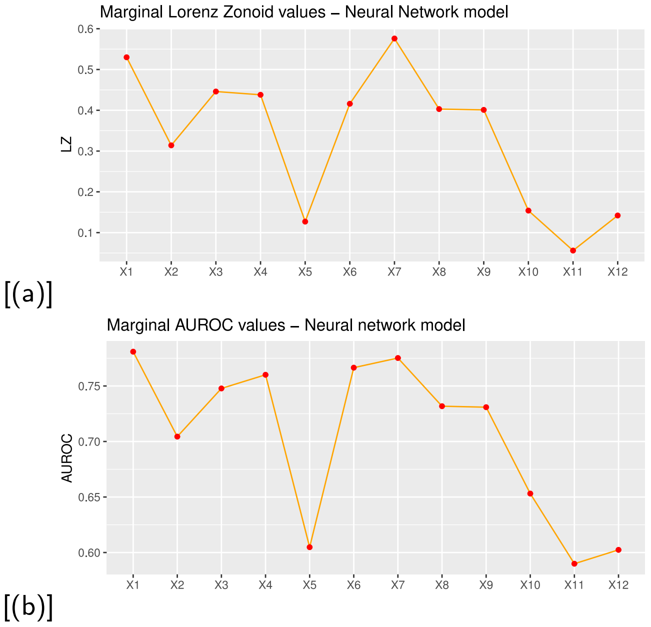
We now report the results of model comparison, for a neural network model built (without loss of generality) with five neurons in the hidden layer. Specifically, we apply a feedforward multilayer perceptron neural network characterised by straight forward networks (see, e.g. [33]).

The behaviour of the $LZ(X_k)$ and of the AUROC for each explanatory variable is shown in Figs. 14 (a) and (b), respectively.

From Fig. 14(a), the variables that contribute the most are variables X_7 and X_1, and similarly in Fig. 14(b), although in a reversed order. Additionally, Fig. 14(b) indicates a high importance also for variable X_6. In both cases, the least important results to be X_{11}.

The results of the stepwise procedure for the neural network models are reported, respectively, in Table 7, for the $LZ(X_k)$ measure; and in Table 8, for the AUROC measure.

From Table 7 we obtain that, similarly to what occurs for logistic regression models, the neural network procedure selects two variables, and one is X_1. However, the second variable is X_6 and not X_3. From a financial viewpoint, the results are indeed similar, as both X_3 and X_6 measure profitability, whereas X_1 indicates financial leverage.


Marginal Lorenz Zonoid values – Neural Network model

Marginal AUROC values – Neural network model

Fig. 14. [(a)] Neural network model - Marginal Lorenz Zonoid [(b)] Neural network model - Marginal AUROC.

Table 7
Neural network model (forward stepwise) - Marginal contributions ($LZ(\hat{Y}_{X_k})$); additional contributions (*pay-off* (X_k)); significance (p-value) of the additional contributions; F_1 metric.

ID	Variable	$LZ(\hat{Y}_{X_k})$	ID of the included variables	*pay-off* (X_k)	p-value	F_1
1	TA/TL	0.5343	1	–	–	–
6	IP/(PBT+IP)	0.4684	1, 6	0.0212	<0.001	0.4154
7	EBITDA/IP	0.4574	1, 6, 7	0.0009	0.7806	0.400

Legend: TA/TL = Total assets/Total Liabilities; IP/(PBT+IP) = Interest paid/(Profit before taxes + Interest paid); EBITDA/IP = EBITDA/Interest paid.

Table 8
Neural network model (forward stepwise) - Marginal contributions (AUROC); additional contributions in terms of AUROC difference; significance (p-value) of the additional contribution; F_1 metric.

ID	Variable	$AUROC_{X_k}$	ID of the included variables	*pay-off* (X_k)	p-value	F_1
1	TA/TL	0.7809	1	–	–	–
7	EBITDA/IP	0.7752	1, 7	0.0219	0.0426	0.4366
6	IP/(PBT+IP)	0.7665	1, 7, 6	0.0013	0.8348	0.4000

Legend: TA/TL = Total assets/Total Liabilities; EBITDA/IP = EBITDA/Interest paid; IP/(PBT+IP) = Interest paid/(Profit before taxes+Interest paid).

Similar conclusions can be derived when the AUROC metric is employed in place of the Lorenz Zonoid pay-off. Table 8 shows that, again, two explanatory variables are included in the selected model. While the first one is confirmed to be X_1, the second is X_7, instead of X_6: another function of the profitability. These results are confirmed when a backward selection procedure is implemented, for robustness.

In summary, the application of the procedure to neural networks shows that both the Lorenz Zonoid and the AUROC model selection lead to choose a model with two variables (one measuring leverage and one measuring profitability), which represents a very good trade-off between explainability and accuracy. On one hand, the model is more explainable than the full model, as the response depends significantly only on two variables, and we know which ones (whereas a full neural

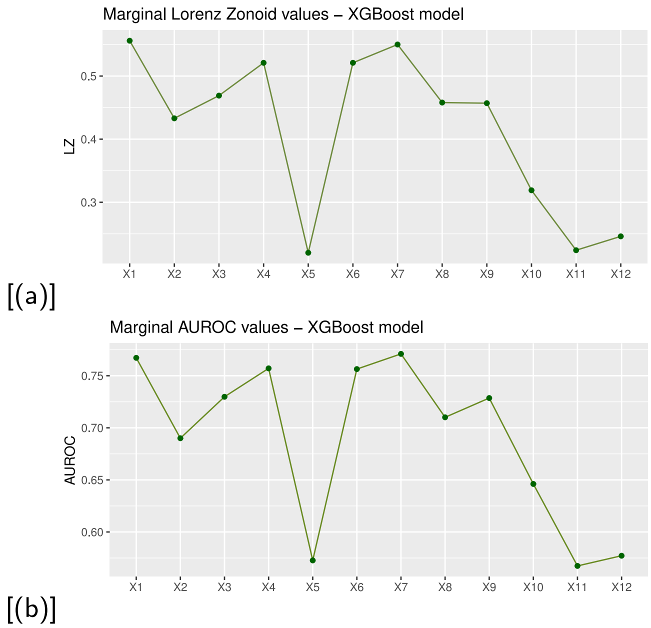


Fig. 15. [(a)] XGBoost model - Marginal Lorenz Zonoid [(b)] XGBoost model - Marginal AUROC.

Table 9
XGBoost model (forward stepwise)- Marginal contribution in terms of each single explanatory variable ($LZ(\hat{Y}_{X_k})$); marginal contribution in terms of any additional explanatory variable (*pay-off* (X_k)); the marginal contribution significance (*p*-value); F_1 metric.

ID	Variable	$LZ(\hat{Y}_{X_k})$	ID of the included variables	*pay-off* (X_k)	*p*-value	F_1
1	TA/TL	0.5565	1	–	–	–
7	EBITDA/IP	0.5496	1, 7	0.0747	<0.001	0.4170
6	IP/(PBT+IP)	0.5212	1, 7, 6	0.0052	<0.001	0.4386
4	ROE	0.5210	1, 7, 6, 4	0.0035	0.0758	0.4390

Legend: TA/TL = Total assets/Total Liabilities; EBITDA/IP = EBITDA/Interest paid; IP/(PBT+IP) = Interest paid/(Profit before taxes+Interest paid); ROE = Return on Equity.

Table 10
XGBoost model (forward stepwise) - Marginal contributions (AUROC); additional contributions in terms of AUROC difference; significance (*p*-value) of the additional contribution; F_1 metric.

ID	Variable	$LZ(\hat{Y}_{X_k})$	ID of the included variables	*pay-off* (X_k)	*p*-value	F_1
7	EBITDA/IP	0.7710	7	–	–	–
1	TA/TL	0.7672	7, 1	0.0362	<0.001	0.4170
4	ROE	0.5210	7, 1, 4	0.0068	0.1282	0.4105

Legend: EBITDA/IP = EBITDA/Interest paid; TA/TL = Total assets/Total Liabilities; ROE = Return on Equity.

network model is a black-box); on the other hand, the model is accurate as its predictive accuracy is not significantly improved making it more complex (adding more variables).

We can apply our procedure, in the same way, to another type of machine learning model: the XGBoost, which belongs to the class of tree models. The results are illustrated, from a graphical view point, in Fig. 15(a) and (b); and are specified with numerical details in Tables 9 and 10.

Fig. 15(a) shows that variables X_1 and X_7, followed by X_6, are the factors with the highest impact on the probability of default. Fig. 14(b) shows a similar result, swapping X_1 with X_7 and replacing X_6 with X_4.

Table 11
Predictive accuracy of the selected and full models.

	AUROC selected model	AUROC full model
Logistic regression model	0.8037	0.8045
Neural network model	0.7800	0.7810
XGBoost	0.8110	0.8557

In terms of model selection, both procedures lead to select a model that contains X_1 and X_7. Additionally, the Lorenz Zonoid based procedure includes also X_6, leading to a more complex model, with three significant contributions. We remark that also in this case, the backward model search confirms the selected variables.

The conclusions that can be drawn from the XGBoost model selection procedure are in line with those from the neural network model. Overall, the empirical findings from our analysis can be summarised with the conclusion that the proposed model selection procedure, based on the Lorenz Zonoids, is able to simplify a black-box machine learning model into an explainable model.

From a financial viewpoint, all models indicate that the most important variables for credit scoring are: a measure of financial leverage and a measure of profitability, confirming the previous analysis of [5,32] on the same data.

A natural question that arises is: which of the three model champions is the best model overall, both in absolute terms (predictive accuracy) and in relative terms, with respect to the full model (explainability)? To answer this question, the logistic regression, neural network and XGBoost models selected with the Lorenz Zonoid approach are compared in terms of the predictive accuracy of their full model and selected model. To achieve an "external" evaluation, predictive accuracy is evaluated using the AUROC measure. The results can be found in Table 11.

From Table 11 note that the best machine learning model, in terms of predictive accuracy, is the XGBoost model, with an AUROC of 0.811; whereas the neural network model is the worst one, with an AUROC of 0.78. On the other hand, the XGboost model is the least explainable model: differently from what occurs for the logistic regression and neural networks, the AUROC of the full model reduces substantially and in a significant way (*p*-value greater than 0.05) moving to the reduced model.

5. Concluding remarks

The paper proposes to improve machine learning models by means of a model selection methodology, based on the Lorenz Zonoids, which allows to maintain a high predictive accuracy, explaining the predictions with a parsimonious set of explanatory variables.

We remark that our approach is quite general, and can be applied to different types of machine learning models: support vector machines, neural networks and deep learning, random forest and gradient boosting.

In the case of a binary response, the approach is consistent with the results that can be obtained applying the well known AUROC accuracy measure.

Further advantages of our proposed procedure are: its generality (in the paper we have considered a binary response, but the same tool can be applied for ordinal or continuous response, differently from what occurs for the AUROC); its computational efficiency (we do not need to calculate the Lorenz Zonoids of all models, but only of those considered in the stepwise path, differently from what occurs with the Shapley value approach to explainability).

The application of the proposal to simulated data has shown that it is capable to select the correct underlying model, and to take into account the sample size. The application of the proposal to a real credit scoring database has shown its capability to identify, as relevant variables, those that concern the profitability and the financial leverage of the companies asking for credit, in line with the subject matter literature.

We believe that the proposed method could be employed as a use case to improve the compliance of Artificial Intelligence applications in finance to the emerging regulations, such as the European AI act (http:artificialintelligenceact.eu).

Further research may focus on the application of the methodology to other machine learning applications, that involve different type of variables: ordinal or continuous. The generality of the proposed measure allows to do so, differently from what occurs with available metrics such as the AUROC and the MSE.

CRediT authorship contribution statement

Paolo Giudici: Conceptualization, Methodology, Supervision, Final approval of the revised version to be submitted. **Alex Gramegna:** Software, Data curation. **Emanuela Raffinetti:** Conceptualization, Methodology, Formal analysis, Revision of the article.

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.

Data availability

The data that has been used is confidential.

Acknowledgements

The authors acknowledge support from the European Horizon2020 Periscope programme, contract number n. 101016233, and from the European xAIM (eXplainable Artificial Intelligence in healthcare Management) project supported by the CEF Telecom under Grant Agreement No. INEA/CEF/ICT/A2020/2276680.

References

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

[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. [Accessed 31 October 2022].

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

[4] Molnar C. Interpretable machine learning - A guide for making black box models explainable. 2020, https://cristophm.github.io/interpretable-ml-book. [Accessed 31 October 2022].

[5] Bussmann N, Giudici P, Marinelli D, Papenbrock J. Explainable machine learning in credit risk management. Comput Econ 2020;57:203–16. http://dx.doi.org/10.1007/s10614-020-10042-0.

[6] Mantegna RN, Stanley HE. Introduction to econophysics: correlations and complexity in finance. Cambridge University Press; 1999.

[7] Giudici P, Raffinetti E. Lorenz model selection. J Classif 2020;37:754–68. http://dx.doi.org/10.1007/s00357-019-09358-w.

[8] Rossi R, Murari A, Gaudio P, Gelfusa M. Upgrading model selection criteria with goodness of fit tests for practical applications. Entropy 2020;22:1–13. http://dx.doi.org/10.3390/e22040447.

[9] Diebold F, Mariano R. Comparing predictive accuracy. J Bus Econ Stat 1995;13:253–63. http://dx.doi.org/10.1080/07350015.1995.10524599.

[10] Giudici P, Raffinetti E. Shapley-Lorenz explainable artificial intelligence. Expert Syst Appl 2021;167:1–9. http://dx.doi.org/10.1016/j.eswa.2020.114104.

[11] Angelini E, Tollo G, Roli A. A neural network approach for credit risk evaluation. Q Rev Econ Financ 2008;48:733–55. http://dx.doi.org/10.1016/j.qref.2007.04.001.

[12] Hsieh NC, Hung LP. A data driven ensemble classifier for credit scoring analysis. Expert Syst Appl 2010;37:534–45. http://dx.doi.org/10.1016/j.eswa.2009.05.059.

[13] Ravi Kumar P, Ravi V. Bankruptcy prediction in banks and firms via statistical and intelligent techniques–A review. Eur J Oper Res 2007;180:1–28. http://dx.doi.org/10.1016/j.ejor.2006.08.043.

[14] Pacelli V, Azzollini M. An artificial neural network approach for credit risk management. J Intel Learn Syst Appl 2011;3:103–12. http://dx.doi.org/10.4236/jilsa.2011.32012.

[15] Zhang C, Ma Y. Ensemble machine learning methods and applications. Springer; 2012.

[16] Friedman J. Greedy boosting approximation: A gradient boosting machine. Ann Stat 2001;29:1189–232. http://dx.doi.org/10.1214/aos/1013203451.

[17] Koshevoy G, Mosler K. The Lorenz zonoid of a multivariate distribution. J Am Stat Assoc 1996;91:873–82. http://dx.doi.org/10.1080/01621459.1996.10476955.

[18] Lorenz MO. Methods of measuring the concentration of wealth. Publ Am Stat Assoc 1905;70:209–19. http://dx.doi.org/10.1080/15225437.1905.10503443.

[19] Kent JT, O'Quigley J. Measures of dependence for censored survival data. Biometrika 1988;75:525–34. http://dx.doi.org/10.1093/biomet/75.3.525.

[20] Schemper M. Predictive accuracy and explained variation. Stat Med 2003;22:2299–308. http://dx.doi.org/10.1111/j.0006-341x.2000.00249.x.

[21] Lerman R, Yitzhaki S. A note on the calculation and interpretation of the Gini index. Econom Lett 1984;15:363–8. http://dx.doi.org/10.1016/0165-1765(84)90126-5.

[22] Koroljuk VS, Borovskich YV. Theory of u-statistics. Mathematics and its applications, Dordrecht: Springer; 1994.

[23] Schechtman E, Yitzhaki S. A measure of association based on Gini's mean difference. Commun Stat-Theor M 1987;16:207–31. http://dx.doi.org/10.1080/03610928708829359.

[24] Sang Y, Dang X, Zhao Y. Jackknife empirical likelihood methods for Gini correlations and their equality testing. J Stat Plan Infer 2019;199:45–59. http://dx.doi.org/10.1016/j.jspi.2018.05.004.

[25] Kendall MG, Stuart A. The advanced theory of statistics. London: Griffin; 1969.

[26] Yitzhaki S, Schechtman E. The Gini methodology: a primer on a statistical methodology. New York: Springer; 2013.

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

[28] Arvesen JN. Jackknifing U-statistics. Ann Math Statist 1969;40:2076–100. http://dx.doi.org/10.1214/aoms/1177697287.

[29] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics 1988;44:837–45. http://dx.doi.org/10.2307/2531595.

[30] Page EB. Ordered hypotheses for multiple treatments a significance test for linear ranks. J Amer Statist Assoc 1963;58:216–30. http://dx.doi.org/10.2307/2282965.

[31] Heathcote A, Brown S, Wagenmakers EJ, Eidels A. Distribution-free tests of stochastic dominance for small samples. J Math Psychol 2010;54:454–63. http://dx.doi.org/10.1016/j.jmp.2010.06.005, (2010).

[32] Giudici P, Hadji-Misheva B, Spelta A. Network based credit risk models. Qual Eng 2020;32:199–211. http://dx.doi.org/10.1080/08982112.2019.1655159.

[33] Schmidhuber J. Deep learning in neural networks: An overview. Neural Netw 2015;61:85–117. http://dx.doi.org/10.1016/j.neunet.2014.09.003.

Paolo Giudici is Full Professor of Statistics at the Department of Economics and Management of the University of Pavia. His main research contributions are in: models for financial stability, correlation networks for financial technologies, operational risk management models, and model diagnostics, predictive accuracy, and explainability. He is coordinator of 12 funded scientific projects, among which the European Horizon2020 projects "PERISCOPE: Pan-European response to the impacts of covid-19 and future pandemics and epidemics (2020–2023)" and "FIN-TECH: Financial supervision and Technological compliance" (2019–2020). He is chief editor of Artificial Intelligence in Finance, Frontiers, and AE of Digital Finance, Springer, and of Risks MDPI.

Alex Gramegna is a Ph.D. student at the University of Pavia. His main research interests are: Artificial Neural Networks, machine learning algorithms and Explainable Artificial Intelligence (XAI).

Emanuela Raffinetti is Assistant Professor of Statistics at the Department of Economics and Management of the University of Pavia. Her research activity is on: Explainable Artificial Intelligence (XAI) methods; predictive accuracy measures; Machine Learning model validation methods; assessment of operational and cyber risks; dependence analysis; sub-sampling methods; inequality measures for income distributions. She is Associate Editor of the Frontiers in Artificial Intelligence journal.