Co-financed by the Connecting Europe
Facility of the European Union

Explainability of Artificial Intelligence methods

Trustworthy AI - Lecturer: Emanuela Raffinetti; Python instructor: Alex Gramegna
E-mail: emanuela.raffinetti@unipv.it; alex.gramegna01@universitadipavia.it

Artificial intelligence: risks

- Cyber technologies bring also risks, such as cyber risks and model risks.
- Policy makers, regulators, supervisors and standardisation bodies around the world are promoting AI risk management, measuring AI risks to make it sustainable.
- A reference model is the European Artificial Intelligence Act, which requires risk management of high risk AI applications.

Requirements for a trustworthy AI

For a trustworthy AI the key principles of accuracy, sustainability, fairness, explainability have to be developed.

Model comparison

- Machine Learning models can be compared in terms of their associated predictive capability (accuracy).
- Complex Machine Learning models can be evaluated in terms of their interpretability (explainability).

Evaluation of a specific model

- A specific model can be evaluated in terms of its robustness with respect to perturbed data (sustainability).
- A specific model can be evaluated in terms of equality with respect to the different groups (gender, ethnicity, . . .) composing the population (fairness).

Explainable Artificial Intelligence

Black box Artificial Intelligence (AI) is not suitable in regulated financial services. Thus, eXplainable AI (XAI) methods are necessary.

Definition

Explainability means that an interested stakeholder can comprehend the main drivers of a model-driven decision.

Problem

- "Simple" machine learning-models provide a high interpretability but, possibly, a limited predictive accuracy.
- "Complex" machine learning-models provide a high predictive accuracy at the expense of a limited interpretability.

Solution

Boosting highly accurate machine learning-models with novel methodologies that can explain their predictive output (local and global explanation based-approaches).

Local explanation based-approach - The Shapley-value based approach

Shapley values were originally introduced by Shapley (1953) to measure the contribution of each explanatory variable for each point prediction of a machine learning model.

Premises

Let:

- $i = 1, \ldots, n$ be a statistical unit, whose (multivariate) characteristics Y_i are to be predicted with a machine learning model;
- $\hat{Y}_i = \hat{f}(X_i)$ indicate the predicted value for the response vector Y_i, based on an explanatory vector of characteristics X_i, obtained with a machine learning model.

Definition

Given K explanatory variables, the marginal contribution of a variable $X_k, (k = 1, \ldots, K)$ can be expressed as

$$\phi(\hat{f}(X_i)) = \sum_{X' \subseteq \mathscr{C}(X) \setminus X_k} \frac{|X'|!(K - |X'| - 1)!}{K!} [\hat{f}(X' \cup X_k)_i - \hat{f}(X')_i], \qquad (1)$$

where $\mathscr{C}(X) \setminus X_k$ 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, $\hat{f}(X' \cup X_k)_i$ and $\hat{f}(X')_i$ are the predictions associated with all the possible model configurations including variable X_k and excluding variable X_k, both calculated for the unit i.

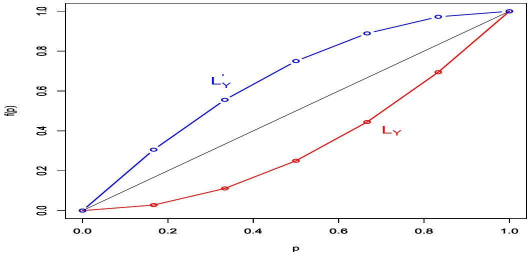
Global explanation-based approach - Lorenz Zonoid

Lorenz Zonoids were introduced by Koshevoy and Mosler (1996) as a generalization of the Lorenz curve in d dimensions. When $d = 1$, the Lorenz Zonoid corresponds with the well known Gini coefficient.

Main steps to build the Lorenz Zonoid

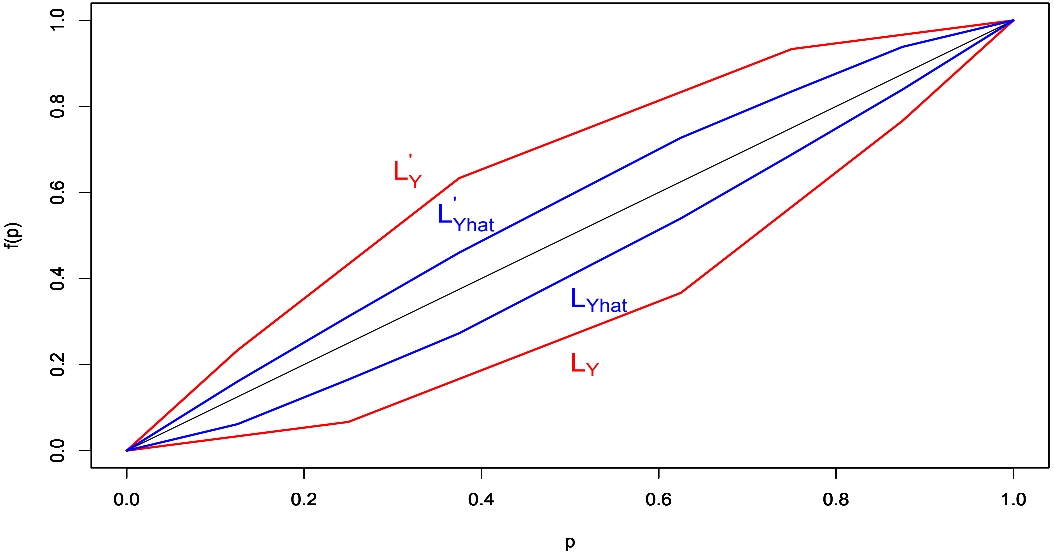
Given a variable Y and n observations,

- build the Y Lorenz curve (L_Y) by re-ordering the Y values in non-decreasing sense, whose points have coordinates $(i/n, \sum_{j=1}^{i} y_{r_j}/(n\bar{y}))$, for $i = 1, \ldots, n$, where r and $\bar{y}$ indicate the (non-decreasing) ranks of Y and the Y mean value, respectively;
- build the Y dual Lorenz curve (L'_Y) by re-ordering the Y values in a non-increasing sense, whose points have 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 correspond to the Lorenz Zonoid.



The inclusion property

Consider a MLM, such that $\hat{Y} = \hat{f}(X_1, \ldots, X_K)$. It results that $LZ(\hat{Y}) \subseteq LZ(Y)$.



Features and formalisation of the Lorenz Zonoids

Features:

- $LZ(\cdot)$ is a measure of the mutual variability that characterizes a phenomenon of interest.
- $LZ(\hat{\cdot})$ is used to assess the contribution of additional independent variables in explaining the variability of the response variable.

Formalization

Let:

- $LZ(Y)$ be the Lorenz Zonoid of the response variable Y;
- X_1 be an independent variable such that $\hat{Y}_{X_1} = f(X_1)$;
- $LZ(\hat{Y}_{X_1})$ be the Lorenz Zonoid of $\hat{Y}_{X_1}$;
- X_2 be an additional independent variable such that $\hat{Y}_{X_2} = f(X_2)$;
- $LZ(\hat{Y}_{X_2})$ be the Lorenz Zonoid of $\hat{Y}_{X_2}$.

The Lorenz Zonoid of a variable may be expressed by resorting to the covariance operator, i.e.

$$LZ(Y) = \frac{2Cov(Y, r(Y))}{n\mu}, LZ(\hat{Y}_{X_1}) = \frac{2Cov(\hat{Y}_{X_1}, r(\hat{Y}_{X_1}))}{n\mu}$$

$$\text{and } LZ(\hat{Y}_{X_2}) = \frac{2Cov(\hat{Y}_{X_2}, r(\hat{Y}_{X_2}))}{n\mu}.$$

where $\mu = E(Y)$ and $r(\cdot)$ are rank scores.

The Lorenz Zonoids for sample data

Given a sample data of size n, the Lorenz Zonoids may be re-expressed as:

$$LZ(y) = \frac{2Cov(y, r(y))}{n\bar{y}}, LZ(\hat{y}_{x_1}) = \frac{2Cov(\hat{y}_{x_1}, r(\hat{y}_{x_1}))}{n\bar{y}}$$

$$\text{and } LZ(\hat{y}_{x_2}) = \frac{2Cov(\hat{y}_{x_2}, r(\hat{y}_{x_2}))}{n\bar{y}}$$

where y, $\hat{y}_{x_1}$ and $\hat{y}_{x_2}$ are the vectors of the observed and estimated values, $r(y)$, $r(\hat{y}_{x_1})$ and $r(\hat{y}_{x_2})$ are the ranks of the observed and estimated values, and $\bar{y}$ is the sample mean.

The Shapley-Lorenz decomposition

The mathematical derivation of the Shapley-Lorenz decomposition can be obtained through the following steps:

- replace $LZ(\cdot)$ in place of $\hat{f}(\cdot)$ in the Shapley expression in (1);
- define the marginal contribution associated with the additional variable X_k as:

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

where $LZ(\hat{Y}_{X' \cup X_k})$ and $LZ(\hat{Y}_{X'})$ describe the (mutual) variability explained by the models including the $X^{'} \cup X_k$ variables and the $X^{'}$ variables, respectively.

Application to Bitcoin prediction

We aim to build a model able to predict bitcoin prices.

Data description

- The data provide the daily bitcoin prices of the Coinbase crypto exchange, from 18 May, 2016 to 30 April, 2018;
- The Coinbase price is the response variable to be predicted by the available explanatory variables;
- The candidate explanatory variables are Oil, Gold and SP500 prices are taken into account.

The Shapley-Lorenz decomposition

A linear regression model is implemented as our selected candidate machine learning model. The Shapley Lorenz marginal contributions, associated with the inclusion of SP500, Gold and Oil, are determined as follows:

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

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

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

The variance decomposition

The variance decomposition associated with the same variables, under the assumption of a linear model, is provided by:

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

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

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

Results

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

Conclusions:

- employing the Lorenz Shapley approach, variable SP500 provides the highest marginal contribution in the prediction of the Bitcoin prices, while the other two give a minimal contribution;
- findings from the Shapley Lorenz approach are quite similar to those obtained with the linear R^2-based Shapley approach, indicating their robustness;
- the Global Shapley values, obtained summing the Shapley variable contributions across all units, are the least interpretable.

Model selection based on Lorenz Zonoids

The Shapley-Lorenz approach appears computationally intensive, especially if dealing with huge datasets involving a large number of predictors.

In order to meet the sustainability requirement, a pre-selection of the most important predictors has to be set.

By exploiting the Lorenz Zonoid inclusion property, marginal and partial contributions provided by each predictor can be determined giving rise to 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.

We remark that, as well as the Shapley-Lorenz decomposition, the proposed approach is model agnostic not depending on the type of target variable and data to be analised.

Marginal Gini contribution

Let Y and X_k, for $k = 1, \ldots, K$, the response and the k-th explanatory variables, respectively.

The Marginal Gini Contribution associated with the k-th explanatory variable is defined as

$$
\begin{aligned}
MGC_{Y|X_k} &= \frac{LZ(\hat{Y}_{X_k})}{LZ(Y)} = \frac{2Cov(\hat{Y}_{X_k}, r(\hat{Y}_{X_k}))/n\mu}{2Cov(Y, r(Y))/n\mu} \\
&= \frac{Cov(\hat{Y}_{X_k}, r(\hat{Y}_{X_k}))}{Cov(Y, r(Y))}, \qquad (3)
\end{aligned}
$$

whose sample version is

$$
MGC_{y|x_k} = \frac{Cov(\hat{y}_{x_k}, r(\hat{y}_{x_k}))}{Cov(y, r(y))} = \frac{\frac{2}{n\bar{y}}\left[\frac{1}{n}\sum_{i=1}^{n} i\hat{y}_{(x_k i)} - \frac{n(n+1)}{2n}\bar{y}\right]}{\frac{2}{n\bar{y}}\left[\frac{1}{n}\sum_{i=1}^{n} iy_{(i)} - \frac{n(n+1)}{2n}\bar{y}\right]}.
$$

Partial Gini contribution

The additional contribution related to the inclusion of covariate X_k can be determined in terms of a relative index

$$
PGC_{Y,X_k|X_1,\dots,X_{k-1}} = \frac{LZ(\hat{Y}_{X_1,\dots,X_k}) - LZ(\hat{Y}_{X_1,\dots,X_{k-1}})}{LZ(Y) - LZ(\hat{Y}_{X_1,\dots,X_{k-1}})}
$$

$$
= \frac{\frac{2}{n\mu}Cov(\hat{Y}_{X_1,\dots,X_k}, r(\hat{Y}_{X_1,\dots,X_k})) - \frac{2}{n\mu}Cov(\hat{Y}_{X_1,\dots,X_{k-1}}, r(\hat{Y}_{X_1,\dots,X_{k-1}}))}{\frac{2}{n\mu}Cov(Y, r(Y)) - \frac{2}{n\mu}Cov(\hat{Y}_{X_1,\dots,X_{k-1}}, r(\hat{Y}_{X_1,\dots,X_{k-1}}))}
$$

$$
= \frac{Cov(\hat{Y}_{X_1,\dots,X_k}, r(\hat{Y}_{X_1,\dots,X_k})) - Cov(\hat{Y}_{X_1,\dots,X_{k-1}}, r(\hat{Y}_{X_1,\dots,X_{k-1}}))}{Cov(Y, r(Y)) - Cov(\hat{Y}_{X_1,\dots,X_{k-1}}, r(\hat{Y}_{X_1,\dots,X_{k-1}}))}.
$$

It can be shown that $PGC_{Y,X_h|X_1,\dots,X_{k-1}}$ computed on sample data can be expressed as:

$$
PGC_{y,x_h|x_1,\dots,x_{k-1}} = \frac{\sum_{i=1}^{n} i(\hat{y}_{(x_1,\dots,x_k i)} - \hat{y}_{(x_1,\dots,x_{k-1} i)})}{\sum_{i=1}^{n} i(y_{(i)} - \hat{y}_{(x_1,\dots,x_{k-1} i)})}.
$$

Procedure

Steps:

- A stepwise model comparison procedure can be implemented considering the term $LZ(\hat{Y}_{X' \cup X_k}) - LZ(\hat{Y}_{X'})$ 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. The procedure continues by fitting, at each step, a more complex model that includes the predictor which provides the highest contribution.
- 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. The procedure continues by fitting, at each step, a simpler model obtained deleting the predictor with the lowest contribution.

Measuring the predictive gain

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 (3) to all such univariate models will provide a ranking of the candidate predictors, in terms of their (marginal) importance. At each step, a model with also an additional ranked variable is fitted and the predictive gain can be calculated as:

$$\textit{pay-off}(X_k) = LZ(\hat{Y}_{X' \cup X_k}) - LZ(\hat{Y}_{X'}), \tag{4}$$

where $LZ(\hat{Y}_{X' \cup X_k})$ and $LZ(\hat{Y}_{X'})$ describe the (mutual) variability of the response variable Y explained by the models which, respectively, include $X^{'} \cup X_k$ predictors or only $X^{'}$ predictors.

The procedure can continue until the predictive gain defined in (4) is found not significant.

Evaluating the significance of the predictive gain

As $r(\cdot)/n$ is the empirical transformation of the cumulative distribution function $F(\cdot)$, the pay-off can be re-expressed as:

$$LZ(\hat{Y}_{X'\cup X_k}) - LZ(\hat{Y}_{X'}) = \frac{2Cov(\hat{Y}_{X'\cup X_k}, F(\hat{Y}_{X'\cup X_k}))}{E(\hat{Y}_{X'\cup X_k})} - \frac{2Cov(\hat{Y}_{X'}, F(\hat{Y}_{X'}))}{E(\hat{Y}_{X'})}, \qquad (5)$$

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

Assuming the reasonable approximation, equation (5), which describes the marginal contribution (MC) provided by X_k, can be simplified as follows:

$$\boxed{E(\hat{Y}_{X'\cup X_k}) = E(\hat{Y}_{X'}) = \mu}$$

$$MC = \frac{2Cov(\hat{Y}_{X'\cup X_k}, F(\hat{Y}_{X'\cup X_k}))}{\mu} - \frac{2Cov(Y_{X'}, F(\hat{Y}_{X'}))}{\mu}. \qquad (6)$$

In line with the previous mathematical derivations, we propose γ as an adjusted version of equation (6), i.e.

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

The null hypothesis $H_0 : \xi(\hat{Y}_{X'\cup X_k}) = \xi(\hat{Y}_{X'})$ can be tested by the test statistic: $Z = \dfrac{\hat{\gamma}}{\sqrt{\widehat{Var(\hat{\gamma})}}} \to N(0,1)$

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

Application to default risk

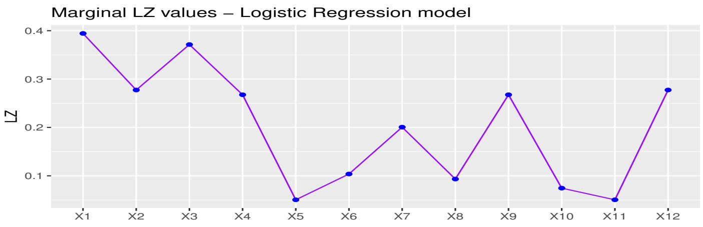
We data supplied by Modefinance, a European Credit Assessment Institution (ECAI) that specializes in credit scoring for P2P platforms focused on SME commercial lending.

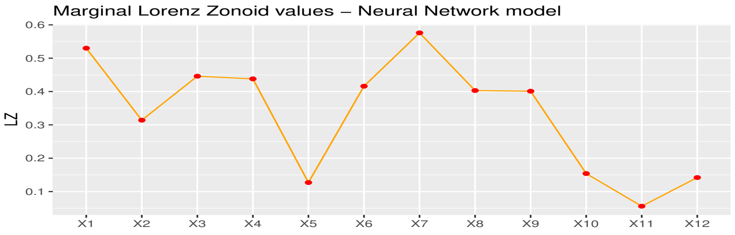
The 12 considered explanatory variables selected are: Total Assets/Total Liabilities (X_1); Current Assets/Current Liabilities (X_2); (Profit or Loss before tax+Interest paid)/Total Assets (X_3); Return on Equity (X_4); Operating Revenues/Total Assets (X_5); Interest paid/(Profit before taxes+Interest 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}).

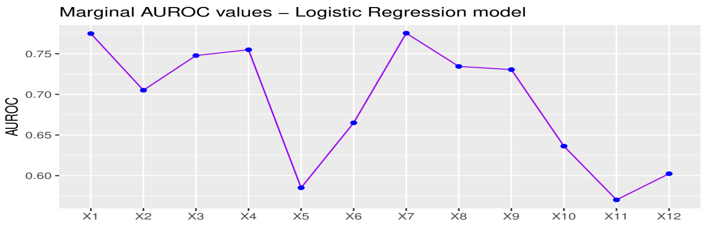
The data on the above mentioned explanatory variables are extracted from the balance-sheets of 15,045 SMEs, mostly based in Southern Europe, for the year 2015. The data on the response variable are obtained from information about the status ($0 =$ active, $1 =$ defaulted) of each SME one year later (2016),

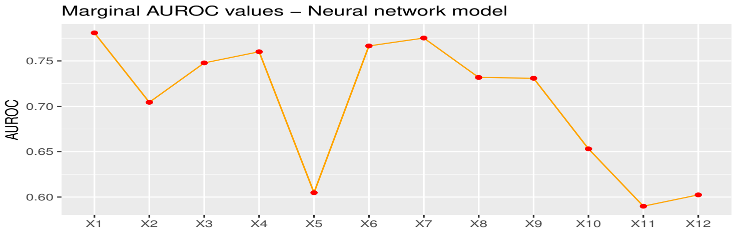
Co-financed by the Connecting Europe Facility of the European Union

Results









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

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

ID	Variable	$AUROC_{X_k}$	ID of the included variables	*pay-off* (X_k)	*p*-value
7	EBITDA/IP	0.7753	7	–	–
1	TA/TL	0.7748	7, 1	0.0016	0.9050

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

Reference

- Giudici P., Raffinetti E.: Shapley-Lorenz eXplainable Artificial Intelligence, Expert Systems With Applications, 167, 114104 (2021) and the references therein.

- Giudici P., Gramegna A., Raffinetti E.: Machine learning classification model comparison, Socio-Economic Planning Sciences (2023)

- Koshevoy G, Mosler K. The Lorenz Zonoid of a Multivariate Distribution. Journal of the American Statistical Association 91, pp 873-882 (1996)

- Shapley L.S.: A value for *n*-person games. In: Kuhn H, Tucker A, editors. Contributions to the Theory of Games II. Princeton University Press: Princeton, pp 307-317 (1953)