S.A.F.E. Artificial Intelligence

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

How to deal with high-risk applications of AI?

As mentioned in the previous sessions, complex ML models can reach high predictive performance at the expense of interpretability.

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.

The AI Act introduces a risk-based approach to AI applications, defining an AI risk taxonomy with four risk categories: unacceptable risk, high risk, limited risk, and minimal risk.

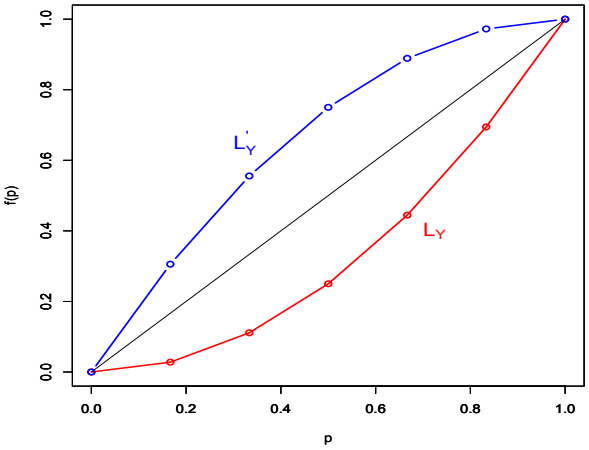
The requirements established for high-risk applications include those about sustainability, accuracy, fairness and explainability.

State of the art

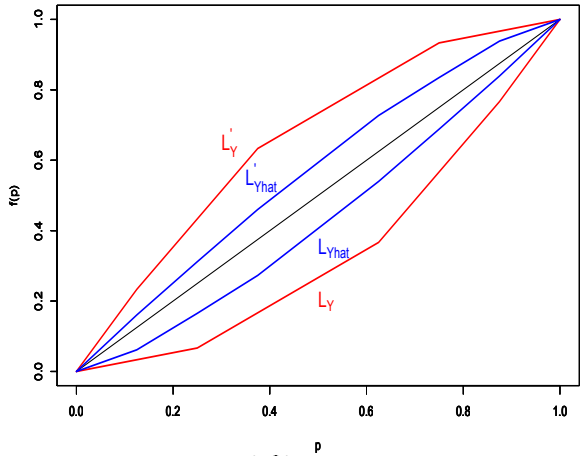
There exists no such set of integrated metrics that can establish not only whether but also how much the requirements are satisfied over time.

Recap on the Lorenz Zonoids

Inclusion Property



The Lorenz curve (L_Y) and the dual Lorenz curve (L_Y')



The inclusion property $LZ(\hat{Y}) \subset LZ(Y)$

Computation

The Lorenz Zonoid-value of a generic variable $\cdot$ (such as the response variable, or the predicted response variable) is calculated as

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

where $r(\cdot)$ are the rank-scores associated with the $\cdot$ variable and $E(\cdot)$ is its expected value.

Recap on the Shapley-Lorenz values:

- define the marginal contribution associated with the additional variable X_k as:

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

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.

Explainability Score - *Ex-Score*

Given a ML model with K predictors, we can thus measure its explainability score as in the following definition.

Definition

Explainability score. The score for explainability can be calculated on the whole sample as:

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

where $LZ(Y)$ corresponds to the response variable Y Lorenz Zonoid-value, and SL_k denotes the Shapley-Lorenz values associated with the k-th predictor.

Accuracy Score - *Ac-Score*

To measure accuracy, a more robust metric with respect to the RMSE (used for contuinuous response variable) and AUROC (used for binary the response variable) is the Lorenz Zonoid, which can be calculated on the test set in the same way for binary, ordered categorical and continuous responses.

Given a ML model with $k \leq K$ predictors, and a test sample from the dataset, we can measure its accuracy score as in the following definition.

Definition

The score for accuracy can be defined as:

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

where $LZ(\hat{Y}_{X_1,\ldots,X_k})$ is the Lorenz Zonoid of the predicted response variable, obtained using k predictors on the test set, and $LZ(Y_{test})$ is the Y response variable Lorenz Zonoid value computed on the same test set.

Fairness Score I - *Fair-Score*

Fairness is a property that essentially requires that AI applications do not present biases among different population groups.

To measure fairness we propose to extend the Gini coefficient in terms of the Shapley-Lorenz values.

Let:

- $m = 1, \ldots, M$ be the considered population groups;
- K be the number of the available predictors;
- $v^{SL}_{mX_k}$ be the Shapley-Lorenz value associated with the k-th predictor in the m-th population;
- k^*, where $k^* = 1, \ldots, k$ and such that $k^* < K$, be the number of predictors which compose the model selected by the stepwise procedure based on the application of the Lorenz-Zonoid.

Fairness Score II - *Fair-Score*

With the purpose of measuring the explainability and accuracy provided by each explanatory variable included into the final model, we consider the vector V_M^{SL*} defined as

$$V_M^{SL*} = \{v_1^{SL*}, \ldots, v_m^{SL*}, \ldots, v_M^{SL*}\},$$

where $v_m^{SL*} = v_{mX_1}^{SL} + \ldots + v_{mX_{k^*}}^{SL}$ represents the sum of the Shapley-Lorenz values related to the predictors $X_1, \ldots, X_{k^*}$.

The Gini coefficient can be applied to the vector V_M^{SL*}, obtaining a measure of concentration of the variables' importance among different population groups.

Remark

For a given set of selected explanatory variables,

- Shapley-Lorenz values which are similar in the M populations lead to a Gini coefficient close to 0;
- a Gini coefficient close to 1 indicates that the variables' effect largely depend on some groups, highlighting biasness.

Fairness Score III - *Fair-Score*

Given a ML model with k^* and M population groups, we can measure its fairness score as in the following definition.

Definition

The score for fairness can be defined as:

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

where $LZ(V_M^{SL*})$ denotes the Lorenz Zonoid (Gini coefficient) computed on the vector V_M^{SL*} whose elements correspond to the sum of the selected predictors' Shapley-Lorenz values in each population.

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

Sustainability Score I - *Sust-Score*

The results from a ML model may be altered by the presence of "extreme" data points, deriving from anomalous events.

We propose to verify sustainability by comparing predictive accuracy, as measured by Shapley-Lorenz values, in different ordered subset of the data, possibly altered artificially by anomalous ones.

We proceed by:

- ordering the predicted response values (in the test set) in terms of their predictive accuracy, from the most accurate to the lowest;
- dividing the ordered predictions in $g = 1, \ldots, G$ equal size groups (such as the deciles of the distribution);
- build a vector including the sum of the Shapley-Lorenz values of the predictors composing the final model, i.e. $V_G^{SL*} = \{v_1^{SL*}, \ldots, v_g^{SL*}, \ldots, v_G^{SL*}\}$, where $v_g^{SL*} = v_{gX_1}^{SL} + \ldots + v_{gX_{k^*}}^{SL}$ represents the sum of the Shapley-Lorenz values related to the predictors $X_1, \ldots, X_{k^*}$.

Sustainability Score II - *Sust-Score*

Definition

The score for sustainability can then be defined as:

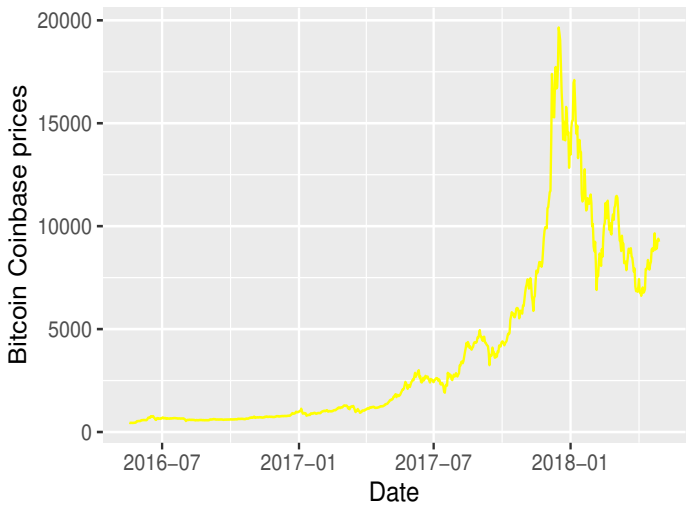
$$Sust\text{-}Score = 1 - LZ(V_G^{SL*}),$$

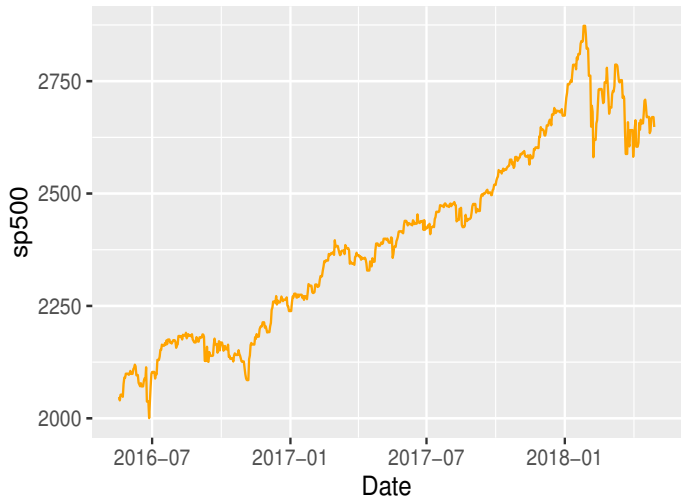
where $LZ(V_G^{SL*})$ indicates the Lorenz Zonoid (Gini coefficient) calculated on the vector V_G^{SL*} whose elements correspond to the sum of the selected predictors' Shapley-Lorenz values in each group.

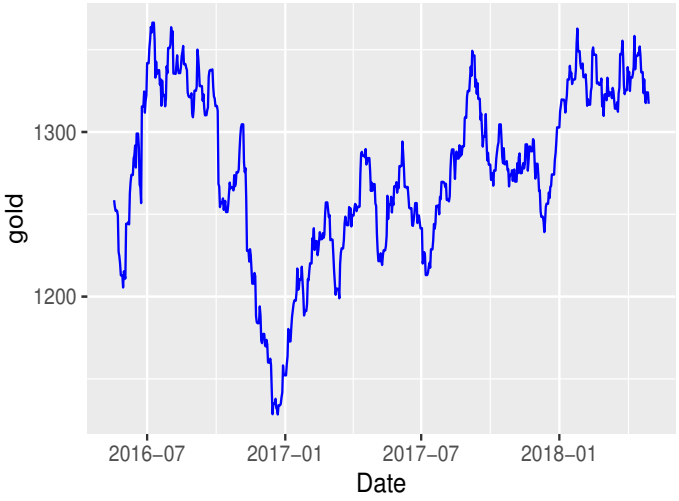
Application to Bitcoin price prediction

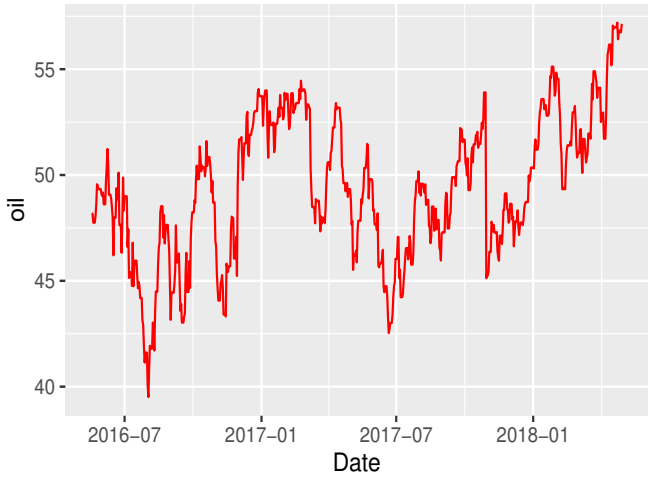
- As data we consider the bitcoin price in the Coinbase exchange, from 18 May 2016 to 30 April 2018, as a target variable, both as continuous and as binarised; the prices of sp500, gold and oil, and the exchange rates USD/EUR and USD/YUAN as predictors.
- As a predictive model we consider a neural network model with 5 hidden neurons.
- As training data we consider the time series until December 31st, 2017, while as test data the 2018 time series.
- We deal with both a continuous target variable and a binarised target variable (assigning value equal to 0 to the negative bitcoin *returns* and value equal to 1 otherwise).
- For both continuous and binary cases, we calculate the S.A.F.E. scores of the model, using Lorenz Zonoid values, Shapley-Lorenz values and on the corresponding Gini coefficient across population percentiles and groups.

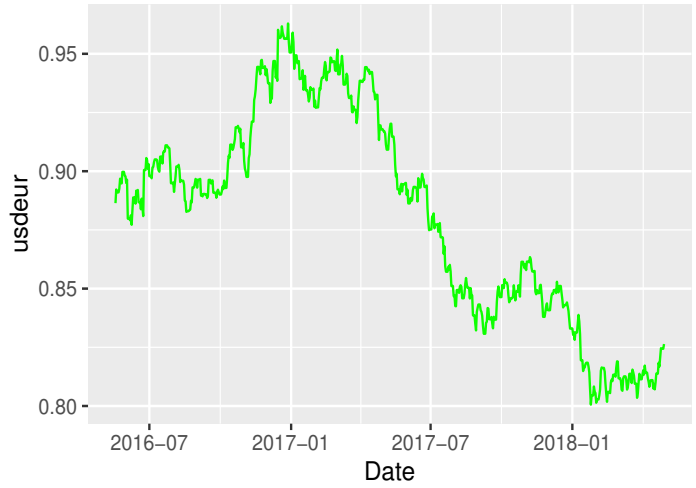
Explorative Analysis

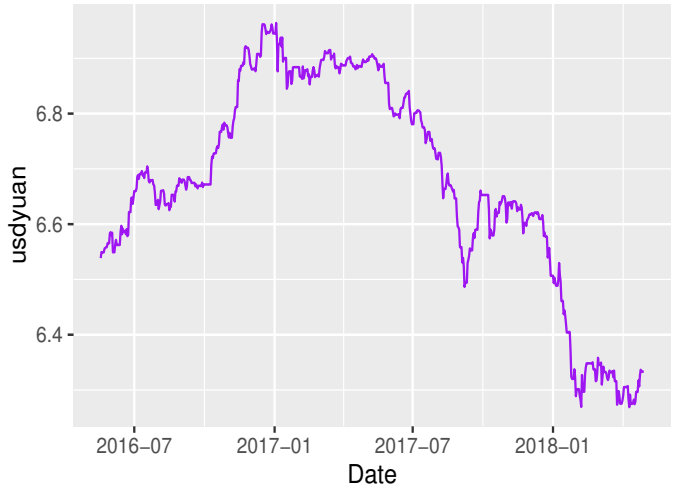




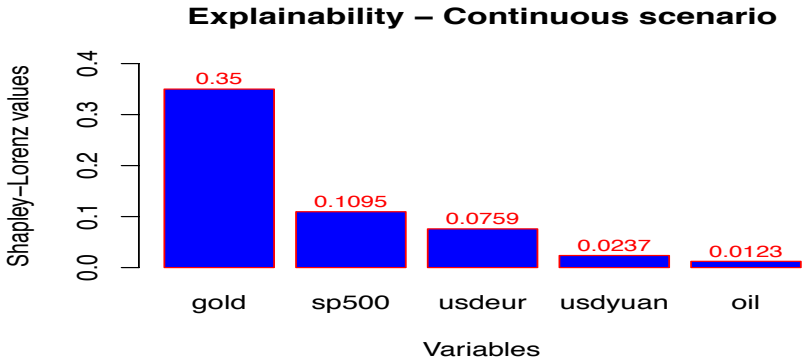


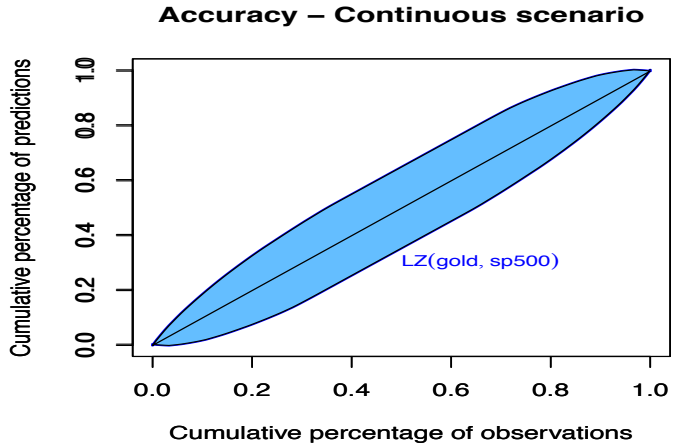






Results - continuous target variable (1)





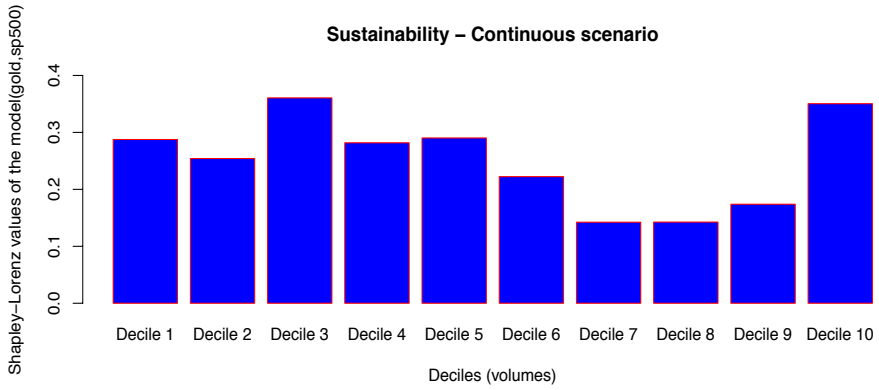
$$\text{Ex-Score} = \frac{\sum_{k=1}^{5} SL_k}{LZ(returns)} = 0.5714$$

$$\text{Ac-Score} = \frac{LZ(\hat{returns}_{gold,sp500})}{LZ(returns_{test})} = 0.3279$$

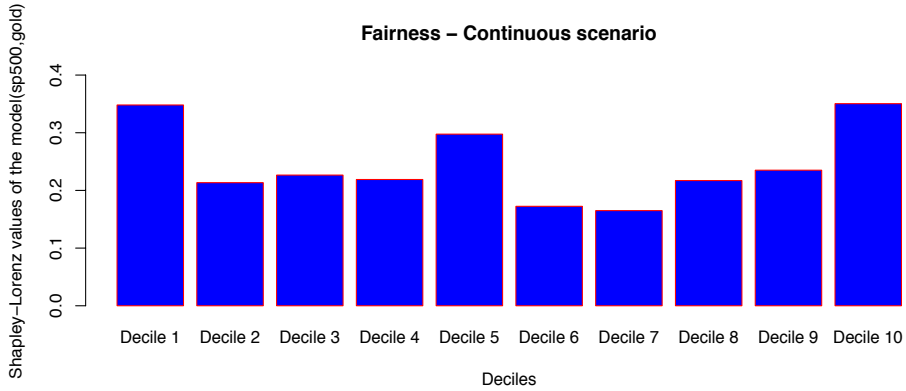
Co-financed by the Connecting Europe Facility of the European Union

eXplainable Artificial Intelligence in healthcare Management
2020-EU-IA-0098

Results - continuous target variable (2)



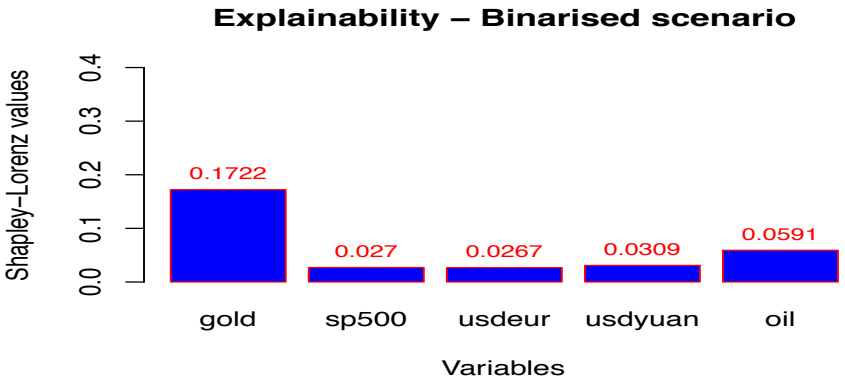
To assess sustainability, we have ordered the test data response according to how well is predicted by the model (from the best to the worst predictions) and, accordingly, subdivided it into ten deciles.
We have then calculated the Lorenz Zonoid of the model, separately in each decile.

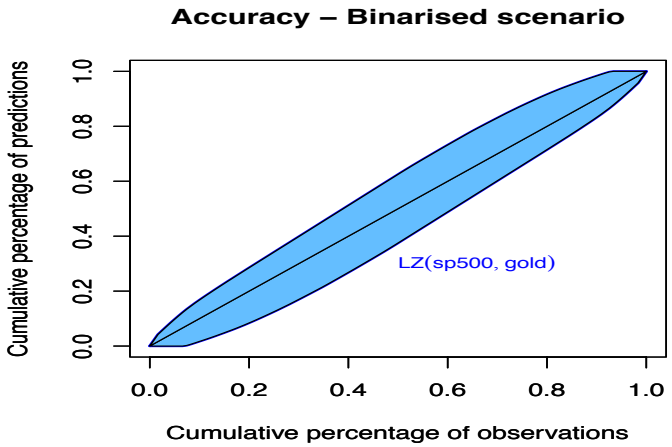


To measure fairness we have ordered the test data response in terms of the corresponding trading volumes (from the lowest to the highest) and, accordingly, subdivided it into ten deciles.
We have then calculated the Lorenz Zonoid of the model, separately in each decile.

$$\text{Sust-Score} = 1 - Gini(SL_{gold+sp500}) = 0.8314$$
$$\text{Fair-Score} = 1 - Gini(SL_{gold+sp500}) = 0.8617$$

Results - binarised target variable (1)



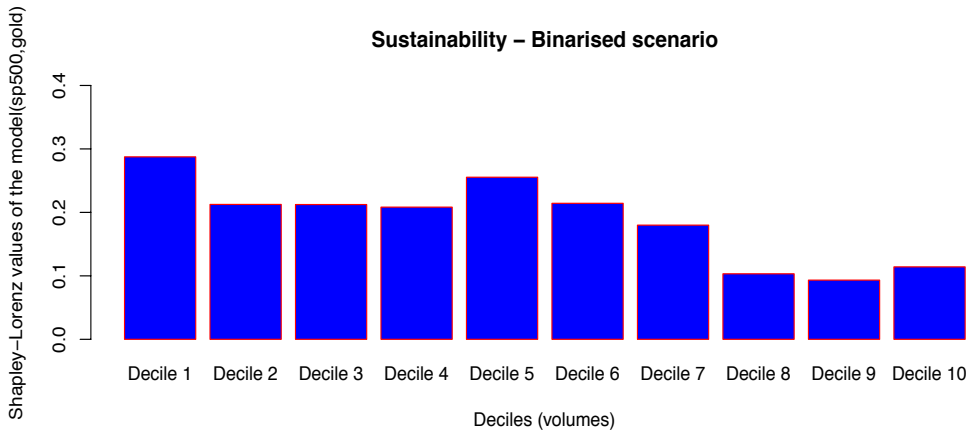


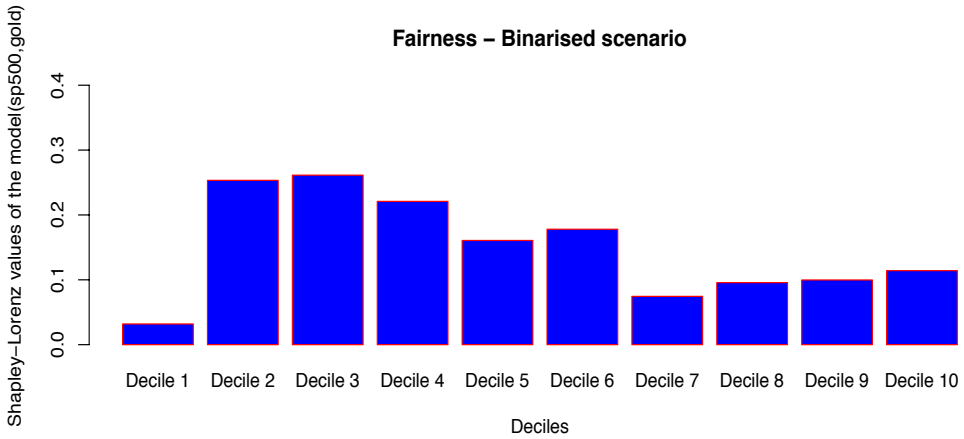
$$\text{Ex-Score} = \frac{\sum_{k=1}^{5} SL_k}{LZ(b.returns)} = 0.3160$$

$$\text{Ac-Score} = \frac{LZ(\hat{b.returns}_{sp500,gold})}{LZ(b.returns_{test})} = 0.4088$$

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

Results - binarised target variable (2)





$$\text{Sust-Score} = 1 - Gini(SL_{sp500+gold}) = 0.8184$$
$$\text{Fair-Score} = 1 - Gini(SL_{sp500+gold}) = 0.7165$$

Reference

- Giudici P, Raffinetti E.: SAFE Artificial Intelligence in Finance (February 17, 2023). Available at SSRN: https://ssrn.com/abstract=4362034 or http://dx.doi.org/10.2139/ssrn.4362034