Shapley Value Decomposition of Evaluation Metrics for Regression and Classification Models*

Sullivan Hué† Christophe Hurlin‡ Christophe Pérignon§ Sébastien Saurin¶

March 7, 2022

Abstract

Although extremely popular in the industry due to their high predictive performance, the lack of interpretability of machine learning algorithms raises concerns from practitioners and regulators. We propose an original model-agnostic method aiming to unravel the opacity regarding black boxes' decision process. Specifically, the method mesures the contribution of input features to the predictive performance of a model through the decomposition of an evaluation metric. The approach can be applied to any type of model, econometrics or machine learning, and to a wide class of evaluation metric including the most famous measures such as the R^2 for regression problem or the Area Under the ROC Curve (AUC) for classification problems. We show the properties of our decomposition method and illustrate how feature contributions are estimated, even in a high-dimensional model context. A framework for local analysis is also developed. We highlight the usefulness of the approach using real data of credit scoring applications. Through the decomposition of several evaluation metrics, we illustrate how the decomposition can be used in practice to improve the decision making associated to black boxes.

Keywords: Shapley; Credit markets; Machine Learning; Artificial intelligence

*We thank the ACPR Chair in Regulation and Systemic Risk, the Fintech Chair at Dauphine-PSL University, and the French National Research Agency (Ecodec ANR-11-LABX-0047, F-STAR ANR-17-CE26-0007-01) for supporting our research.

†Aix-Marseille University (Aix-Marseille School of Economics), CNRS & EHESS. Email: sullivan.hue@univ-amu.fr

‡University of Orléans, LEO, Rue de Blois, 45067 Orléans, France. Email: christophe.hurlin@univ-orleans.fr

§HEC Paris, 1 Rue de la Libération, 78350 Jouy-en-Josas, France. Email: perignon@hec.fr

¶University of Orléans, LEO, Rue de Blois, 45067 Orléans, France. Email: sebastien.saurin@univ-orleans.fr

1 Introduction

2 Framework and Evaluation Metrics

We consider a classification or a regression problem involving a target variable denoted y which takes values in $\mathcal{Y}$, which is either defined as $\mathcal{Y} = \{0, 1\}$, in case of classification, or as $\mathcal{Y} \subset \mathbb{R}$, in case of regression. The q-vector $\mathbf{x} \in \mathcal{X}$ refers to input (explanatory) features with $\mathcal{X} \subset \mathbb{R}^q$. Here, we consider continuous features for ease of comprehension but the same framework could be applied to discrete features. We denote by $f : \mathbf{x} \to \hat{y}$ an econometric model or a machine learning algorithm, where $\hat{y} \in \mathcal{Y}$ is either a classification output, or regression output, such as $\hat{y} = f(\mathbf{x})$. In case of classification problem, we assume that the classifier also produces conditional probabilities denoted $P(\mathbf{x}) = \mathbb{P}(\hat{y} = 1|\mathbf{x})$. We impose no constraint on the model form $f(.)$. For instance, the model may be parametric or not, linear or not, individual or ensemble classifier, etc. In case of a parametric model, we exclude the parameters from the notation, $f(\mathbf{x}) \equiv f(\mathbf{x}; \theta)$.

The model is estimated (for a parametric model) or trained (for a machine learning algorithm) **once for all** on a training (estimation) sample $\{\mathbf{x}_j, y_j\}_{j=1}^T$. The training sample size T is considered as fixed and we impose no constraint on it.[1] The estimated model can be written interchangeably as $\hat{f}(.)$ or $\hat{f}_T(.)$. The statistical performance of the model is evaluated on a test sample S_n of n individuals, indexed by $i = 1, ..., n$, for which we observe $(\mathbf{x}_i, y_i, \hat{f}(\mathbf{x}_i))$, such as $S_n = \{\mathbf{x}_i, y_i, \hat{f}(\mathbf{x}_i)\}_{i=1}^n$.

We define an **evaluation metric** (EM) as an assessment measure of the statistical model performance. For instance, Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared are typical evaluation metrics for regression models, whereas Area Under the Curve (AUC), Brier Score (BS), and Gini index are standard evaluation metrics for classification models. We can also consider alternative metrics such as information criteria (AIC, BIC, etc.) or any loss function (Qlike,etc.).

Definition 1. *A sample evaluation metric $EM_n \in \Theta \subseteq \mathbb{R}$ associated to the model $\hat{f}(.)$ and a test sample S_n is a scalar defined as*[2]*:*

$$\begin{aligned} EM_n &= \tilde{G}_n(y_1, ..., y_n; \hat{y}_1, ..., \hat{y}_n) = \tilde{G}_n(y_1, ..., y_n; \hat{f}(\mathbf{x_1}), ..., \hat{f}(\mathbf{x_n})) \\ &= G_n(y_1, ..., y_n; \mathbf{x_1}, ..., \mathbf{x_n}) = G_n(\mathbf{y}; \mathbf{X}), \end{aligned} \tag{1}$$

where $\mathbf{y} = (y_1, ..., y_n)^T$ and $\mathbf{X} = (\mathbf{x}_1, .., \mathbf{x}_n)^T$.

Assumption 1. *The sample evaluation metrics increases with statistical performance of the model.*

For instance, the R-squared and the AUC satisfy this assumption, whereas we have to consider the opposite of the MSE and the Brier Score as sample evaluation metric.

Assumption 2. *(i) The sample evaluation metric $G_n(\mathbf{y}; \mathbf{X})$ converges to the population evaluation metric $\mathbb{E}_{y,\mathbf{x}}(G(y; \mathbf{x}))$, where $\mathbb{E}_{y,\mathbf{x}}(.)$ refers to the expected value with respect to the joint distribution of y and $\mathbf{x}$. (ii) $\mathbb{E}_{y,\mathbf{x}}(G(y; \mathbf{x}))$ exists and is finite.*

[1] Interpretable machine learning literature (Molnar, 2020) aims to identify feature contributions to model outcomes where the model is estimated (trained) on a given set of information. Here, we adopt the same approach and consider a fixed training sample size. Thus, we do not assess the impact of the training sample (in-sample) size on the model performance analysis.

[2] For instance, $\Theta = [0, 1]$ for AUC, R-squared, Brier Score, and $\Theta = \mathbb{R}^+$ for MSE, MAE, etc.

To illustrate our framework, let us consider typical applications.

Example 1. In case of a linear regression model $\hat{f}(.)$ with i.i.d assumptions, if we consider the (opposite) MSE as evaluation metric then:

$$G_n(\mathbf{y};\mathbf{X}) = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(\mathbf{x}_i))^2,$$
$$\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x})) = -\mathbb{E}_{y,\mathbf{x}}\left[\left(y - \hat{f}(\mathbf{x})\right)^2\right].$$

Example 2. In case of logistic regression model $\hat{f}(.)$ with i.i.d features, if we consider the accuracy as evaluation metric then:

$$G_n(\mathbf{y};\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}(\mathbf{x}_i)y_i + (1-\hat{f}(\mathbf{x}_i))(1-y_i)\right),$$
$$\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x})) = \mathbb{E}_{y,\mathbf{x}}\left(\hat{f}(\mathbf{x})y + (1-\hat{f}(\mathbf{x}))(1-y)\right).$$

3 Shapley value decomposition of evaluation metric

Our objective is to identify the contribution of the features to the evaluation metric. To do so, we use a standard Shapley value decomposition (Shapley, 1953). The Shapley values, a concept issued from game theory, fairly distribute the evaluation metric (payoff) among the model features (players). It corresponds to the average marginal impact of a given feature on the evaluation metric evaluated while controlling for the effect of combinations of other features (coalitions). This interpretability method is theorically founded and has interesting properties (dummy, symmetry, etc.).

Property 1. *(Efficiency) The sum of the Shapley values ϕ_j, $\forall j = 1, ..., q$, is equal to the difference between the population evaluation metric $\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x}))$ and its benchmark ϕ_0 such as:*

$$\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x})) = \phi_0 + \sum_{j=1}^{q}\phi_j, \tag{2}$$

where $\phi_0 = \mathbb{E}_{\mathbf{x}}\mathbb{E}_y\left(G(y;\mathbf{x})\right)$ corresponds to the evaluation metric associated to a population where the target variable is independent from all features considered in the model.

In other words, ϕ_0 is the evaluation metric obtained for a model without any predictive ability. Thereafter, we refer to ϕ_0 as the benchmark of the evaluation metric.

If a feature x_j enhances the predictive ability of the model compared to the benchmark ϕ_0, then its contribution ϕ_j to the evaluation metric is positive. This contribution is formally defined as a Shapley value. Let denote $\mathbf{x}^S$ as the vector of features included in coalition S and $\mathbf{x}^{\overline{S}}$ the vector of features excluded from coalition S, such as $\{\mathbf{x}\} = \{\mathbf{x}^S\} \cup \{\mathbf{x}^{\overline{S}}\} \cup \{x_j\}$.

Definition 2. *The contribution of feature x_j to the evaluation metric is:*

$$\phi_j = \sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\}\setminus\{x_j\})} w_S\left[\mathbb{E}_{\mathbf{x}^S}\mathbb{E}_{y,x_j,\mathbf{x}^{\overline{S}}}\left(G\left(y;\mathbf{x}\right)\right) - \mathbb{E}_{x_j,\mathbf{x}^S}\mathbb{E}_{y,\mathbf{x}^{\overline{S}}}\left(G\left(y;\mathbf{x}\right)\right)\right], \tag{3}$$

$$w_S = \frac{|S|!\,(q-|S|-1)!}{q!}, \tag{4}$$

with S a coalition, i.e., a subset of features, excluding the feature of interest x_j, $|S|$ the number of features in the coalition, and $\mathcal{P}(\{\mathbf{x}\} \setminus \{x_j\})$ the partition of the set $\{\mathbf{x}\} \setminus \{x_j\}$.

The Shapley value ϕ_j associated to the feature x_j measures its weighted average marginal contribution to the evaluation metric over all feature coalitions. This marginal contribution is defined as the difference betweeen expected values of the evaluation metric obtained while considering or not the dependence between the target variable y and x_j. Formally, in equation (12), $\mathbb{E}_{\mathbf{x}^S}\mathbb{E}_{y,x_j,\mathbf{x}^{\overline{S}}}\left(G\left(y;\mathbf{x}\right)\right)$ refers to the expected value of the evaluation metric considering that the coalitions variables $\mathbf{x}^S$ are independent from $(y, x_j, \mathbf{x}^{\overline{S}})$. Similarly, $\mathbb{E}_{x_j,\mathbf{x}^S}\mathbb{E}_{y,\mathbf{x}^{\overline{S}}}\left(G\left(y;\mathbf{x}\right)\right)$ is the expected value of the evaluation metric when the features $(\mathbf{x}^S, x_j)$ are independent from $(y, \mathbf{x}^{\overline{S}})$. To illustrate the Shapley value computation, we consider a model with three features and x_1 the feature of interest. In Table 1, we report all the coalitions among the set $\{x_2, x_3\}$ (column 1), the associated weights according to equation (4) (column 2) and the marginal contributions (column 3) used to compute the Shapley value ϕ_1.

Table 1: Computation of the Shapley value ϕ_1 in a three-feature model.

S	w_S	$\mathbb{E}_{\mathbf{x}^S}\mathbb{E}_{y,x_1,\mathbf{x}^{\overline{S}}}\left(G\left(y;\mathbf{x}\right)\right) - \mathbb{E}_{x_1,\mathbf{x}^S}\mathbb{E}_{y,\mathbf{x}^{\overline{S}}}\left(G\left(y;\mathbf{x}\right)\right)$
$\{\emptyset\}$	$1/3$	$\mathbb{E}_{y,x_1,x_2,x_3}\left(G\left(y;\mathbf{x}\right)\right) - \mathbb{E}_{x_1}\mathbb{E}_{y,x_2,x_3}\left(G\left(y;\mathbf{x}\right)\right)$
$\{x_2\}$	$1/6$	$\mathbb{E}_{x_2}\mathbb{E}_{y,x_1,x_3}\left(G\left(y;\mathbf{x}\right)\right) - \mathbb{E}_{x_1,x_2}\mathbb{E}_{y,x_3}\left(G\left(y;\mathbf{x}\right)\right)$
$\{x_3\}$	$1/6$	$\mathbb{E}_{x_3}\mathbb{E}_{y,x_1,x_2}\left(G\left(y;\mathbf{x}\right)\right) - \mathbb{E}_{x_1,x_3}\mathbb{E}_{y,x_2}\left(G\left(y;\mathbf{x}\right)\right)$
$\{x_2, x_3\}$	$1/3$	$\mathbb{E}_{x_2,x_3}\mathbb{E}_{y,x_1}\left(G\left(y;\mathbf{x}\right)\right) - \mathbb{E}_{x_1,x_2,x_3}\mathbb{E}_{y}\left(G\left(y;\mathbf{x}\right)\right)$

Note: This table displays details of Shapley value computation, i.e., the coalitions (column 1), the associated weights according to equation (4) (column 2) and the marginal contributions (column 3).

The Shapley value ϕ_1 is computed from Table 1 by multiplying the weights (column 2) to the marginal contributions (column 3) and summing over all coalitions, such as:

$$\begin{aligned}
\phi_1 &= \frac{1}{3}\left(\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x})) - \mathbb{E}_{x_1}\mathbb{E}_{y,x_2,x_3}\left(G\left(y;\mathbf{x}\right)\right)\right) \\
&+ \frac{1}{6}\left(\mathbb{E}_{x_2}\mathbb{E}_{y,x_1,x_3}\left(G\left(y;\mathbf{x}\right)\right) - \mathbb{E}_{x_1,x_2}\mathbb{E}_{y,x_3}\left(G\left(y;\mathbf{x}\right)\right)\right) \\
&+ \frac{1}{6}\left(\mathbb{E}_{x_3}\mathbb{E}_{y,x_1,x_2}\left(G\left(y;\mathbf{x}\right)\right) - \mathbb{E}_{x_1,x_3}\mathbb{E}_{y,x_2}\left(G\left(y;\mathbf{x}\right)\right)\right) \\
&+ \frac{1}{3}\left(\mathbb{E}_{x_2,x_3}\mathbb{E}_{y,x_1}\left(G\left(y;\mathbf{x}\right)\right) - \phi_0\right).
\end{aligned}$$

Using the same approach we can compute ϕ_2, ϕ_3 and then verified the efficiency property as follows:

$$\sum_{j=1}^{3}\phi_j = \underbrace{\sum_{j=1}^{3}\frac{1}{3}\left(\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x})) - \phi_0\right)}_{\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x}))-\phi_0} + \underbrace{\sum_{j=1}^{3}M_j(G\left(y;\mathbf{x}\right))}_{=\,0} \tag{5}$$

where $M_j(G\left(y;\mathbf{x}\right))$ includes expected values of the evaluation metric (see Appendix A).

The Shapley value satisfies a set of axioms (efficiency, dummy, symmetry, monotonicity) particularly relevant for statistical performance analysis.

Property 2. *(Dummy) If the model feature x_j does not have any impact on the evaluation metric $\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x}))$, then its Shapley value ϕ_j is null, i.e., $\phi_j = 0$.*

Property 3. *(Monotonicity) If a feature x_j contributes more to the evaluation metric $\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x}))$ than a feature x_s then $\phi_j > \phi_s$.*

Property 4. *(Symmetry) If two features x_j and x_s contribute equally to the evaluation metric $\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x}))$ across all coalitions then $\phi_j = \phi_s$.*

Illustration (MSE and regression model). Consider a linear regression model $\hat{f}(\mathbf{x}_i) = \sum_{j=1}^{q} \beta_j x_{i,j}$, and the MSE (opposite) as sample evaluation metric. We consider model parameters β_j as fixed.[3] As concerned the Data Generating Process (DGP) for the validation sample $S_n = \{\mathbf{x}_i, y_i, \hat{f}(\mathbf{x}_i)\}_{i=1}^{n}$, we assume that the features satisfy $\mathbb{E}(\mathbf{x}) = 0_q$ and $\mathbb{V}(\mathbf{x}) = diag(\sigma_{x_j}^2)\ \forall j = 1, ..., q$. We denote by σ_y^2 the variance of the target variable and by σ_{y,x_j} the covariance between the feature x_j and the target variable. Then, the contributions ϕ_j of features x_j, $\forall j = 1, ..., n$, to the MSE (opposite) satisfy the efficiency property such that:

$$\underbrace{2\sum_{j=1}^{q}\beta_j\sigma_{y,x_j} - \sum_{j=1}^{q}\beta_j^2\sigma_{x_j}^2 - \sigma_y^2}_{\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x}))} = \underbrace{-\sum_{j=1}^{q}\beta_j^2\sigma_{x_j}^2 - \sigma_y^2}_{\phi_0} + \sum_{j=1}^{q}\underbrace{2\beta_j\sigma_{y,x_j}}_{\phi_j} \tag{6}$$

Formally, the Shapley value ϕ_j depends on the model parameter β_j (estimation sample) and the covariance between x_j and the target variable y (validation sample), i.e., σ_{y,x_j}. The Shapley values ϕ_j, $\forall j = 1, ..., n$, are positive or null. [4] The dummy property $\phi_j = 0$ is either satisfied if the feature has no impact on the model ($\beta_j = 0$) or if the feature is uncorrelated with the target variable on the validation sample ($\sigma_{y,x_j} = 0$). Similarly, a variable x_j has a larger MSE contribution than a feature x_s as soon as $\beta_j\sigma_{y,x_j} > \beta_s\sigma_{y,x_s}$, meaning that x_j is more related to the target variable than x_s both in-sample (through β_j) and out-sample (through σ_{y,x_s}). The benchmark ϕ_0 corresponds to the MSE that we would obtain by applying the model to data generated by a DGP where the target variable is independent from the features.

Illustration (Accuracy and classification model). Consider a logistic regression model and the accuracy as sample evaluation metric with

$$\hat{f}(\mathbf{x}_i) = \begin{cases} 1 & \text{if } \hat{P}(\mathbf{x}_i) = 1/\left[1 + \exp\left(-\mathbf{x}_i\beta\right)\right] > \pi \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where $\pi \in [0, 1]$ refers to a cutoff value. We denote by $\sigma_{y,\hat{f}(\mathbf{x})}$ the covariance between the target

[3] In a traditional in-sample/out-of-sample estimation framework, β_j refers to the parameter estimates obtained on the estimation sample.

[4] If the DGPs of the estimation and validation sample are similar, we expect the model parameters β_j and covariances σ_{y,x_j}, $\forall j = 1, ..., n$ to have the same sign. For instance, if the covariance between the target variable and a feature is negative in-sample and out-of-sample, i.e., $\beta_j < 0$ and $\sigma_{y,x_j} < 0$, then $\phi_j > 0$.

variable and the classification output. Then, the efficiency property becomes:

$$\underbrace{2\sigma_{y,\hat{f}(\mathbf{x})} + 2P(\mathbf{x})\hat{P}(\mathbf{x}) + 1 - P(\mathbf{x}) - \hat{P}(\mathbf{x})}_{\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x}))} = \underbrace{2P(\mathbf{x})\hat{P}(\mathbf{x}) + 1 - P(\mathbf{x}) - \hat{P}(\mathbf{x})}_{\phi_0} + \underbrace{2\sigma_{y,\hat{f}(\mathbf{x})}}_{\sum_{j=1}^{q}\phi_j} \tag{8}$$

with $P(\mathbf{x}) = \mathbb{P}(y = 1|\mathbf{x})$. Unlike the example of the MSE, the Shapley values ϕ_j, $\forall j = 1, ..., n$ do not have any analytical expressions. As expected, the Shapley values depend on the covariance between the target variable y and the classification output $\hat{f}(\mathbf{x})$. The benchmark ϕ_0 corresponds to the accuracy that we would obtain by applying the model to data generated by a DGP where the target variable is independent from the features.

4 Shapley value estimation

In this section we develop the estimation of Shapley values ϕ_j, $j = 1, .., q$ and identify among them individual contributions $\phi_{i,j}$, $i = 1, .., n$, $j = 1, .., q$.

4.1 Global Analysis

The estimation of feature contributions to the evaluation metric requires to make the following assumption.

Assumption 3. *The sample evaluation metric satisfies an* ***additive property*** *such that:*

$$G_n(\mathbf{y}; \mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} G(y_i; \mathbf{x}_i; \delta_n), \tag{9}$$

where $G(y_i; \mathbf{x}_i; \delta_n)$ denotes an individual contribution to the evaluation metric and $\delta_n(y_i, \mathbf{x}_i)$ is a normalization factor which depends on the test sample S_n.

For ease of presentation, we only consider models for which the outcome y_i for cross-sectional unit i only depends on its own features $\mathbf{x}_i$. For regression or classification models with cross-sectional interactions (e.g., spatial econometrics model) or time series dependence, notations have to be adjusted such that $\hat{y}_i = \widehat{f}(\mathbf{w}_i)$ where $\mathbf{x}_i \subseteq \mathbf{w}_i$, $\exists j \neq i : \mathbf{x}_j \subseteq \mathbf{w}_i$ and/or $y_j \subseteq \mathbf{w}_i$. Then, the additive property becomes $G_n(\mathbf{y}; \mathbf{X}) = n^{-1}\sum_{i=1}^{n} G(y_i; \mathbf{w}_i; \delta_n)$.

The additive property is satisfied for many pairs of models and evaluation metrics as illustrated by the following examples.

Example 3. Consider a linear regression model $\hat{f}(.)$ with i.i.d assumptions and the (opposite) MSE as evaluation metric then:

$$G_n(\mathbf{y}; \mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} G(y_i; \mathbf{x}_i; \delta_n) = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(\mathbf{x}_i))^2,$$

with $G(y_i; \mathbf{x}_i; \delta_n) \equiv G(y_i; \mathbf{x}_i) = -(y_i - \hat{f}(\mathbf{x}_i))^2$.

Example 4. Consider a logistic regression model $\hat{f}(.)$ with i.i.d features and the accuracy as evaluation metric then:

$$G_n(\mathbf{y};\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} G(y_i;\mathbf{x}_i;\delta_n) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}(\mathbf{x}_i)y_i + (1-\hat{f}(\mathbf{x}_i))(1-y_i)\right),$$

with $G(y_i;\mathbf{x}_i;\delta_n) \equiv G(y_i;\mathbf{x}_i) = \hat{f}(\mathbf{x}_i)y_i + (1-\hat{f}(\mathbf{x}_i))(1-y_i)$.

Example 5. Consider a linear regression model $\hat{f}(.)$ with i.i.d assumptions and the R^2 as evaluation metric then:

$$G_n(\mathbf{y};\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} G(y_i;\mathbf{x}_i;\delta_n) = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{f}(\mathbf{x}_i))^2}{\frac{1}{n}\sum_{j=1}^{n}(y_j-\bar{y})^2},$$

with $G(y_i;\mathbf{x}_i;\delta_n) = 1-\delta_n^{-1}(y_i-\hat{f}(\mathbf{x}_i))^2$ and $\delta_n = \frac{1}{n}\sum_{j=1}^{n}(y_j-\bar{y})^2$.

Under regular assumptions and the additive property (assumption 9), the Shapley value ϕ_j can be estimated as a weighted average of individual contributions differences.

Definition 3. *A consistent estimator of the contribution of feature x_j to evaluation metric is:*

$$\hat{\phi}_j = \sum_{S\subseteq\mathcal{P}(\{\mathbf{x}\}\setminus\{x_j\})} w_S\left[\frac{1}{n^2}\sum_{u=1}^{n}\sum_{v=1}^{n} G\left(y_v;x_{v,j},\mathbf{x}_u^S,\mathbf{x}_v^{\overline{S}};\delta_n\right) - \frac{1}{n^2}\sum_{u=1}^{n}\sum_{v=1}^{n} G\left(y_v;x_{u,j},\mathbf{x}_u^S,\mathbf{x}_v^{\overline{S}};\delta_n\right)\right], \tag{10}$$

$$w_S = \frac{|S|!\,(q-|S|-1)!}{q!}, \tag{11}$$

with S a coalition, i.e., a subset of features, excluding the feature of interest x_j, $|S|$ the number of features in the coalition, and $\mathcal{P}(\{\mathbf{x}\}\setminus\{x_j\})$ the partition of the set $\{\mathbf{x}\}\setminus\{x_j\}$.

Similarly to the examples reported in Table 1 we illustrate the computation of the estimated Shapley value $\hat{\phi}_1$ in a model with three features. Table 2 reports the estimated marginal contributions (column 3) of feature x_1 and the corresponding weights (column 2) for all coalitions of other features (column 1).

Table 2: Computation of the Shapley value $\hat{\phi}_1$ in a three-feature model.

S	w_S	$\frac{1}{n^2}\sum_{u=1}^{n}\sum_{v=1}^{n} G\left(y_v;x_{v,j},\mathbf{x}_u^S,\mathbf{x}_v^{\overline{S}};\delta_n\right) - \frac{1}{n^2}\sum_{u=1}^{n}\sum_{v=1}^{n} G\left(y_v;x_{u,j},\mathbf{x}_u^S,\mathbf{x}_v^{\overline{S}};\delta_n\right)$
$\{\emptyset\}$	1/3	$\frac{1}{n}\sum_{v=1}^{n} G\left(y_v;x_{v,1},x_{v,2},x_{v,3};\delta_n\right) - \frac{1}{n^2}\sum_{u=1}^{n}\sum_{v=1}^{n} G\left(y_v;x_{u,1},x_{v,2},x_{v,3};\delta_n\right)$
$\{x_2\}$	1/6	$\frac{1}{n^2}\sum_{u=1}^{n}\sum_{v=1}^{n} G\left(y_v;x_{v,1},x_{u,2},x_{v,3};\delta_n\right) - \frac{1}{n^2}\sum_{u=1}^{n}\sum_{v=1}^{n} G\left(y_v;x_{u,1},x_{u,2},x_{v,3};\delta_n\right)$
$\{x_3\}$	1/6	$\frac{1}{n^2}\sum_{u=1}^{n}\sum_{v=1}^{n} G\left(y_v;x_{v,1},x_{v,2},x_{u,3};\delta_n\right) - \frac{1}{n^2}\sum_{u=1}^{n}\sum_{v=1}^{n} G\left(y_v;x_{u,1},x_{v,2},x_{u,3};\delta_n\right)$
$\{x_2,x_3\}$	1/3	$\frac{1}{n^2}\sum_{u=1}^{n}\sum_{v=1}^{n} G\left(y_v;x_{v,1},x_{u,2},x_{u,3};\delta_n\right) - \frac{1}{n^2}\sum_{u=1}^{n}\sum_{v=1}^{n} G\left(y_v;x_{u,1},x_{u,2},x_{u,3};\delta_n\right)$

Note: This table displays details of empirical Shapley value computation, i.e., the coalitions (column 1), the associated weights (column 2) and the estimated marginal contributions (column 3).

4.2 Local Analysis

Feature contributions to the evaluation metric can be distributed among individuals. We detail the estimation of individual i contribution to the Shapley value ϕ_j as follows.

Definition 4. *Under assumption 3, individual i contribution to the Shapley value ϕ_j is:*

$$\phi_{i,j} = \sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\} \setminus \{x_j\})} w_S \left[\mathbb{E}_{\mathbf{x}^S} \left(G\left(y; \mathbf{x}\right) \right) - \mathbb{E}_{x_j, \mathbf{x}^S} \left(G\left(y; \mathbf{x}\right) \right) \right]. \tag{12}$$

Definition 5. *A consistent estimator of individual i contribution to the estimated Shapley value $\hat{\phi}_j$ is:*

$$\hat{\phi}_{i,j} = \sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\} \setminus \{x_j\})} w_S \left[\frac{1}{n} \sum_{u=1}^{n} G\left(y_i; x_{i,j}, \mathbf{x}_u^S, \mathbf{x}_i^{\overline{S}}; \delta_n\right) - \frac{1}{n} \sum_{u=1}^{n} G\left(y_i; x_{u,j}, \mathbf{x}_u^S, \mathbf{x}_i^{\overline{S}}; \delta_n\right) \right], \tag{13}$$

$$\hat{\phi}_j = \frac{1}{n} \sum_{i=1}^{n} \hat{\phi}_{i,j} \tag{14}$$

with S a coalition, i.e., a subset of features, excluding the feature of interest x_j, $|S|$ the number of features in the coalition, and $\mathcal{P}(\{\mathbf{x}\} \setminus \{x_j\})$ the partition of the set $\{\mathbf{x}\} \setminus \{x_j\}$.

$$\hat{\phi}_{i,0} = \frac{1}{n} \sum_{u=1}^{n} G\left(y_i; x_{u,1}, x_{u,2}, x_{u,3}; \delta_n\right) \tag{15}$$

Table 3: Computation of the Shapley value $\hat{\phi}_{i,1}$ in a three-feature model.

S	w_S	$\frac{1}{n}\sum_{u=1}^{n} G\left(y_i; x_{i,j}, \mathbf{x}_u^S, \mathbf{x}_i^{\overline{S}}; \delta_n\right) - \frac{1}{n}\sum_{u=1}^{n} G\left(y_i; x_{u,j}, \mathbf{x}_u^S, \mathbf{x}_i^{\overline{S}}; \delta_n\right)$
$\{\emptyset\}$	1/3	$G\left(y_i; x_{i,1}, x_{i,2}, x_{i,3}; \delta_n\right) - \frac{1}{n}\sum_{u=1}^{n} G\left(y_i; x_{u,1}, x_{i,2}, x_{i,3}; \delta_n\right)$
$\{x_2\}$	1/6	$\frac{1}{n}\sum_{u=1}^{n} G\left(y_i; x_{i,1}, x_{u,2}, x_{i,3}; \delta_n\right) - \frac{1}{n}\sum_{u=1}^{n} G\left(y_i; x_{u,1}, x_{u,2}, x_{i,3}; \delta_n\right)$
$\{x_3\}$	1/6	$\frac{1}{n}\sum_{u=1}^{n} G\left(y_i; x_{i,1}, x_{i,2}, x_{u,3}; \delta_n\right) - \frac{1}{n}\sum_{u=1}^{n} G\left(y_i; x_{u,1}, x_{i,2}, x_{u,3}; \delta_n\right)$
$\{x_2, x_3\}$	1/3	$\frac{1}{n}\sum_{u=1}^{n} G\left(y_i; x_{i,1}, x_{u,2}, x_{u,3}; \delta_n\right) - \frac{1}{n}\sum_{u=1}^{n} G\left(y_i; x_{u,1}, x_{u,2}, x_{u,3}; \delta_n\right)$

Note: This table displays details of empirical Shapley value computation, i.e., the coalitions (column 1), the associated weights (column 2) and the estimated marginal contributions (column 3).

4.3 Illustrations

4.3.1 Classification model

Consider a probit model and the accuracy as sample evaluation metric with

$$y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{16}$$

where y_i^* refers to a latent variable defined as:

$$y_i^* = \mathbf{x}_i \beta + \varepsilon_i, \tag{17}$$

with $\varepsilon_i \sim \mathcal{N}(0,1)$ an i.i.d random variable.

We generate $q = 3$ features $x_{i,j}$, $j = 1, 2, 3$, for $n = 100$ individuals, $i = 1, \ldots, 100$, and assume the following Gaussian distributions, $x_1 \sim \mathcal{N}(0, 8)$, $x_2 \sim \mathcal{N}(0, 4)$, and $x_3 \sim \mathcal{N}(0, 1)$. Conditional probability of observing the event $y_i = 1$ is obtained as follows:

$$P(\mathbf{x}_i) = \Phi(\mathbf{x}_i \beta), \tag{18}$$

where $\Phi(.)$ represents the cumulative distribution function of the standard normal distribution and $\{\beta_0, \beta_1, \beta_2, \beta_3\} = \{0.05, 0.5, 0.5, 0\}$ the vector of parameters. Finally, the estimated probabilities are defined as:

$$\hat{P}(\mathbf{x}_i) = \Phi(\mathbf{x}_i \hat{\beta}), \tag{19}$$

with $\{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3\} = \{0.06, 0.7, 0.3, 0.08\}$ the vector of estimated parameters.

The AUC and features contributions obtained from our simulations are the following:

$$G_n(\mathbf{y}; \mathbf{X}) \simeq 0.9804$$

$$\underbrace{0.4697}_{\hat{\phi}_0} + \underbrace{0.3988}_{\hat{\phi}_1} + \underbrace{0.1139}_{\hat{\phi}_2} + \underbrace{-0.002}_{\hat{\phi}_3} \simeq 0.9804$$

As we can see the benchmark $\hat{\phi}_0$ obtained from our simulations is very close to the well-known AUC value of 0.5. As a reminder, an AUC equal to 0.5 means that the estimated model is as good as a model making random guesses of the target variable. Hence, our benchmark AUC is very close to the one obtained from of a model making random prediction of y_i. This result is expected as ϕ_0 represents the AUC provided by a model without any predictive ability. As our estimated model has an AUC close to 1 ($\simeq 0.9804$) the features contain important information allowing the model to make correct predictions most of the time. Using our methodology we find that among those features, the first one above all explain the predictive ability of the model ($0.3988/0.9804 \simeq 41\%$). At the opposite, as the last feature has a low impact on the estimated probabilities, i.e., $\hat{\beta}_3 = 0.08$ and $\sigma^2_{x_3} = 1$, its Shapley value $\hat{\phi}_3$ is close to 0 (-0.002).[5]

Local analyses of feature contributions to individual performances $G(y_i; \mathbf{x}_i; \delta_n)$ is detailed in Table 4. In binary classification, the benchmark $\hat{\phi}_{i,0}$ only takes two values as it corresponds to the evaluation metric computed for each target value from a model without any predictive ability. In our case, we can see that $\hat{\phi}_{i,0}$ is either equal to 0.4748 or 0.4648. Individual performances superior to their benchmark, i.e., $G(y_i; \mathbf{x}_i; \delta_n) > \hat{\phi}_{i,0}$, show that knowledge of individual's characteristics empower the model to make more accurate predictions than simple random guesses. In other words, if $G(y_i; \mathbf{x}_i; \delta_n) > \hat{\phi}_{i,0}$ it means that for individual i features x_j, $j = 1, ..., q$ contributes to its correct classification. As expected for most individuals feature x_1 mainly contributes to their good classification, i.e., $\hat{\phi}_1 \gg 0$ and $G(y_i; \mathbf{x}_i; \delta_n) \gg \hat{\phi}_{i,0}$. However, local analysis allows us to see that features impact on performances is not homogeneous among individuals. Indeed, for some individuals the second feature plays a more important role than the first one to promote their correct classification. For instance, for individual 5 feature x_2 contributes to 28.3% of his good performance ($G(y_i; \mathbf{x}_i; \delta_n) \gg \hat{\phi}_{i,0}$) whereas x_1 share is about 21.8%.

[5] Note that even if in the DGP used to simulate the data the third feature does not have any effect on the predicted probabilities it does not implies that $\hat{\phi}_3 = 0$.

Table 4: Illustration of AUC Shapley values in a three-fold logit model.

	$G(y_i;\mathbf{x}_i;\delta_n)$	$\hat{\phi}_{i,0}$	$\hat{\phi}_{i,1}$	$\hat{\phi}_{i,2}$	$\hat{\phi}_{i,3}$
$i=1$	1.0204	0.4748	0.4464	0.0884	0.0108
$i=2$	1.0204	0.4748	0.5137	0.0329	-0.0010
$i=3$	1.0204	0.4748	0.4672	0.0815	-0.0031
$i=4$	1.0004	0.4748	0.4742	0.0612	-0.0097
$i=5$	0.9804	0.4648	0.2133	0.2782	0.0240
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$i=996$	0.9804	0.4648	0.4971	0.0260	-0.0074
$i=997$	0.9204	0.4648	0.2335	0.2146	0.0076
$i=998$	0.9804	0.4648	0.4574	0.0711	-0.0129
$i=999$	0.8203	0.4648	0.1741	0.2170	-0.0355
$i=1,000$	1.0204	0.4748	0.3721	0.1635	0.0100

Note: This table displays individual contributions to the AUC, individual benchmarks, and Shapley values associated to each feature x_j, $j=1,2,3$, in a three-fold logit model.

4.3.2 Regression model

Consider a linear regression model $\hat{f}(\mathbf{x}_i) = \hat{\beta}_0 + \sum_{j=1}^{3}\hat{\beta}_j x_{i,j}$, and the R^2 as sample evaluation metric. We assume that the features are normally distributed and satisfy $\mathbb{E}(\mathbf{x}) = 0_q$ and $\mathbb{V}(\mathbf{x}) = diag(\sigma^2_{x_j})\ \forall j = 1,2,3$.

We generate $q=3$ predictive features $x_{i,j}$, $j=1,\ldots,q$, for $n=1,000$ individuals, $i=1,\ldots,n$, and assume the following Gaussian distributions, $x_1 \sim \mathcal{N}(0,8)$, $x_2 \sim \mathcal{N}(0,4)$, and $x_3 \sim \mathcal{N}(0,1)$. We simulate the true target variable y_i, and predicted target variable $\hat{y}_i$ as follows:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \varepsilon_i, \tag{20}$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2} + \hat{\beta}_3 x_{i,3}, \tag{21}$$

with $\{\beta_0,\beta_1,\beta_2,\beta_3\} = \{\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3\} = \{0.2,0.5,0.5,0\}$ the vector of parameters, and $\varepsilon_i \sim \mathcal{N}(0,4)$ the error term.

The R^2 and features contributions obtained from our simulations are the following:

$$G_n(\mathbf{y};\mathbf{X}) \simeq 0.3948$$

$$\underbrace{-0.3973}_{\hat{\phi}_0} + \underbrace{0.5358}_{\hat{\phi}_1} + \underbrace{0.2563}_{\hat{\phi}_2} + \underbrace{0}_{\hat{\phi}_3} \simeq 0.3948$$

Table 5: Illustration of R^2 Shapley values in a three-fold standard linear model.

	$G(y_i; \mathbf{x}_i; \delta_n)$	$\hat{\phi}_{i,0}$	$\hat{\phi}_{i,1}$	$\hat{\phi}_{i,2}$	$\hat{\phi}_{i,3}$
$i = 1$	-1.0893	-0.9820	-0.3794	0.2721	0.0
$i = 2$	0.5751	0.6006	-0.2142	0.1887	0.0
$i = 3$	0.9018	-0.0310	0.6548	0.2780	0.0
$i = 4$	-1.0454	-1.7017	1.0307	-0.3744	0.0
$i = 5$	0.8222	0.5626	0.2827	-0.0231	0.0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$i = 996$	0.9995	0.2820	0.5569	0.1606	0.0
$i = 997$	-1.4782	-1.0698	2.0814	-2.4898	0.0
$i = 998$	0.9411	0.5965	0.2129	0.1317	0.0
$i = 999$	-0.5503	0.0593	-0.8088	0.1992	0.0
$i = 1,000$	0.9253	0.5173	0.2741	0.1340	0.0

Note: .

5 Feasible Shapley values

6 Empirical Application

A Additive property in a three-fold model

In a three-fold model the additive property is illustrated as follows:

$$\sum_{j=1}^{3}\phi_j = \underbrace{\sum_{j=1}^{3}\frac{1}{3}\left[\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x})) - \phi_0\right]}_{\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x}))-\phi_0} + \underbrace{\sum_{j=1}^{3} M_j(G\left(y;\mathbf{x}\right))}_{=\,0} \tag{22}$$

with

$$\begin{aligned}
M_1(G\left(y;\mathbf{x}\right)) &= \frac{1}{6}\left(\mathbb{E}_{x_2}\mathbb{E}_{y,x_1,x_3}\left(G\left(y;\mathbf{x}\right)\right) - \mathbb{E}_{x_1,x_2}\mathbb{E}_{y,x_3}\left(G\left(y;\mathbf{x}\right)\right)\right)\\
&+ \frac{1}{6}\left(\mathbb{E}_{x_3}\mathbb{E}_{y,x_1,x_2}\left(G\left(y;\mathbf{x}\right)\right) - \mathbb{E}_{x_1,x_3}\mathbb{E}_{y,x_2}\left(G\left(y;\mathbf{x}\right)\right)\right)\\
&+ \frac{1}{3}\left(\mathbb{E}_{x_2,x_3}\mathbb{E}_{y,x_1}\left(G\left(y;\mathbf{x}\right)\right) - \mathbb{E}_{x_1}\mathbb{E}_{y,x_2,x_3}\left(G\left(y;\mathbf{x}\right)\right)\right)\\
\\
M_2(G\left(y;\mathbf{x}\right)) &= \frac{1}{6}\left(\mathbb{E}_{x_1}\mathbb{E}_{y,x_2,x_3}\left(G\left(y;\mathbf{x}\right)\right) - \mathbb{E}_{x_1,x_2}\mathbb{E}_{y,x_3}\left(G\left(y;\mathbf{x}\right)\right)\right)\\
&+ \frac{1}{6}\left(\mathbb{E}_{x_3}\mathbb{E}_{y,x_1,x_2}\left(G\left(y;\mathbf{x}\right)\right) - \mathbb{E}_{x_2,x_3}\mathbb{E}_{y,x_1}\left(G\left(y;\mathbf{x}\right)\right)\right)\\
&+ \frac{1}{3}\left(\mathbb{E}_{x_1,x_3}\mathbb{E}_{y,x_2}\left(G\left(y;\mathbf{x}\right)\right) - \mathbb{E}_{x_2}\mathbb{E}_{y,x_1,x_3}\left(G\left(y;\mathbf{x}\right)\right)\right)\\
\\
M_3(G\left(y;\mathbf{x}\right)) &= \frac{1}{6}\left(\mathbb{E}_{x_2}\mathbb{E}_{y,x_1,x_3}\left(G\left(y;\mathbf{x}\right)\right) - \mathbb{E}_{x_2,x_3}\mathbb{E}_{y,x_1}\left(G\left(y;\mathbf{x}\right)\right)\right)\\
&+ \frac{1}{6}\left(\mathbb{E}_{x_1}\mathbb{E}_{y,x_2,x_3}\left(G\left(y;\mathbf{x}\right)\right) - \mathbb{E}_{x_1,x_3}\mathbb{E}_{y,x_2}\left(G\left(y;\mathbf{x}\right)\right)\right)\\
&+ \frac{1}{3}\left(\mathbb{E}_{x_1,x_2}\mathbb{E}_{y,x_3}\left(G\left(y;\mathbf{x}\right)\right) - \mathbb{E}_{x_3}\mathbb{E}_{y,x_1,x_2}\left(G\left(y;\mathbf{x}\right)\right)\right).
\end{aligned}$$

B Additive property additional examples

Example 6. Consider a logistic regression model $\hat{f}(.)$ with i.i.d features and the sensitivity as evaluation metric then:

$$G_n(\mathbf{y};\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} G(y_i;\mathbf{x}_i;\delta_n) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i\hat{f}(\mathbf{x}_i)}{\frac{1}{n}\sum_{j=1}^{n} y_j}\right],$$

with $G(y_i;\mathbf{x}_i;\delta_n) = \frac{y_i\hat{f}(\mathbf{x}_i)}{\delta_n}$ and $\delta_n = \frac{1}{n}\sum_{j=1}^{n} y_j$.

Example 7. Consider a logistic regression model $\hat{f}(.)$ with i.i.d features and the specificity as evaluation metric then:

$$G_n(\mathbf{y};\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} G(y_i;\mathbf{x}_i;\delta_n) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1-y_i)(1-\hat{f}(\mathbf{x}_i))}{\frac{1}{n}\sum_{j=1}^{n}(1-y_j)}\right],$$

with $G(y_i; \mathbf{x}_i; \delta_n) = \frac{(1-y_i)(1-\hat{f}(\mathbf{x}_i))}{\delta_n}$ and $\delta_n = \frac{1}{n}\sum_{j=1}^{n}(1-y_j)$.

Example 8. Consider a logistic regression model $\hat{f}(.)$ with i.i.d features and the precision as evaluation metric then:

$$G_n(\mathbf{y}; \mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} G(y_i; \mathbf{x}_i; \delta_n) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i\hat{f}(\mathbf{x}_i)}{\frac{1}{n}\sum_{j=1}^{n}\hat{f}(\mathbf{x}_j)}\right],$$

with $G(y_i; \mathbf{x}_i; \delta_n) = \frac{y_i\hat{f}(\mathbf{x}_i)}{\delta_n}$ and $\delta_n = \frac{1}{n}\sum_{j=1}^{n}\hat{f}(\mathbf{x}_j)$.