Measuring the Driving Forces of Predictive Performance: Application to Credit Scoring

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

February 23, 2024

Abstract

In credit scoring, machine learning models are known to outperform standard parametric models. As they condition access to credit, banking supervisors and internal model validation teams need to monitor their predictive performance and to identify the features with the highest impact on performance. To facilitate this, we introduce the XPER methodology to decompose a performance metric (e.g., AUC, R^2) into specific contributions associated with the various features of a classification or regression model. XPER is theoretically grounded on Shapley values and is both model-agnostic and performance metric-agnostic. Furthermore, it can be implemented either at the model level or at the individual level. Using a novel dataset of car loans, we decompose the AUC of a machine-learning model trained to forecast the default probability of loan applicants. We show that a small number of features can explain a surprisingly large part of the model performance. Furthermore, we find that the features that contribute the most to the predictive performance of the model may not be the ones that contribute the most to individual forecasts (SHAP). We also show how XPER can be used to deal with heterogeneity issues and significantly boost out-of-sample performance.

Keywords: Explainability; Credit scoring; Performance metrics; Shapley values

*Aix Marseille Univ, CNRS, AMSE, Marseille, France. Email: sullivan.hue@univ-amu.fr

†University of Orléans (LEO) and Institut Universitaire de France (IUF), 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

Why is the AUC of a given machine learning model equal to 0.7? Which features mainly explain this performance? What are the contributions of the different features to the MSE of a regression model? To answer these questions, we develop and apply a general methodology, called eXplainable PERformance (XPER), which measures the marginal contribution of a particular feature to the predictive performance of a regression or classification model.

Being able to identify the driving forces of the performance of a predictive model is crucial in the context of credit scoring, where complex and opaque machine learning models are increasingly being used by banks and fintechs. Such decomposition is of primary importance both for banking supervisors and internal model validation teams at banks as they need to understand why a given model is working or not, and for which types of borrowers. Furthermore, it also permits to address heterogeneity issues by identifying groups of individuals for which the features have similar effects on performance. One can then estimate group-specific models to improve overall performance and make sure the credit scoring model operates effectively for all borrowers. In this paper, we propose an application where we use XPER to decompose the AUC of a scoring model forecasting loan defaults and to identify the features with the strongest impact on the AUC.

The XPER framework is based on Shapley values (Shapley (1953)). While the latter decomposes a *payoff* among *players* in a *game*, XPER values decomposes a *performance metric* (e.g., AUC, R^2) among *features* in a *model*. More precisely, XPER breaks down the difference between a performance metric and a benchmark value among the various features of the model. Formally, an XPER value is defined as the weighted average marginal contribution of a given feature to a performance metric, obtained in a set of coalitions of other features. For instance, evaluating the XPER value of x_1 in a three-feature model implies to assess the incremental performance due to x_1 by successively considering four subsets or coalitions of features: one coalition including no features, only x_2, only x_3, and both x_2 and x_3.

The application of the Shapley methodology in the context of model performance explanation is not trivial. Indeed, many Shapley values can be defined given the assumptions made on the model, the data, and the features excluded from the coalitions (see for instance Strumbelj and Kononenko (2010), Redell (2019), Kumar et al. (2020), Sundararajan and Najmi (2020), and Aas et al. (2021)). In this paper, we define XPER values by considering the expectation of the performance metric with respect to the joint distribution of the target variable and of the features which are excluded from the coalition. A key advantage of this definition is that the benchmark value has a meaningful interpretation: it corresponds to the performance metric that we would obtain on a hypothetical sample in which the target variable is independent from all the features included in the model, i.e., a fully misspecified model with irrelevant features. Moreover, our method does not require to re-estimate the model (à la Grömping (2007)) nor to pick ad-hoc values for features excluded from the coalition (à la Israeli (2007)).

XPER offers several other advantages. First, as the XPER decomposition is based on Shapley values, it is theoretically grounded and XPER values satisfy several desirable axioms. Second, XPER is model-agnostic as it permits to interpret the predictive performance of any econometric or machine learning model. Third, it is metric-agnostic as it can break down any performance metric: e.g., predictive accuracy (AUC, Gini, accuracy), goodness of fit (R^2), information criterion (AIC, BIC), statistical loss function (MSE, MAE, Q-like), or economic performance metric (profit-and-loss function). Fourth, XPER can be implemented either at the global level or at the local level. At the global level, the XPER value of a given feature measures its contribution to the performance of the model. At the local level, the XPER value of a given feature measures its contribution to the model performance that comes from a given individual. Finally, as the number of coalitions grows fast with the number of features, we propose two estimation procedures: an exact one when the number of features remains moderate and an approximated one, which is a modified version of the Kernel SHAP method of Lundberg and Lee (2017).

We apply our methodology in the credit scoring context. Using a novel, balanced dataset of auto loans provided by an international bank, we demonstrate the usefulness of XPER for decomposing the AUC and other performance metrics of a machine-learning credit scoring model. We show that a small number of features can explain a surprisingly large part of the model performance. Furthermore, the empirical analysis confirms that XPER values differ from standard feature importance metrics. More importantly, we show how to use XPER to boost model performance. To do so, we build homogeneous groups of individuals by clustering them based on their individual XPER values. By construction, within a given group, the features tend to have similar effects on performance (same sign, same strength). We then show that estimating group-specific models yields to a much higher predictive accuracy than with a one-fits-all model.

Our paper contributes to the burgeoning literature on eXplainable Artificial Intelligence (XAI) (Murdoch et al., 2019; Molnar, 2020; Gelman and Vehtari, 2021). One well-known limitation of AI and machine learning methods comes from their opacity and lack of explainability. Most of these algorithms are considered as black boxes in the sense that the corresponding outcomes cannot be easily explained to final users nor related to the initial features (Sun and Wang, 2021). Recently, many explainability methods have been designed to explain black-box models by measuring which features most affect its predictions (see Molnar (2020); Zhao and Hastie (2021); Horel and Giesecke (2022); Liu and Ročková (2023)). One of the most impactful XAI methods is the Shapley additive explanation (SHAP) of Lundberg and Lee (2017), which distributes the prediction of a model among its features (see also Sundararajan et al. (2017); Lundberg et al. (2018); Sundararajan et al. (2020); Casalicchio et al. (2019); Bowen and Ungar (2020); Aas et al. (2021)). While model predictive performance obviously depends on predictions, the contribution of a feature to the performance metric also depends on the true value of the target variable. Our contribution to this literature is to decompose any performance metric using an XAI approach.

Our paper also contributes to the statistical literature on the decomposition of performance metrics. Numerous methods have been proposed for the MSE (Theil, 1971; Ahlburg, 1984), for various inequality measures (Bourguignon, 1979; Shorrocks, 1980, 1982), and for the R^2 (see Grömping (2015) for a survey and Lindeman et al. (1980), Kruskal (1987), Chevan and Sutherland (1991), Johnson (2000), Lahaye and Neely (2020)). Other studies break down performance metrics using Shapley values (Stufken, 1992; Israeli, 2007; Grömping, 2007; Redell, 2019; Casalicchio et al., 2019; Williamson and Feng, 2020; Williamson et al., 2021a,b; Borup et al., 2022). Most of these decompositions are specific to a performance metric and to a model. Furthermore, some of them imply to re-estimate the model with different subsets of features, which may lead to omitted variable bias. Others use ad-hoc values for the features excluded from the coalitions. In contrast, XPER does not require such an assumption, nor any re-estimation, and is metric/model agnostic.

The rest of this paper is structured as follows. We introduce our framework of analysis and the concept of performance metric in Section 2. Section 3 introduces the XPER value decomposition of performance metrics and Section 4 provides some simulation results. Section 5 presents the empirical application and Section 6 concludes the paper and discusses further applications.

2 Framework and performance metrics

We consider a classification or regression problem involving a target variable, denoted y, taking values in $\mathcal{Y} = \{0, 1\}$ or $\mathcal{Y} \subseteq \mathbb{R}$ given the problem considered. The q-vector $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^q$ refers to the input 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 a regression output, such as $\hat{y} = f(\mathbf{x})$. We impose no constraint on the model: it may be parametric or not, linear or not, a weak learner or an ensemble method, etc. For simplicity, for parametric models we exclude the parameters from the notation, i.e., $f(\mathbf{x}) \equiv f(\mathbf{x}; \theta)$. The model is estimated (parametric model) or trained (machine learning algorithm) once for all on a training sample $S_T = \{\mathbf{x}_j, y_j\}_{j=1}^T$. The sample size T is considered as fixed. The trained model, denoted $\hat{f}(.)$, is evaluated on a test sample $S_n = \{\mathbf{x}_i, y_i, \hat{f}(\mathbf{x}_i)\}_{i=1}^n$ with a performance metric (PM). A PM is defined as an assessment measure of the model predictive performance, e.g., MSE, MAE, or R^2 for regression models or, AUC, Brier score, or Gini index, for classification models. More generally, any information criteria, loss function, or economic performance indicator can be considered as PM.

Definition 1. *A sample performance metric* $PM_n \in \Theta \subseteq \mathbb{R}$ *associated to the model* $\hat{f}(.)$ *and the test sample* S_n, *is defined as* $PM_n = \tilde{G}_n(y_1, ..., y_n; \hat{f}(\mathbf{x_1}), ..., \hat{f}(\mathbf{x_n})) = G_n(\mathbf{y}; \mathbf{X})$, *where* $\mathbf{y} = (y_1, ..., y_n)'$ *and* $\mathbf{X} = (\mathbf{x}_1, .., \mathbf{x}_n)'$.

For instance, $\Theta = [0, 1]$ for AUC or R^2, and $\Theta = \mathbb{R}^+$ for MSE or MAE. We introduce the following three assumptions on the PM.

Assumption 1. *The sample PM increases over* Θ *with the predictive performance.*

Assumption 1 simplifies the interpretation of the PM. For instance, the AUC satisfies this assumption, as the higher it is, the better the model is at distinguishing between positive and negative classes. Differently, when dealing with PMs that are negatively correlated with performance, e.g. MSE, we need to consider the opposite of the metric.

Assumption 2. *The sample PM satisfies the following additive assumption:*

$$G_n(\mathbf{y}; \mathbf{X}) = \frac{1}{n} \sum_{i=1}^{n} G(y_i; \mathbf{x}_i; \hat{\delta}_n), \qquad (1)$$

where $G(y_i; \mathbf{x}_i; \hat{\delta}_n)$ *denotes an individual contribution to the PM, and* $\hat{\delta}_n$ *is a nuisance parameter which depends on the test sample* S_n.

For simplicity, we only consider models for which the outcome y_i for instance i only depends on features $\mathbf{x}_i = (x_{i,1}, ..., x_{i,n})'$. 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 assumption becomes $G_n(\mathbf{y}; \mathbf{X}) = n^{-1} \sum_{i=1}^{n} G\left(y_i; \mathbf{w}_i; \hat{\delta}_n\right)$.

Assumption 3. *The sample metric $G_n(\mathbf{y}; \mathbf{X})$ converges to the population performance metric $\mathbb{E}_{y,\mathbf{x}}(G(y; \mathbf{x}; \delta_0))$, where $\mathbb{E}_{y,\mathbf{x}}(.)$ refers to the expected value with respect to the joint distribution of y and $\mathbf{x}$, and $\delta_0 = plim\ \hat{\delta}_n$. In addition, $\mathbb{E}_{y,\mathbf{x}}(G(y; \mathbf{x}; \delta_0))$ exists and is finite.*

These assumptions are consistent with a wide range of PMs. In Appendix A, we detail $G(y; \mathbf{x}; \delta_0)$ for standard PMs associated to regression or classification models. For instance, in the case of a linear regression model $\hat{f}(\mathbf{x}) = \mathbf{x}\hat{\beta}$, when the R^2 is used as PM, we have:

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

with $G(y_i; \mathbf{x}_i; \hat{\delta}_n) = 1 - \hat{\delta}_n^{-1}(y_i - \mathbf{x}_i\hat{\beta})^2$ and $\hat{\delta}_n = n^{-1} \sum_{j=1}^{n} (y_j - \bar{y})^2$. The corresponding population R^2 is defined as $\mathbb{E}_{y,\mathbf{x}}(G(y; \mathbf{x}; \delta_0)) = 1 - \frac{1}{\sigma_y^2} \mathbb{E}_{y,\mathbf{x}}\left((y - \mathbf{x}\hat{\beta})^2\right)$, with $G(y; \mathbf{x}; \delta_0) = 1 - \delta_0^{-1}(y - \mathbf{x}\hat{\beta})^2$ and $\delta_0 = \sigma_y^2$ the variance of the target variable.

3 XPER values

3.1 Definition

Our objective is to identify the contribution of the model's features to its predictive performance, as evaluated by a PM on a given sample. We measure this contribution through Shapley values (Shapley, 1953), a method used in game theory to fairly distribute a payoff $Val(x_1, ..., x_q)$ among several players $x_1, ..., x_q$. The Shapley value ϕ_j measures the marginal impact of a player x_j on the payoff by assessing the changes in $Val(x_1, ..., x_q)$ when this player is added or not to a coalition of players already in the game. Denote by $\mathbf{x}^S$ the vector of players included in a coalition S and $\mathbf{x}^{\overline{S}}$ the vector of excluded players, such that $\{\mathbf{x}\} = \{\mathbf{x}^S\} \cup \{\mathbf{x}^{\overline{S}}\} \cup \{x_j\}$ and $\mathbf{x} = (x_1, ..., x_q)$. The Shapley value is then defined as the weighted average of the marginal contributions associated with all possible coalitions S.

Definition 2 (Shapley (1953))**.** *The Shapley value contribution of player x_j to the payoff is:*

$$\phi_j = \sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\} \setminus \{x_j\})} \omega_S \left[Val(\mathbf{x}^S \cup \{x_j\}) - Val(\mathbf{x}^S) \right], \tag{3}$$

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

with $Val(.)$ the payoff, S a coalition of players, excluding the player of interest x_j, $|S|$ the number of players in the coalition, and $\mathcal{P}(\{\mathbf{x}\} \setminus \{x_j\})$ the powerset of the set $\{\mathbf{x}\} \setminus \{x_j\}$.

By analogy, we decompose the performance metric $\mathbb{E}_{y,\mathbf{x}}(G(y; \mathbf{x}; \delta_0))$ (the "payoff") among the features $x_1, ..., x_q$ (the "players") of the model $\hat{f}(\mathbf{x})$ (the "game"). The main difference with the previous notations is that the payoff $Val(\mathbf{x}; y) = \mathbb{E}_{y,\mathbf{x}}(G(y; \mathbf{x}; \delta_0))$ depends on the joint distribution of the features $x_1, ..., x_q$ and the target variable y. Formally, we define XPER as the Shapley value associated to the performance metric $\mathbb{E}_{y,\mathbf{x}}(G(y; \mathbf{x}; \delta_0))$ and to the model $\hat{f}(\mathbf{x})$.

Definition 3 (XPER)**.** *The XPER value associated to the feature x_j is defined as:*

$$\phi_j = \sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\} \setminus \{x_j\})} \omega_S \left[\underbrace{\mathbb{E}_{y,x_j,\mathbf{x}^S}}_{averaging} \underbrace{\mathbb{E}_{\mathbf{x}^{\overline{S}}}}_{marginalisation} (G(y; \mathbf{x}; \delta_0)) - \underbrace{\mathbb{E}_{y,\mathbf{x}^S}}_{averaging} \underbrace{\mathbb{E}_{x_j \mathbf{x}^{\overline{S}}}}_{marginalisation} (G(y; \mathbf{x}; \delta_0)) \right],$$

with S a coalition of features, excluding the feature of interest x_j, $\mathcal{P}(\{\mathbf{x}\} \setminus \{x_j\})$ the powerset of the set $\{\mathbf{x}\} \setminus \{x_j\}$, and ω_S the coalition weight.

The XPER value ϕ_j measures the weighted average marginal contribution of the feature x_j to the PM over all features coalitions. For each coalition, the marginal contribution of x_j is defined as the difference between the expected values of (1) the PM obtained while including this feature in the coalition and (2) the PM obtained while excluding this feature from the coalition. In Definition 3, the term $\mathbb{E}_{\mathbf{x}^{\overline{S}}}$ refers to the marginalisation with respect to the features which are excluded from the coalition S. The second expectation $\mathbb{E}_{y,x_j,\mathbf{x}^S}(G(y; \mathbf{x}; \delta_0))$ refers to an averaging effect, i.e., expectation with respect to the joint distribution of the features $\mathbf{x}^S$ included in the coalition with x_j, and the target variable y.

As an illustration, we consider a three-feature model $\hat{f}(x_1, x_2, x_3)$ and pick x_1 as the feature of interest. In Table 1, we report all the coalitions among the set $\{x_2, x_3\}$ (column 1), the associated weights computed according to Equation 4 (column 2), and the marginal contributions (column 3) used to compute the XPER value ϕ_1. For instance, when considering a coalition with only x_1, we first compute the expectation of the PM with respect to the joint distribution of excluded features x_2 and x_3, i.e., $\mathbb{E}_{x_2,x_3}(\tilde{G}(y, \hat{f}(x_1, x_2, x_3)))$. This corresponds to a marginalisation of the PM with respect to the features which are excluded from the coalition. Then, we consider the expectation of the PM with respect to the joint distribution of the features included in the coalition and the target, i.e., y and x_1 in our example. Thus, we get an expected value $\mathbb{E}_{y,x_1}\mathbb{E}_{x_2,x_3}(\tilde{G}(y, \hat{f}(x_1, x_2, x_3)))$, where the first expectation refers to the averaging effect, whereas the second one refers to the marginalisation effect. The Shapley value is then computed by summing the weighted differences in the expected PM obtained with or without the feature of interest, for all the coalitions of other features.

Table 1: Components of ϕ_1 in a three-feature model

S	ω_S	$\mathbb{E}_{y,x_1,\mathbf{x}^S}\mathbb{E}_{\mathbf{x}^{\overline{S}}}\left(G\left(y;\mathbf{x};\delta_0\right)\right) - \mathbb{E}_{y,\mathbf{x}^S}\mathbb{E}_{x_1,\mathbf{x}^{\overline{S}}}\left(G\left(y;\mathbf{x};\delta_0\right)\right)$
$\{\emptyset\}$	1/3	$\mathbb{E}_{y,x_1}\mathbb{E}_{x_2,x_3}\left(G\left(y;\mathbf{x};\delta_0\right)\right) - \mathbb{E}_{y}\mathbb{E}_{x_1,x_2,x_3}\left(G\left(y;\mathbf{x};\delta_0\right)\right)$
$\{x_2\}$	1/6	$\mathbb{E}_{y,x_1,x_2}\mathbb{E}_{x_3}\left(G\left(y;\mathbf{x};\delta_0\right)\right) - \mathbb{E}_{y,x_2}\mathbb{E}_{x_1,x_3}\left(G\left(y;\mathbf{x};\delta_0\right)\right)$
$\{x_3\}$	1/6	$\mathbb{E}_{y,x_1,x_3}\mathbb{E}_{x_2}\left(G\left(y;\mathbf{x};\delta_0\right)\right) - \mathbb{E}_{y,x_3}\mathbb{E}_{x_1,x_2}\left(G\left(y;\mathbf{x};\delta_0\right)\right)$
$\{x_2, x_3\}$	1/3	$\mathbb{E}_{y,x_1,x_2,x_3}\left(G\left(y;\mathbf{x};\delta_0\right)\right) - \mathbb{E}_{y,x_2,x_3}\mathbb{E}_{x_1}\left(G\left(y;\mathbf{x};\delta_0\right)\right)$

Note: This table provides details about the XPER value computation, i.e., the coalitions (column 1), the associated weights according to Equation 4 (column 2), and the marginal contributions (column 3).

One advantage of our marginalisation-based approach is that there is no need to re-estimate any sub-model for each subset of features. We consider the expected value of the PM with respect to the features $\mathbf{x}^{\overline{S}}$ excluded from the coalitions, while leaving the model $\hat{f}(\mathbf{x})$ unchanged. An alternative solution would consist in estimating a submodel for each subset of features. For example, for a model with three features x_1, x_2, x_3, the computation of the Shapley value associated to x_1 would require to estimate four sub-models, namely without any feature, with x_2 only, with x_3 only, and with x_2, x_3, and then to re-estimate the same sub-models while including x_1 as an additional feature. This approach was adopted for instance by Israeli (2007) to decompose the R^2 of a linear model. However, re-estimating a submodel with only a subset of the initial features can lead to model specification errors, e.g., omitted variable bias in the case of linear regression. This specification error necessarily distorts the Shapley value and thus may lead to an unreliable decomposition of the performance metric.

3.2 Axioms

The XPER values satisfy different axioms associated to Shapley values. These axioms are particularly relevant in the context of statistical performance analysis.

Axiom 1 (Efficiency)**.** *The sum of the XPER values ϕ_j, $\forall j = 1, ..., q$, is equal to the difference between the performance metric $\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0))$ and a benchmark ϕ_0 such as:*

$$\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0)) = \phi_0 + \sum_{j=1}^{q} \phi_j, \tag{5}$$

where $\phi_0 = \mathbb{E}_{\mathbf{x}}\mathbb{E}_y\left(G(y;\mathbf{x};\delta_0)\right)$ *corresponds to the performance metric associated to a population where the target variable is independent from all features considered in the model.*

One of the main advantages of the XPER decomposition is that the benchmark value ϕ_0 has an insightful interpretation: it corresponds to the PM that we would obtain on a hypothetical sample in which the target variable y is independent from all model features $\mathbf{x}$, i.e., in a case where the model $\hat{f}(\mathbf{x})$ is fully misspecified. As an example, for the AUC, the benchmark ϕ_0 corresponds to the AUC associated to a random predictor and is equal to 0.5. For the sensitivity (true positive rate), the benchmark corresponds to the probability $\Pr(\hat{y} = 1)$, for the specificity (true negative rate) the benchmark is $\Pr(\hat{y} = 0)$, etc. Thus, we can decompose any PM into two parts: (i) a base value ϕ_0 obtained in a hypothetical case where y and $\mathbf{x}$ would be independent, and (ii) a component determined by the XPER feature contributions, which depends on their dependence with the target, i.e., their relevance.

$$\underbrace{\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0)}_{\text{population performance metric}} = \underbrace{\mathbb{E}_y\mathbb{E}_{\mathbf{x}}(G(y;\mathbf{x};\delta_0)}_{\text{expected value under independence}} + \underbrace{\sum_{j=1}^{q}\phi_j}_{\text{feature contributions}}.$$

The XPER values also satisfy the other main axioms of the Shapley values:

Axiom 2 (Symmetry)**.** *If for all subsets* $S \subseteq \mathcal{P}(\{\mathbf{x}\} \setminus \{x_j, x_k\})$

$$\mathbb{E}_{y,x_j,\mathbf{x}^S}\mathbb{E}_{\mathbf{x}^{\overline{S}}}\left(G\left(y;\mathbf{x},x_j;\delta_0\right)\right) = \mathbb{E}_{y,x_k,\mathbf{x}^S}\mathbb{E}_{\mathbf{x}^{\overline{S}}}\left(G\left(y;\mathbf{x},x_k;\delta_0\right)\right),$$

then $\phi_j = \phi_k$. *It means that if two features* x_j *and* x_k *contribute equally to the performance* *of the model then their effects must be the same.*

Axiom 3 (Linearity)**.** *For any two performance metrics PM_1 and PM_2*

$$\phi_j(PM_1 + PM_2) = \phi_j(PM_1) + \phi_j(PM_2),$$

where $\phi_j(PM_1 + PM_2)$ refers to the XPER value of feature x_j measuring its effect on the performance metric defined as the sum of PM_1 and PM_2.

Axiom 4 (Null effects)**.** *If for all subsets $S \subseteq \mathcal{P}(\{\mathbf{x}\} \setminus \{x_j\})$*

$$\mathbb{E}_{y,x_j,\mathbf{x}^S}\mathbb{E}_{\mathbf{x}^{\overline{S}}}\left(G\left(y;\mathbf{x},x_j;\delta_0\right)\right) = \mathbb{E}_{y,\mathbf{x}^S}\mathbb{E}_{x_j,\mathbf{x}^{\overline{S}}}\left(G\left(y;\mathbf{x},x_j;\delta_0\right)\right),$$

then $\phi_j = 0$. A feature without any impact on the performance of the model $\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0))$ has an XPER value equal to 0.

To illustrate these axioms, consider a linear regression model $\hat{f}(\mathbf{x}_i) = \sum_{j=1}^{q}\hat{\beta}_j x_{i,j}$ estimated on a training set, and the R^2 as PM. For simplicity, we assume that the data generating process (DGP) of the test sample $S_n = \{\mathbf{x}_i, y_i, \hat{f}(\mathbf{x}_i)\}_{i=1}^{n}$ satisfies $\mathbb{E}\left(\mathbf{x}\right) = \mu_q$ and $\mathbb{V}(\mathbf{x}) = \Sigma$ a positive semi-definite matrix, with $\Sigma_{k,j} = \sigma_{x_k,x_j}$ the covariance between feature x_k and x_j. Then, the XPER contribution ϕ_j of feature x_j to the R^2 is:

$$\phi_j = \frac{2\hat{\beta}_j\sigma_{y,x_j}}{\sigma_y^2}, \qquad \forall j = 1, ..., q, \tag{6}$$

with σ_y^2 the variance of the target variable and σ_{y,x_j} its covariance with feature x_j. See

Appendix E.1 for the proof. Formally, ϕ_j depends on the estimated parameter $\hat{\beta}_j$ (training set) and the covariance (test set) between x_j and the target variable y, i.e., σ_{y,x_j}. If the DGP of the training and test samples are identical, model parameters $\hat{\beta}_j$ and covariance σ_{y,x_j} have the same sign, which means $\phi_j > 0$. Otherwise, XPER values may be negative. Similarly, if $\hat{\beta}_j = 0$ (i.e., the variable is useless in the model) or if the feature is uncorrelated with the target variable on the test sample ($\sigma_{y,x_j} = 0$), then $\phi_j = 0$. Finally, a feature x_j has a larger contribution to the R^2 than a feature x_s if $\hat{\beta}_j \sigma_{y,x_j} > \hat{\beta}_s \sigma_{y,x_s}$, meaning that x_j is more related to the target variable than x_s both in-sample (through $\hat{\beta}_j$) and out-of-sample (through σ_{y,x_s}).

Most of the time, there is no analytical expression for the XPER values ϕ_j. Indeed, the function $G(y; \mathbf{x}; \delta_0)$ is in general non-linear in y and $\mathbf{x}$, as the model $\hat{f}(.)$ may be highly non-linear in $\mathbf{x}$ (e.g., machine learning model) and/or the performance metric may be non-linear in y and $\mathbf{x}$. One exception is the case of quadratic loss functions, e.g., MSE, R^2, etc., associated to the linear regression model (see Appendix B).

Finally, the XPER values ϕ_j are theoretical and unobserved quantities. In Appendix C, we propose two estimation methods for ϕ_j to accommodate models with either few features or with a large number of features.

3.3 Individual XPER values

The XPER framework can be used to conduct a global analysis of the model predictive performance through feature contributions ϕ_j, or a local analysis at the individual level.

Definition 4. *(Individual XPER) The individual XPER value $\phi_{i,j}$ associated to individual i and model feature j satisfies $\phi_j = \mathbb{E}_{y,\mathbf{x}}(\phi_{i,j}(y_i;\mathbf{x}_i))$, where the random variable $\phi_{i,j}(y_i;\mathbf{x}_i)$ corresponds to individual i and feature j contribution to the performance metric defined as:*

$$\phi_{i,j}(y_i;\mathbf{x}_i) = \sum_{S\subseteq\mathcal{P}(\{\mathbf{x}\}\backslash\{x_j\})} w_S \left[\mathbb{E}_{\mathbf{x}^{\overline{S}}}\left(G\left(y_i;x_{i,j},\mathbf{x}_i^S,\mathbf{x}^{\overline{S}};\delta_0\right)\right) - \mathbb{E}_{x_j,\mathbf{x}^{\overline{S}}}\left(G\left(y_i;x_j,\mathbf{x}_i^S,\mathbf{x}^{\overline{S}};\delta_0\right)\right)\right].$$

The individual XPER value can be interpreted as the contribution of individual i and model feature j to the performance metric. For a given realisation $(y_i,\mathbf{x}_i)$, the corresponding individual contribution to the performance metric can be broken down into:

$$G(y_i;\mathbf{x}_i;\delta_0) = \phi_{i,0} + \sum_{j=1}^{q}\phi_{i,j}, \tag{7}$$

where $\phi_{i,j}$ is the realisation of $\phi_{i,j}(y_i;\mathbf{x}_i)$ and $\phi_{i,0}$ is the realisation of $\phi_{i,0}(y_i) = \mathbb{E}_{\mathbf{x}}(G(y_i;\mathbf{x};\delta_0))$. The benchmark $\phi_{i,0}$ corresponds to the contribution of an individual to the performance metric when the target variable y_i is independent from the features $\mathbf{x}_i$. Therefore, the difference between the individual contribution to the performance metric $G(y_i;\mathbf{x}_i;\delta_0)$ and the individual benchmark $\phi_{i,0}$ is explained by individual XPER values $\phi_{i,j}$.

As an illustration, consider the R^2 of a linear regression. Then, the individual XPER value $\phi_{i,j}$ can be expressed as:

$$\phi_{i,j} = \sigma_y^{-2} \left[\hat{\beta}_j (x_{i,j} - \mathbb{E}(x_j)) A - \hat{\beta}_j^2 (x_{i,j}^2 - \mathbb{E}(x_j^2)) + \sum_{\substack{k=1 \\ k \neq j}}^{q} \hat{\beta}_k \hat{\beta}_j \sigma_{x_k, x_j} \right]. \tag{8}$$

with $A = \left(2y_i - \sum_{\substack{k=1 \\ k \neq j}}^{q} \hat{\beta}_k (x_{i,k} + \mathbb{E}(x_k)) \right)$. See Appendix E.2 for the proof. The value $\phi_{i,j}$ measures the contribution of feature x_j to $G(y_i; \mathbf{x}_i; \delta_0) = 1 - \sigma_y^{-2}(y_i - \hat{f}(\mathbf{x}_i))^2$. A positive individual XPER value $\phi_{i,j}$ means that incorporating the information included in feature x_j improves model prediction for individual i compared to the benchmark, hence increases R^2. Several comments can be made here. First, if $\hat{\beta}_j = 0$, i.e., the feature has no impact on the model outcome, the feature x_j does not have any impact on the R^2, for all individuals. Second, the closer the realization of feature x_j is to its expected value $\mathbb{E}(x_j)$, the lower is the contribution of this feature to the performance metric, for all individuals. A similar result occurs when x_j^2 is close to its expected value. Indeed, when the characteristics of an individual are close to the mean values over the population, his or her contribution to the predictive performance of the model is also close to the average contribution of other individuals.

4 Simulations

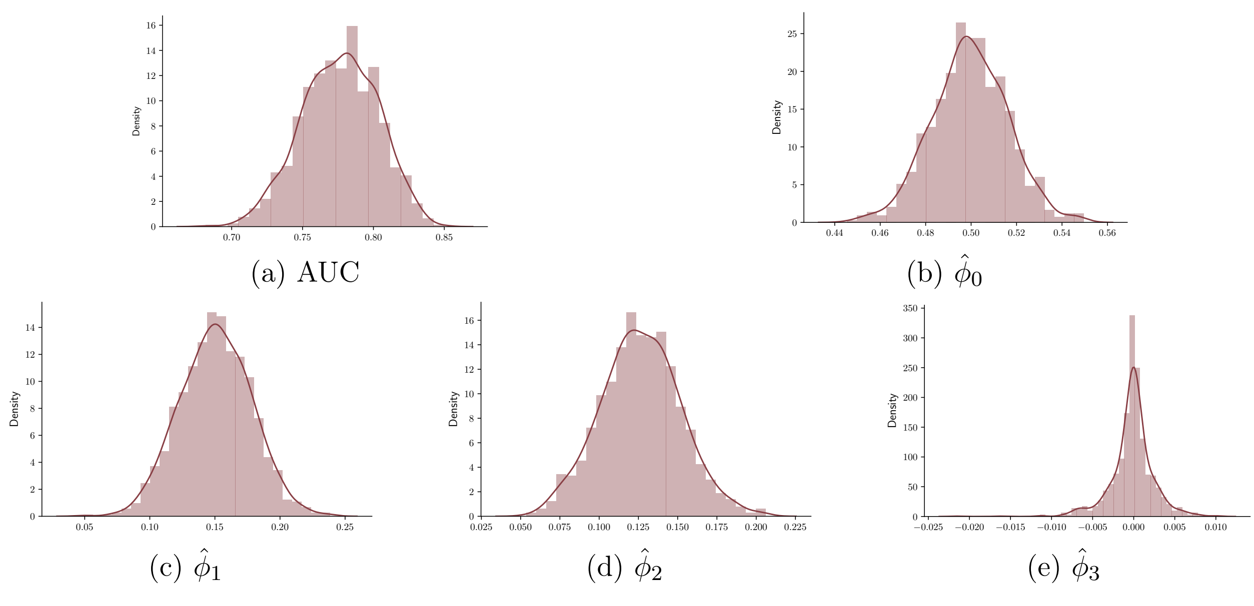
In this section, we conduct a Monte Carlo simulation experiment to illustrate the XPER methodology and to highlight some of its axioms.

In this experiment, XPER values are used to explain the predictive performance of a probit regression model, measured by the AUC computed on a test set S_n. To understand the mechanisms of the XPER decomposition, we consider a white-box model for which the predictions are explainable. The DGP is given by a latent variable model such that $y_i = \mathbf{1}(y_i^* > 0)$, where $\mathbf{1}(.)$ is the indicator function, $y_i^* = \omega_i\beta + \varepsilon_i$, $\omega_i = (1 : \mathbf{x}_i^{'})$ and ε_i an i.i.d. error term with $\varepsilon_i \sim \mathcal{N}(0, 1)$. We consider three i.i.d. features such that $\mathbf{x}_i = (x_{i,1}, x_{i,2}, x_{i,3})^{'} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ with $diag(\Sigma) = (1.2, 1, 1)$. The true vector of parameters is $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^{'} = (0.05, 0.5, 0.5, 0)^{'}$ with β_0 the intercept.

We simulate $K = 5,000$ pseudo-samples $\{y_i^s, \mathbf{x}_i^s\}_{i=1}^{T+n}$ of size 1,000, for $s = 1, \ldots, K$. For each pseudo-sample, we use the first $T = 700$ observations to estimate a probit model and the remaining ones as test set S_n to compute the AUC and the corresponding XPER values according to Equation A1. For instance, the estimated parameters obtained for the simulation $s = 1$ are equal to $\{\hat{\beta}_0^s, \hat{\beta}_1^s, \hat{\beta}_2^s, \hat{\beta}_3^s\} = \{0.0109, 0.4943, 0.5234, 0.0688\}$ and the AUC is equal to 0.7775. The associated feature contributions are the following:

$$\underbrace{0.7775}_{AUC} = \underbrace{0.4984}_{\hat{\phi}_0} + \underbrace{0.1716}_{\hat{\phi}_1} + \underbrace{0.1098}_{\hat{\phi}_2} + \underbrace{(-0.0023)}_{\hat{\phi}_3}.$$

As expected, the estimated benchmark $\hat{\phi}_0$ is close to 0.5. As a reminder, an AUC equal to 0.5 is associated to a random predictor. Hence, the benchmark $\hat{\phi}_0$ corresponds to the AUC of the model $\hat{f}(\mathbf{x}_i) = \mathbf{x}_i^{'}\hat{\beta}$ that we would obtain on a virtual test sample in which the target variable is independent from all features. The difference between the estimated AUC and this hypothetical benchmark is explained by the feature contributions. We verify that these contributions are positive or null for all features. The feature with the largest variance (x_1) has also the largest contribution to the predictive ability of the model ($0.1716/(0.7775 - 0.4984) \simeq 62\%$). On the contrary, as the third feature is excluded from the model, its contribution to the AUC is close to zero. Figure 1 displays the empirical distributions of AUC, benchmark values $\hat{\phi}_0$, and XPER values associated to features x_1, x_2, and x_3, computed from the K simulations. It confirms the robustness of our analysis, but also illustrates the possibility to make inference on XPER values by Bootstrap or other numerical methods.



(a) AUC (b) $\hat{\phi}_0$

(c) $\hat{\phi}_1$ (d) $\hat{\phi}_2$ (e) $\hat{\phi}_3$

Figure 1: Empirical distributions of AUC and XPER values

Note: This figure displays the empirical distributions of the AUC and XPER values on the test sample according to the DGP detailed in Illustration 1. The solid red lines refer to kernel density estimations.

The third column of Table 2 displays the individual benchmark. For a binary classification

Table 2: Illustration of AUC XPER values in a three-fold probit model

	$G(y_i;\mathbf{x}_i;\hat{\delta}_n)$	$\hat{\phi}_{i,0}$	$\hat{\phi}_{i,1}$	$\hat{\phi}_{i,2}$	$\hat{\phi}_{i,3}$	y_i	$\hat{\mathbb{P}}(y_i=1\|\mathbf{x}_i)$
i=1	0.9000	0.5001	0.2450	0.1771	-0.0221	1	0.6614
i=2	1.0067	0.5001	0.2142	0.3203	-0.0279	1	0.8785
...	...	...	...	...	...	...	...
i=300	0.2533	0.4967	-0.1232	-0.1237	0.0035	0	0.7616
	0.7775	0.4984	0.1716	0.1098	-0.0023	0.4967	0.4941

Note: This table displays individual contributions to the AUC, individual benchmarks, and XPER values associated to each feature x_j, $j=1,2,3$, in a three-fold probit model. The last row of the table reports average values of the columns.

model, the benchmark $\hat{\phi}_{i,0} \equiv \hat{\phi}_{i,0}(y_i)$ only takes two values. In our simulations, for individuals $y_i = 1$ (respectively $y_i = 0$), this value is equal to 0.5001 (respectively 0.4967). Remind that the individual benchmark corresponds to the contribution to the AUC obtained from a random predictor for an individual with a target value $y_i = 1$ or $y_i = 0$. Consider the first instance $i = 1$ with $y_i = 1$, $G(y_1;\mathbf{x}_1;\hat{\delta}_n) = 0.9$ and $\hat{\phi}_{1,0} = 0.5001$. As $G(y_1;\mathbf{x}_1;\hat{\delta}_n) > \hat{\phi}_{1,0}$, it means that the features allow the probit model to better predict the event y_1 for this instance than a random predictor.

Finally, columns 4, 5 and 6 report the XPER values associated to features x_1, x_2, x_3. First, we verify that feature x_3 has close to no impact on the AUC for all instances. Second, the heterogeneity of contributions to the AUC depends on individual characteristics. Remind that at the global level, the contribution of feature x_1 is higher than the one of feature x_2. However, at the local level, we observe for instance $i = 2$ that the contribution of feature x_1 is smaller than the one of feature x_2, i.e., $0.2142 < 0.3203$. Third, some individual contributions are negative. For instance, the negative contribution of feature x_2 for individual

$i = 300$ implies that, for this instance, this feature disturbs the model to predict the true target value $y_i = 0$.

To further illustrate the XPER methodology, we provide in Appendix F two additional Monte Carlo simulation experiments illustrating how XPER values can be used to detect the origin of overfitting.

5 Empirical application

5.1 Data and model

We implement our methodology on a proprietary database of auto loans provided by an international bank. For each borrower, we know whether he or she has eventually defaulted ($y = 1$) or not ($y = 0$) on the loan. Given the sensitive nature of the data, we had to randomly under-sample individuals to set the default rate to an arbitrary 20% level. Besides benefits in terms of confidentiality, setting a high arbitrary default rate also protects us against concerns arising from using an unbalanced database. After under-sampling, our database includes 7,440 borrowers. Besides the default target variable, we have access to ten features on the loan (funding amount, funded duration, vehicle price, down-payment) and on the borrower (job tenure, age, marital status, monthly payment in percentage of income, home ownership status, credit event). We divide the database into a stratified training (70%) and test (30%) samples to have the same default rate in both subsamples.

We provide in Table A4 some summary statistics about the features and the target variable. In our dataset, a typical loan amounts to around 11,500 euros, finances a 13,000 euro car, and lasts for 56 months. A typical borrower is 45 years old, married, not owning his or her home, has spent nine years in the same job, experienced no credit event over the past six months, provides less than a 50% down payment, and allocates 10% of his or her monthly income to reimburse the car loan. To get a first sense of the role of each feature on default, we display in Figure A3 their distributions separately for defaulting and non-defaulting borrowers. This preliminary test indicates that the list of discriminating feature includes age, credit event, down payment, marital and ownership status.

Using the training sample, we estimate an XGBoost model to predict default. We selected this type of model because it is recognised as one of the most powerful scoring engines (Gunnarsson et al., 2021). Another reason for using an XGBoost is its black-box nature. Indeed, while the XPER methodology is model-agnostic, we believe it is interesting to assess its usefulness when used with a particularly complex and opaque algorithm. We select the hyperparameters of the XGBoost using stratified five-fold cross-validation based on a balanced accuracy criteria and a random search algorithm (Bergstra and Bengio, 2012).

Table 3 displays the values of six performance metrics for the XGBoost model obtained on the training and test samples. Specifically, (1) the AUC measures the *discriminatory* *ability* of the model, (2) the Brier Score evaluates the *accuracy of probabilities*, and (3) the accuracy, Balanced Accuracy (BA), sensitivity, and specificity assess the *correctness of categorical predictions*. As shown in Table 3, the XGBoost has an AUC of 0.7521, a Brier Score of 0.1433, and an accuracy of 79.53 on the test sample. We observe some over-fitting in the model as its performances drops slightly from the training sample to the test sample. Overall, the performance metrics of our model are comparable to those displayed in Gunnarsson et al. (2021).

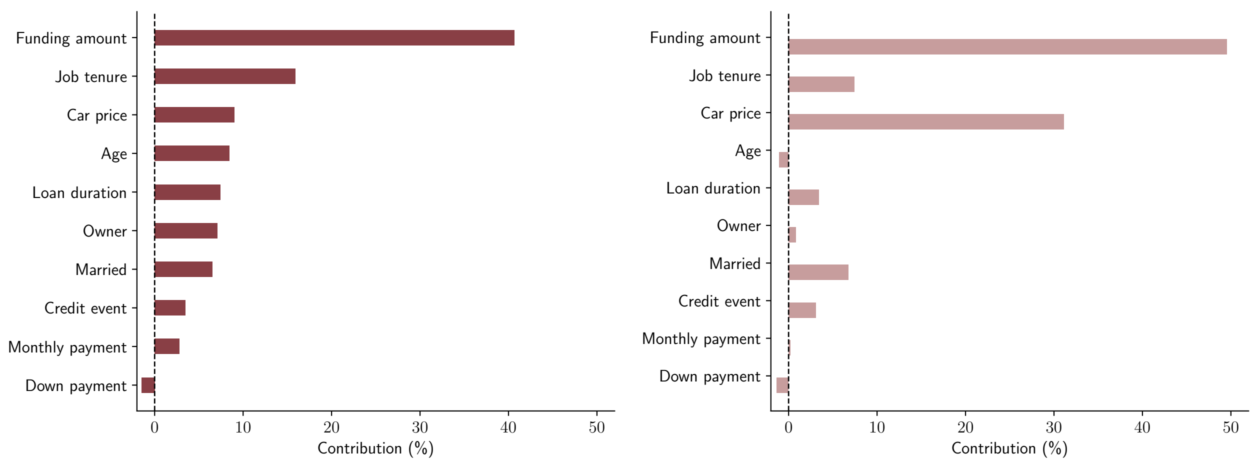
Table 3: XGBoost Performances

Sample	Size (%)	AUC	Brier Score	Accuracy	BA	Sensitivity	Specificity
Training	70	0.8969	0.0958	86.98	72.43	48.18	96.69
Test	30	0.7521	0.1433	79.53	58.69	23.99	93.39

5.2 XPER decomposition

We display in Figure 2a the decomposition of the AUC among the ten features obtained with the estimation method detailed in Appendix C.2. For ease of presentation, we express the feature contributions in percentage of the spread between the AUC and its benchmark (see Equation 5). As shown in Table 3, the AUC in the test sample is equal to 0.7521, which is significantly better than the benchmark value of 0.5 obtained for a random predictor. We see that around 40% of this over-performance is coming from *funding amount*. The second most contributing feature is *job tenure*, which accounts for another 18%. It is interesting to note that with only two features, we can explain more than half of the performance of the model.

Next, we have five features which contribute each for another 8-10% to the performance. At the other side of the spectrum, *down payment* is the feature contributing the least to the AUC. This feature does not help the model to better predict default than a random predictor as its XPER value is close to 0. Note that in Appendix G.3, we also analyse the effect of the various features on the performance for each borrower individually.



(a) XPER values

(b) PI-based feature contributions

Figure 2: XPER decomposition and Permutation Importance (PI)

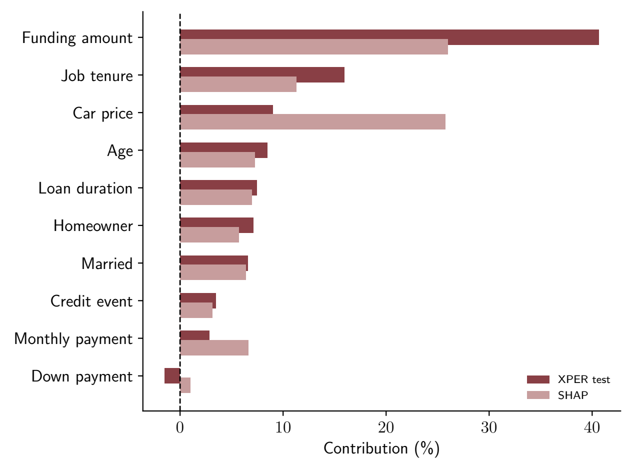
Note: This figure displays in Panel (a) the XPER values for the AUC of the XGBoost model estimated on the test sample and in Panel (b) the feature contributions for the AUC based on Permutation Importance.

We now compare the XPER performance decomposition to standard feature contribution methods commonly used in machine learning to assess the impact of a feature on performance or to explain the output of a black-box model, namely Permutation Importance (PI), feature importance, and SHAP values.

First, we distinguish between the XPER decomposition of the AUC and the permutation importance introduced by Breiman (2001). The latter computes for a feature x_j, the decrease in the accuracy of the model when the values of this feature are randomly reshuffled across instances. As the permutation breaks the dependency between the target variable and the feature, the resulting drop in model accuracy indicates how much the model depends on the feature. In our analysis, the PI results indicate the average decrease in accuracy when the values are reshuffled to obtain more robust results. As for the previous methods, we divide each feature contribution by the sum of all features contributions to compare them to the XPER values. As shown in Figure 2b, XPER values and PI deliver very distinct results. For instance, the contribution of the job tenure is twice as important with XPER than with PI, whereas the contribution of the car price is twice as large with PI than with XPER. Moreover, unlike with XPER, the age of the customer, home-ownership status, and monthly payment loan have virtually no impact on the performance of the model according to PI. Therefore, even if PI also assesses the effect of the features on the performance of the model it delivers different results than XPER.

Second, we contrast in Figure 3a the XPER values of the AUC and the XGBoost-based feature importance. The latter computes for a feature x_j, the average gain across all splits the feature is used in. For ease of comparison, we divide each feature contribution by the sum of the ten feature contributions. As shown in Figure 3a, the result is rather striking. Indeed, the two methodologies lead to very different contributions as some dominating features in a given methodology play a minor role in the other. For instance, *credit event* exhibits the highest feature importance but it is only the $9th$ most contributing feature according to XPER. Differently, funding amount plays a very important role to explain performance but does not contribute much in terms of feature importance.

(a) XPER vs. feature importance



(b) XPER vs. SHAP

Figure 3: XPER vs. other feature contribution methodologies

Note: This figure compares the XPER values of the AUC with (a) the XGBoost-based feature importances and (b) the average absolute SHAP values.

Third, we compare in Figure 3b the XPER decomposition of the AUC with the SHAP values introduced by Lundberg and Lee (2017). In our context, the SHAP values assess the impact of the different features on the probabilities of default of each borrower. As it is the norm, we take the average absolute SHAP values for each feature to assess the feature contribution at the model level. As before, we divide each feature contribution by the sum of all features contributions to express them in percentages. In Figure 3b, we see that SHAP and XPER provide for some features very different information. For instance, the *car price* contribution is more than twice as important for SHAP than for XPER. Similarly, the *funding amount* contribution is around 40% for XPER whereas only 28% for SHAP. We conclude that XPER delivers different, incremental information over SHAP at the model level.

5.3 Using XPER to boost model performance

We now show how XPER can be used to deal with heterogeneity issues and improve out-of-sample performance. In practice, it is often challenging to estimate a single model able to correctly estimate the relationship between the target variable and the features for all individuals in the sample. In this section, we propose an alternative two-step procedure. First, we build homogeneous groups of individuals displaying similar XPER values using a clustering algorithm. Second, within each group, we estimate a group-specific model. Given the definition of XPER, in each group, the features have a similar impact on the target variable. This alternative approach is likely to increase the model performance compared to the one-fits-all model, which has been used so far in the empirical application.

In the first step, we use the KMedoid method to create two clusters of individuals with similar XPER values $\hat{\phi}_{i,j}$ on the training sample. We see in Figure 4 that individuals in group 1 tend to be older, married, home-owners and ask for a mid-range loan to finance a mid-priced car. One striking difference is that in group 2, individuals tend to finance cars with extremely low prices or high prices. Overall, group 1 gathers individuals displaying a lower risk profile compared to those in group 2. This is confirmed by the fact that the default rate is much higher in group 2 (52%) than in group 1 (13%). Moreover, as shown in Figures A7a and A7b in Appendix G.4, the feature distributions are very different for defaulters and non-defaulters in each group. We also see in Figure A8 that in group 2, using the car price deteriorates the performance of the model as its XPER value is close to -20% vs. 20% in group 1. Therefore, these results suggest that estimating group-specific models is likely to boost predictive performance.

In the second step, we estimate using the training dataset one XGBoost model per group to predict default. Then, we aim to compare the resulting out-of-sample performance with the one of the one-fits-all model. To do so, we first assign each individual of the test sample to either group 1 or group 2 using the clustering rule built on the training sample. We see in Table 4 that the performance of our strategy (column 2) is significantly higher than the performance of the initial model (column 1). For instance, the AUC of the model on the test sample increases by 16 percentage points, from 0.752 to 0.912. Another interesting result is that by using two distinct models, the bank would be able to correctly detect 64% of the defaulting borrowers against only 24% with the initial model. Yet, it would still be able to correctly detect more than 95% of the non-defaulting borrowers.

Finally, we benchmark our strategy with a standard clustering approach based on the features themselves. Specifically, we use the K-prototype algorithm to create two clusters of individuals with comparable feature values. As shown in column 3, the performance of the standard approach is very close to the one of the one-fits-all model (AUC = 0.744 vs. 0.752), and is much lower than the performance of the XPER-based models.

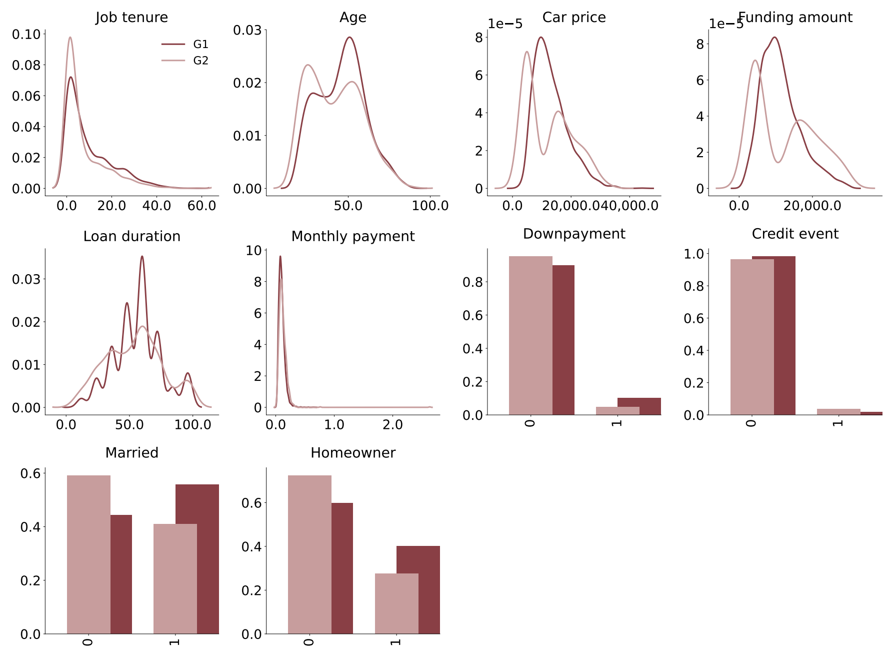


Figure 4: Features distribution by group based on XPER values

Note: This figure displays the distribution of the features on the training sample by group created from individual XPER values using the KMedoids methodology. For continuous features, we use kernel density estimation. Dark red refers to the first group and light red to the second group.

Table 4: Model performances

	Initial (1)	Clusters on XPER values (2)	Clusters on features (3)
AUC	0.752	0.912	0.744
Brier score	0.143	0.080	0.151
Accuracy	79.53	89.11	79.53
Balanced Accuracy	58.69	79.74	59.11
Sensitivity	23.99	64.13	25.11
Specificity	93.39	95.35	93.11

6 Discussion

We have introduced XPER, a methodology designed to measure the feature contributions to the performance of any regression or classification model. XPER is built on Shapley values and interpretability tools developed in machine learning but with the distinct objective of focusing on model performance, such as AUC or R^2, and not on model predictions, $\hat{y}$.

Specifically, XPER breaks down the difference between a performance metric of the model and a benchmark value. The latter corresponds to the performance metric that we would obtain on a hypothetical sample in which the target variable is independent from all the features included in the model. As such, the XPER decomposition only focuses on the part of the performance that directly originate from the features. Other advantages of the method include being implementable either at the model level or at the individual level, not being plagued by model specification error, as it does not require re-estimating the model, and having, as a special case, the standard explainability method in machine learning (SHAP).

We show that identifying the driving forces of the performance of a predictive model is very useful in practice. In a loan default forecasting application, XPER appears to be able to efficiently deal with heterogeneity issues and to significantly boost out-of-sample performance. To do so, we create homogeneous groups of borrowers by clustering them based on their individual XPER values. We find that estimating group-specific models yields to a much higher predictive accuracy than with a one-fits-all model.

Several applications of our method could be envisioned in the future. In medical applications, XPER could be used to identify the variables that have the highest impact on the accuracy of a model forecasting, for instance, the resistance of a virus (Williamson et al., 2021a). In finance, our methodology could identify the main drivers of the financial performance of a portfolio of assets constructed using a large number of state variables. Another natural application would be to choose variables during the selection phase of model development. Indeed, by contrasting the XPER values estimated in the training and validation sets, one could only keep the variables maintaining a high-enough level of importance in the validation test. Furthermore, XPER could go beyond performance. Indeed, as XPER can decompose any function of both $\hat{y}$ and y, one could use it to identify the features responsible for the lack of algorithmic fairness of a given machine learning model (Chen et al., 2023).

References

Aas, K., Jullum, M., and Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*, 298:103502.

Ahlburg, D. A. (1984). Forecast evaluation and improvement using Theil's decomposition. *Journal of Forecasting*, 3(3):345–351.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 12:281–305.

Borup, D., Coulombe, G., Rapach, D. E., Schütte, E. C. M., and Schwenk-Nebbe, S. (2022). The Anatomy of Out-of-Sample Forecasting Accuracy. *Federal Reserve Bank of Atlanta*.

Bourguignon, F. (1979). Decomposable Income Inequality Measures. *Econometrica*, 47(4):901–920.

Bowen, D. and Ungar, L. (2020). Generalized shap: Generating multiple types of explanations in machine learning. *arXiv preprint arXiv:2006.07155*.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Casalicchio, G., Molnar, C., and Bischl, B. (2019). Visualizing the feature importance for black box models. In *Machine Learning and Knowledge Discovery in Databases*, pages 655–670. Springer International Publishing.

Chen, H., Lu, W., Song, R., and Ghosh, P. (2023). On learning and testing of counterfactual fairness through data preprocessing. *Journal of the American Statistical Association*, pages 1–11.

Chevan, A. and Sutherland, M. (1991). Hierarchical partitioning. *The American Statistician*, 45(2):90–96.

Gelman, A. and Vehtari, A. (2021). What are the most important statistical ideas of the past 50 years? *Journal of the American Statistical Association*, 116(536):2087–2097.

Grömping, U. (2015). Variable importance in regression models. *Wiley interdisciplinary reviews: Computational statistics*, 7(2):137–152.

Grömping, U. (2007). Estimators of relative importance in linear regression based on variance decomposition. *The American Statistician*, 61(2):139–147.

Gunnarsson, B. R., vanden Broucke, S., Baesens, B., Óskarsdóttir, M., and Lemahieu, W. (2021). Deep learning for credit scoring: Do or don't? *European Journal of Operational Research*, 295(1):292–305.

Horel, E. and Giesecke, K. (2022). Computationally efficient feature significance and importance for predictive models. In *Proceedings of the Third ACM International Conference on AI in Finance*, pages 300–307.

Israeli, O. (2007). A Shapley-based decomposition of the R-square of a linear regression. *Journal of Economic Inequality*, 5(2):199–212.

Johnson, J. W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate Behavioral Research*, 35(1):1–19.

Kruskal, W. (1987). Relative importance by averaging over orderings. *The American Statistician*, 41(1):6–10.

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., and Friedler, S. A. (2020). Problems with shapley-value-based explanations as feature importance measures. *International Conference on Machine Learning*.

Lahaye, J. and Neely, C. (2020). The role of jumps in volatility spillovers in foreign exchange markets: meteor shower and heat waves revisited. *Journal of Business & Economic Statistics*, 38(2):410–427.

Lindeman, R. H., Merenda, P. F., and Gold, R. Z. (1980). *Introduction to Bivariate and Multivariate Analysis*. Glenview.

Liu, Y. and Ročková, V. (2023). Variable selection via thompson sampling. *Journal of the American Statistical Association*, 118(541):287–304.

Lundberg, S. M., Erion, G. G., and Lee, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

Molnar, C. (2020). *Interpretable machine learning*. Lulu.com.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080.

Perron, P. and Yamamoto, Y. (2021). Testing for changes in forecasting performance. *Journal of Business & Economic Statistics*, 39(1):148–165.

Redell, N. (2019). Shapley decomposition of R-squared in machine learning models. *arXiv preprint arXiv:1908.09718*.

Shapley, L. (1953). The value of a player in n-person games. *Contributions to the Theory of Games*, 2:307–317.

Shorrocks, A. F. (1980). The Class of Additively Decomposable Inequality Measures. *Econometrica*, 48(3):613–625.

Shorrocks, A. F. (1982). Inequality decomposition by factor components. *Econometrica*, 50(1):193–211.

Strumbelj, E. and Kononenko, I. (2010). An Efficient Explanation of Individual Classifications using Game Theory. *Journal of Machine Learning Research*, 11:1–18.

Stufken, J. (1992). On hierarchical partitioning. *The American Statistician*, 46:70–71.

Sun, Y. and Wang, L. (2021). Stochastic tree search for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association*, 116(533):421–432.

Sundararajan, M., Dhamdhere, K., and Agarwal, A. (2020). The shapley taylor interaction index. *International Conference on Machine Learning*, pages 9259–9268.

Sundararajan, M. and Najmi, A. (2020). The many Shapley values for model explanation. *International Conference On Machine Learning*, pages 9269–9278.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. *International Conference on Machine Learning*, pages 3319–3328.

Theil, H. (1971). *Applied Economic Forecasting.* North-Holland, Amsterdam.

Williamson, B. and Feng, J. (2020). Efficient nonparametric statistical inference on population feature importance using shapley values. In *International Conference on Machine Learning*, pages 10282–10291. PMLR.

Williamson, B. D., Gilbert, P. B., Carone, M., and Simon, N. (2021a). Nonparametric variable importance assessment using machine learning techniques. *Biometrics*, 77(1):9–22.

Williamson, B. D., Gilbert, P. B., Simon, N. R., and Carone, M. (2021b). A general framework for inference on algorithm-agnostic variable importance. *Journal of the American Statistical Association*, pages 1–14.

Zhao, Q. and Hastie, T. (2021). Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, 39(1):272–281.

Appendix

A Examples of performance metrics

Table A1: Performance metrics

Panel A: Regression models

Metrics	$G_n(\mathbf{y},\mathbf{x})$	$G(y_i;\mathbf{x}_i;\hat{\delta}_n)$	$\hat{\delta}_n$
MAE	$\frac{1}{n}\sum_{i=1}^{n}\left\lvert y_i-\hat{f}(\mathbf{x}_i)\right\rvert$	$\left\lvert y_i-\hat{f}(\mathbf{x}_i)\right\rvert$	$\emptyset$
MSE	$\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{f}(\mathbf{x}_i)\right)^2$	$\left(y_i-\hat{f}(\mathbf{x}_i)\right)^2$	$\emptyset$
R^2	$1-\frac{\sum_{i=1}^{n}\left(y_i-\hat{f}(\mathbf{x}_i)\right)^2}{\sum_{j=1}^{n}(y_j-\bar{y})^2}$	$1-\hat{\delta}_n^{-1}\left(y_i-\hat{f}(\mathbf{x}_i)\right)^2$	$n^{-1}\sum_{j=1}^{n}(y_j-\bar{y})^2$

Panel B: Classification models

Metrics	$G_n(\mathbf{y},\mathbf{x})$	$G(y_i;\mathbf{x}_i;\hat{\delta}_n)$	$\hat{\delta}_n$
Accuracy	$\frac{1}{n}\sum_{i=1}^{n}\left(y_i\hat{f}(\mathbf{x}_i)+(1-y_i)(1-\hat{f}(\mathbf{x}_i))\right)$	$y_i\hat{f}(\mathbf{x}_i)+(1-y_i)(1-\hat{f}(\mathbf{x}_i))$	$\emptyset$
BA	$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\left[\frac{y_i\hat{f}(\mathbf{x}_i)}{\frac{1}{n}\sum_{j=1}^{n}y_j}+\frac{(1-y_i)(1-\hat{f}(\mathbf{x}_i))}{\frac{1}{n}\sum_{j=1}^{n}(1-y_j)}\right]$	$\frac{1}{2}\left[\hat{\delta}_{n_1}^{-1}\left(y_i\hat{f}(\mathbf{x}_i)\right)+\hat{\delta}_{n_2}^{-1}\left((1-y_i)(1-\hat{f}(\mathbf{x}_i))\right)\right]$	$\hat{\delta}_{n_1}=\frac{1}{n}\sum_{j=1}^{n}y_j$ $\hat{\delta}_{n_2}=\frac{1}{n}\sum_{j=1}^{n}(1-y_j)$
Brier score	$\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{P}(\mathbf{x}_i)\right)^2$	$\left(y_i-\hat{P}(\mathbf{x}_i)\right)^2$	$\emptyset$
Precision	$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i\hat{f}(\mathbf{x}_i)}{\frac{1}{n}\sum_{j=1}^{n}\hat{f}(\mathbf{x}_j)}\right)$	$\hat{\delta}_n^{-1}y_i\hat{f}(\mathbf{x}_i)$	$\frac{1}{n}\sum_{j=1}^{n}\hat{f}(\mathbf{x}_j)$
Sensitivity	$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i\hat{f}(\mathbf{x}_i)}{\frac{1}{n}\sum_{j=1}^{n}y_j}\right)$	$\hat{\delta}_n^{-1}y_i\hat{f}(\mathbf{x}_i)$	$\frac{1}{n}\sum_{j=1}^{n}y_j$
Specificity	$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{(1-y_i)(1-\hat{f}(\mathbf{x}_i))}{\frac{1}{n}\sum_{j=1}^{n}(1-y_j)}\right)$	$\hat{\delta}_n^{-1}(1-y_i)(1-\hat{f}(\mathbf{x}_i))$	$\frac{1}{n}\sum_{j=1}^{n}(1-y_j)$
AUC	$\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(1-y_i)y_jI(\hat{P}(\mathbf{x}_i)<\hat{P}(\mathbf{x}_j))}{\sum_{j=1}^{n}y_j\sum_{j=1}^{n}(1-y_j)}$ $I(\hat{P}(\mathbf{x}_i)<\hat{P}(\mathbf{x}_j))=\begin{cases}0 & \text{if } \hat{P}(\mathbf{x}_i)>\hat{P}(\mathbf{x}_j)\\ 0.5 & \text{if } \hat{P}(\mathbf{x}_i)=\hat{P}(\mathbf{x}_j)\\ 1 & \text{if } \hat{P}(\mathbf{x}_i)<\hat{P}(\mathbf{x}_j)\end{cases}$	$\left((1-y_i)\times\hat{\delta}_{n_1}(\mathbf{x}_i)\right)\hat{\delta}_{n_2}^{-1}$	$\hat{\delta}_{n_1}(\mathbf{x}_i)=\frac{1}{n}\sum_{j=1}^{n}y_jI(\hat{P}(\mathbf{x}_i)<\hat{P}(\mathbf{x}_j))$ $\hat{\delta}_{n_2}=\frac{1}{n^2}\sum_{j=1}^{n}y_j\sum_{j=1}^{n}(1-y_j)$

Note: This table displays the expression of sample performance metrics $G_n(\mathbf{y},\mathbf{x})$, individual contribution to the sample performance metric $G(y_i;\mathbf{x}_i;\hat{\delta}_n)$, and the corresponding nuisance parameter $\hat{\delta}_n$. We distinguish between the performance metric associated to regression models (Panel A) and those related to classification models (Panel B).

B Examples of XPER values decomposition

We provide several examples in Table A2 of the XPER decomposition of regression and classification performance metrics. For regression models, we consider a linear regression model $\hat{f}(\mathbf{x}_i) = \sum_{j=1}^{q} \hat{\beta}_j x_{i,j}$, where we assume that the DGP generating the test sample $S_n = \{\mathbf{x}_i, y_i, \hat{f}(\mathbf{x}_i)\}_{i=1}^{n}$ satisfies $\mathbb{E}\left(\mathbf{x}\right) = 0_q$ and $\mathbb{V}(\mathbf{x}) = diag(\sigma^2_{x_j})\ \forall j = 1, ..., q$, and $\mathbb{E}(y) = 0$. We denote by σ^2_y the variance of the target variable and by σ_{y,x_j} the covariance between the feature x_j and the target variable. For classification models, we consider any binary classification model $\hat{f}(\mathbf{x})$, with $\hat{P}(\mathbf{x}) = \hat{\mathbb{P}}(y = 1|\mathbf{x})$ the estimated probability of belonging to class 1 ($y = 1$). We denote by $\sigma_{y,\hat{f}(\mathbf{x})}$ the covariance between the target variable and the classification output.

Table A2: Examples of XPER values decomposition

Panel A: Regression models

Metrics	$\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0))$	ϕ_0	ϕ_j
MSE	$2\sum_{j=1}^{q}\hat{\beta}_j\sigma_{y,x_j} - \sum_{j=1}^{q}\hat{\beta}_j^2\sigma^2_{x_j} - \sigma^2_y$	$-\sum_{j=1}^{q}\hat{\beta}_j^2\sigma^2_{x_j} - \sigma^2_y$	$2\hat{\beta}_j\sigma_{y,x_j}$
R^2	$\frac{\sigma_{y,\hat{y}}}{\sigma^2_y}$	$-\frac{\sigma_{y,\hat{y}}}{\sigma^2_y}$	$\frac{2\hat{\beta}_j\sigma_{y,x_j}}{\sigma^2_y}$

Panel B: Classification models

Metrics	$\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0))$	ϕ_0	ϕ_j
Accuracy	$2\sigma_{y,\hat{f}(\mathbf{x})} + 2\mathbb{P}(y=1)\hat{P}(\mathbf{x}) + 1 - \mathbb{P}(y=1) - \hat{P}(\mathbf{x})$	$2\mathbb{P}(y=1)\hat{P}(\mathbf{x}) + 1 - \mathbb{P}(y=1) - \hat{P}(\mathbf{x})$	No closed-form
Precision	$\frac{\sigma_{y,\hat{f}(\mathbf{x})}}{\mathbb{P}(\hat{f}(\mathbf{x})=1)} + \mathbb{P}(y=1)$	$\mathbb{P}(y=1)$	No closed-form
Sensitivity	$\frac{\sigma_{y,\hat{f}(\mathbf{x})}}{\mathbb{P}(y=1)} + \mathbb{P}(\hat{f}(\mathbf{x})=1)$	$\mathbb{P}(\hat{f}(\mathbf{x})=1)$	No closed-form
Specificity	$\frac{\sigma_{y,\hat{f}(\mathbf{x})}}{\mathbb{P}(y=0)} + \mathbb{P}(\hat{f}(\mathbf{x})=0)$	$\mathbb{P}(\hat{f}(\mathbf{x})=0)$	No closed-form
AUC	No closed-form	0.5	No closed-form

Note: This table displays the expression of population performance metrics $\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0))$, benchmark values ϕ_0, and XPER values ϕ_j. We distinguish between the performance metric associated to regression models (Panel A) and those related to classification models (Panel B). See Appendix E.3, E.4, and E.5 for the proofs associated to the MSE, the R^2 and the accuracy.

C Estimation

C.1 With a small number of features

In this section, we discuss the estimation procedure for the XPER values ϕ_j and $\phi_{i,j}$, based on a test sample $S_n = \{\mathbf{x}_i, y_i, \hat{f}(\mathbf{x}_i)\}_{i=1}^n$. Below, we assume that the model has a small number of features, typically $q \leq 10$.[1] When the number of features is larger, an approximation of the XPER values is required and we propose a modified version of the Kernel SHAP method of Lundberg and Lee (2017) (see Appendix C.2 for more details).

Under Assumption 1, the XPER value ϕ_j can be estimated by a weighted average of individual contributions differences such as:

$$\hat{\phi}_j = \sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\} \setminus \{x_j\})} w_S \left[\frac{1}{n^2} \sum_{u=1}^{n} \sum_{v=1}^{n} G\left(y_v; x_{v,j}, \mathbf{x}_v^S, \mathbf{x}_u^{\overline{S}}; \hat{\delta}_n\right) - \frac{1}{n^2} \sum_{u=1}^{n} \sum_{v=1}^{n} G\left(y_v; x_{u,j}, \mathbf{x}_v^S, \mathbf{x}_u^{\overline{S}}; \hat{\delta}_n\right) \right], \tag{A1}$$

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

Table A3 illustrates the computation of the estimated value $\hat{\phi}_1$ associated to feature x_1 in a model with three features (x_1, x_2, x_3). For each coalition of features (column 1), we report the corresponding weight (column 2) along with the estimated marginal contribution of feature x_1 to the performance metric (column 3). The intuition is as follows: the sum over index u refers to the marginalisation effect, whereas the sum over index v refers to the averaging effect. For a given instance v, we compute its average performance metric by replacing the features $x^{\overline{S}}$ which are *not included* in the coalition by the corresponding values observed for all the instances of the test sample (marginalisation). Then, we compute the average performance metric for all the instances v (averaging).

[1] The standard estimation framework is only feasible for a small number of features q. Indeed, the computation of the XPER values ϕ_j according to definition 3 becomes cumbersome as the number of features increases. Indeed, it requires to consider $q \times 2^{q-1}$ coalitions of features, e.g., $5,120$ for $q = 10$ and $10,485,760$ for $q = 20$, etc.

Table A3: Computation of the XPER value $\hat{\phi}_1$ in a three-feature model

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

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

Similarly to the estimation of features ϕ_j, we can estimate the *individual* XPER values $\phi_{i,j}$. For any individual i of the test sample of S_n and any feature x_j, a consistent estimator of the individual XPER values $\phi_{i,j}$ is defined as:

$$\hat{\phi}_{i,j} = \sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\} \backslash \{x_j\})} w_S \left[\frac{1}{n} \sum_{u=1}^n G\left(y_i; x_{i,j}, \mathbf{x}_i^S, \mathbf{x}_u^{\overline{S}}; \hat{\delta}_n\right) - \frac{1}{n} \sum_{u=1}^n G\left(y_i; x_{u,j}, \mathbf{x}_i^S, \mathbf{x}_u^{\overline{S}}; \hat{\delta}_n\right) \right]. \tag{A2}$$

By definition, these individual XPER values satisfy:

$$\hat{\phi}_j = \frac{1}{n} \sum_{i=1}^n \hat{\phi}_{i,j}. \tag{A3}$$

C.2 With a large number of features

As shown in section 3, the XPER value ϕ_j relies on every possible coalition $S \subseteq \mathcal{P}(\{\mathbf{x}\} \backslash \{x_j\})$ for each feature. The total number of coalitions required to compute $(\phi_1, \ldots, \phi_q)$ is equal to $q \times 2^{(q-1)}$ coalitions, with $2^{(q-1)}$ the number of coalitions for feature j. As the total number of coalitions quickly increases with the number of features ($5,120$ for $q = 10$ and $10,485,760$ for $q = 20$), XPER values turn out to be cumbersome to compute or estimate in practice. This issue is related to the general concept of Shapley value and is not specific to our approach.

To overcome this issue, a standard approach in the literature is to approximate the exact values. Intuitively, the idea is to rely only on a subset of coalitions to compute XPER values rather than on its total number in order to reduce the number of computations required. To do so, we propose an approximation method of XPER values which is based on the recent model-agnostic approach of Lundberg and Lee (2017) called Kernel SHAP. The main advantage of this approach is that it can be used for any application as it does not depend on the model $f(.)$ nor on the performance metric $G_n(\mathbf{y};\mathbf{X})$. Denote as S_k a coalition randomly drawn for $k=1,\ldots,K$, $K<q\times 2^{(q-1)}$ the total number of coalitions drawn, and $\tilde{S}_k=\{\mathbf{x}^{\bar{S}_k}\}\cup\{\mathbf{x}_j\}$ the vector of features not included in coalition S_k. Formally, the approximation of XPER values $\hat{\phi}_j$ is based on a subset K of coalitions S_k and on the following model

$$G_{n,k}(\mathbf{y};\mathbf{X})=\phi_0+\sum_{j=1}^{q}\phi_j z_{k,j}, \tag{A4}$$

where

$$G_{n,k}(\mathbf{y};\mathbf{X})=\frac{1}{n^2}\sum_{u=1}^{n}\sum_{v=1}^{n}G\left(y_v;\mathbf{x}_v^{S_k},\mathbf{x}_u^{\overline{S}_k};\hat{\delta}_n\right),$$

is a sample performance metric associated to the coalition S_k, and $z_{k,j}$ is equal to 1 if the feature j belongs to the coalition S_k, and 0 otherwise. In matrix notation, we have

$$\mathbf{G_n}(\mathbf{y};\mathbf{X})=\mathbf{Z}\boldsymbol{\phi}$$

$$\underset{(K\times 1)}{\mathbf{G_n}(\mathbf{y};\mathbf{X})}=\begin{pmatrix}G_{n,1}(\mathbf{y};\mathbf{X})\\ \vdots\\ G_{n,K}(\mathbf{y};\mathbf{X})\end{pmatrix},\ \underset{(K\times q+1)}{\mathbf{Z}}=\begin{pmatrix}1 & z_{1,1} & .. & z_{1,q}\\ \vdots & \vdots & \ddots & \vdots\\ 1 & z_{K,1} & .. & z_{K,q}\end{pmatrix},\ \underset{(q+1\times 1)}{\boldsymbol{\phi}}=\begin{pmatrix}\phi_0\\ \phi_1\\ \vdots\\ \phi_q\end{pmatrix}.$$

This approach relies on the comparison of K values of the performance metric obtained for K randomly drawn coalitions involving different subset of features. The features being present in some coalitions and absent in others, the estimation of this regression allows to approximate the impact of each feature on the performance metric that would be obtained for the subset k of features, and thus to approximate estimated XPER values $\hat{\phi}_j$. Moreover, contrary to Equation A1 this approach estimates the impact of every feature all at once, similarly to parameters' estimation in a traditional ordinary least squares regression, and is much faster as it only involves the calculation of K coalitions rather than $q \times 2^{(q-1)}$.

Recall that in Equation A1, the sum of marginal contribution to the performance metric is weighted as some coalitions are more informative than others. To this end, Lundberg and Lee (2017) propose the Kernel Shap which is a function that attributes a higher weight to the most informative coalitions, and which also takes into account the fact that some coalitions are drawn while others are not. The Kernel Shap function attributes a weight ω_{S_k} to coalition S_k such as:

$$\omega_{S_k} = \frac{(q-1)}{\frac{q!}{|S_k|!(q-|S_k|)!} |S_k| \left(q - |S_k|\right)}, \tag{A5}$$

where $|S_k|$ is the number of features included in the coalition S^k. To take into account these weights, Lundberg and Lee (2017) propose to estimate Equation A4 by Weighted Least Squares (WLS).[2] In practice, a relatively large value of K allows to obtain accurate approximations of $\hat{\phi}_j$.

Finally, this approach can also be used to approximate individual contributions to the estimated XPER values based on the following model:

$$G_k\left(y_i; \mathbf{x}_i\right) = \phi_{i,0} + \sum_{j=1}^{q} \phi_{i,j} z_{k,j}, \tag{A6}$$

where

$$G_k\left(y_i; \mathbf{x}_i\right) = \frac{1}{n} \sum_{u=1}^{n} G\left(y_i; \mathbf{x}_i^{S_k}, \mathbf{x}_u^{\tilde{S}_k}; \hat{\delta}_n\right).$$

Equation A6 is then estimated for each observation i of the sample, and the weight w_{S_k} remains the same as defined previously.

[2] Note that $S_k = \{\mathbf{x}\}$ and $S_k = \{\emptyset\}$ are not considered in the set of randomly individual coalitions as they lead to infinite weights. The K coalitions are thus drawn from $2^q - 2$ coalitions. See Lundberg and Lee (2017) for more details on the approximation approach of Shapley values.

D XPER vs. SHAP

In this section, we compare the XPER method with the Shapley additive explanation (SHAP) method of Lundberg and Lee (2017). As SHAP has now become ubiquitous in machine learning, we believe it is important to clearly show the added value of XPER over SHAP. In the latter, the contribution of a feature x_j to the predicted value $\hat{f}(\mathbf{x}_i)$ for individual i, denoted $\phi_{i,j}^{SHAP}$, is defined as:

$$\phi_{i,j}^{SHAP} = \sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\} \backslash \{x_j\})} w_S \left[\mathbb{E}_{\mathbf{x}^{\overline{S}}} \left(\hat{f}(x_{i,j}, \mathbf{x}_i^S, \mathbf{x}^{\overline{S}}) \right) - \mathbb{E}_{x_j, \mathbf{x}^{\overline{S}}} \left(\hat{f}(x_j, \mathbf{x}_i^S, \mathbf{x}^{\overline{S}}) \right) \right], \tag{A7}$$

$$\hat{f}(\mathbf{x}_i) = \phi_{i,0}^{SHAP} + \sum_{j=1}^{q} \phi_{i,j}^{SHAP}, \tag{A8}$$

$$\phi_{i,0}^{SHAP} = \mathbb{E} \left(\hat{f}(\mathbf{x}_i) \right). \tag{A9}$$

Proposition 1. *SHAP is a particular case of XPER where the individual contribution to the performance metric is equal to the predicted value of the model,* $G(y_i; \mathbf{x}_i; \delta_0) = \hat{f}(\mathbf{x}_i)$*.*

See Appendix E.7 for the proof. As stated in Proposition 1, we can show that SHAP is a particular case of XPER where the performance metric does not take into account the target variable. However, as performance metrics generally include at least the target variable to compare the predictions to their true value, SHAP and individual XPER values will differ in most cases. However, one may still wonder whether SHAP and individual XPER values provide the same information. If this were to be true, we should find that for a given feature x_j (1) a positive (negative) SHAP value means that this feature has a positive (negative) effect on the performance, and (2) a large SHAP value implies that this feature has a strong impact on performance. Below, we show that none of these statements are true. Indeed, in the former case, depending on the value of the target variable, a positive XPER value is not necessarily associated to a positive SHAP value. Intuitively, if the target variable is positive, a variable can contribute to increase the predicted value of the model ($\phi_{i,j}^{SHAP} > 0$) which can reduce the spread between the predicted value and the target value ($\phi_{i,j} > 0$). However, if the target variable is negative, a variable which contributes to decrease the predicted value of the model ($\phi_{i,j}^{SHAP} < 0$) can also reduce the prediction error of the model ($\phi_{i,j} > 0$). For instance, when the model only includes one variable, if the performance metric is defined as $G(y_i; \hat{f}(\mathbf{x}_i); \delta_0) = -(y_i - \hat{f}(\mathbf{x}_i))^2$, and if we assume that $\mathbb{E}(\hat{f}(x_1)) = 0$, we can show that:

$$\phi_{i,1} = 2\hat{\varepsilon}_i \phi_{i,1}^{SHAP} + \left(\phi_{i,1}^{SHAP}\right)^2 + \mathbb{V}\left(\hat{f}(x_{i,1})\right), \tag{A10}$$

where $\hat{\varepsilon}_i = y_i - \hat{f}(x_{i,1})$ is the prediction error of the model for individual i. See Appendix E.6 for the proof. As we can see, a positive XPER value can be associated to either a positive or negative SHAP value. If the prediction error and the SHAP value are both positive (negative), it means that this variable contributes to reduce the spread between the predicted value and the target value, which results in a positive XPER value. Therefore, a positive and a negative SHAP value can both lead to a positive XPER value. Moreover, as the individual performance and the predictions of the model does not have the same domain, except in the case where $G(y_i; \mathbf{x}_i; \delta_0) = \hat{f}(\mathbf{x}_i)$, the magnitude of SHAP and XPER values can be very different from one to another.

Although designed at the individual level, the SHAP method is also used to assess the effect of the features at the global level by taking the mean of the absolute SHAP values for each feature. We have:

$$\phi_j^{SHAP} = \mathbb{E}\left(\left|\phi_{i,j}^{SHAP}\right|\right), \tag{A11}$$

where ϕ_j^{SHAP} refers to the Mean Absolute SHAP values of feature j. This raises again the question of how similar these values are from the XPER values. One major difference between

SHAP and XPER is that SHAP values cannot be negative at the global level. This difference turns out to be important as negative XPER values allow to red-flag features that deteriorate the performance of the model on a given sample. For instance, this could help to explain the origin of the overfitting of a model as mentioned in Section 2. Note that according to Equation A8 and A9, it would not be appropriate to take the mean of SHAP values without taking the absolute value, as we would end up decomposing a zero:

$$\mathbb{E}\left(\hat{f}(\mathbf{x}_i)\right) = \phi_{i,0}^{SHAP} + \sum_{j=1}^{q} \mathbb{E}\left(\phi_{i,j}^{SHAP}\right),$$
$$\Leftrightarrow \sum_{j=1}^{q} \mathbb{E}\left(\phi_{i,j}^{SHAP}\right) = \mathbb{E}\left(\hat{f}(\mathbf{x}_i)\right) - \mathbb{E}\left(\hat{f}(\mathbf{x}_i)\right) = 0. \quad \text{(A12)}$$

Moreover, as both methods provide different results at the individual level, there is no reason to think that they would be equivalent at the global level. In the empirical section in Section 5, we confirm that the differences between XPER and SHAP can be substantial in practice.

E Proofs

E.1 Proof of Equation 6

Lemma 1. *The sum of the weights across all coalitions S is equal to 1,* $\sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\} \setminus \{x_j\})} \omega_S = 1.$

Proof. According to the definition of ω_S in Equation 4 and knowing that $\mathcal{P}(\{\mathbf{x}\} \setminus \{x_j\}) = \bigcup_{k=0}^{q-1} \mathcal{P}_k(\{\mathbf{x}\} \setminus \{x_j\})$, where $\mathcal{P}_k(\{\mathbf{x}\} \setminus \{x_j\})$ refers to the collection of all subsets of size k that can be formed from the powerset $\mathcal{P}(\{\mathbf{x}\} \setminus \{x_j\})$, we have:

$$\sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\} \setminus \{x_j\})} \omega_S = \sum_{S \subseteq \bigcup_{k=0}^{q-1} \mathcal{P}_k(\{\mathbf{x}\} \setminus \{x_j\})} \frac{1}{q \times C_{q-1}^{|S|}}.$$

with $C_{q-1}^{|S|}$ the number of $|S|$-combinations of a set with $q-1$ elements. As $\mathcal{P}_k(\{\mathbf{x}\} \setminus \{x_j\}) \cap$

$\mathcal{P}_l(\{\mathbf{x}\} \setminus \{x_j\}) = \emptyset,\ \forall k \neq l$ we derive that:

$$\sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\} \setminus \{x_j\})} \omega_S = \sum_{k=0}^{q-1} \sum_{S \subseteq \mathcal{P}_k(\{\mathbf{x}\} \setminus \{x_j\})} \frac{1}{q \times C_{q-1}^{|S|}}.$$

For each k, $\mathcal{P}_k(\{\mathbf{x}\} \setminus \{x_j\})$ is composed of C_{q-1}^k subsets of size k. As $S \subseteq \mathcal{P}_k(\{\mathbf{x}\} \setminus \{x_j\}$, we know that $|S| = k$, which implies that $C_{q-1}^{|S|} = C_{q-1}^k$. Thus,

$$\sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\} \setminus \{x_j\})} \omega_S = \sum_{k=0}^{q-1} C_{q-1}^k \frac{1}{q \times C_{q-1}^k} = \sum_{k=0}^{q-1} \frac{1}{q} = 1.$$

□

Lemma 2. *Consider a linear regression model $\hat{f}(\mathbf{x}) = \sum_{j=1}^q \hat{\beta}_j x_j$ where we assume that the DGP of the test sample $S_n = \{\mathbf{x}_i, y_i, \hat{f}(\mathbf{x}_i)\}_{i=1}^n$ satisfies $\mathbb{E}(\mathbf{x}) = \mu_q$ and $\mathbb{V}(\mathbf{x}) = \Sigma$ a positive semi-definite matrix, with $\Sigma_{k,j} = \sigma_{x_k,x_j}$ the covariance between feature x_k and x_j. The individual contribution to the R^2, for a coalition $S \subseteq \mathcal{P}(\{\mathbf{x}\} \setminus \{x_j\})$, can be expressed as:*

$$\begin{aligned}
G(y;\mathbf{x}) = 1 - \sigma_y^{-2} & \left[y^2 + \sum_{\substack{l=1 \\ l \in S}}^{q} x_l^2 \hat{\beta}_l^2 + \sum_{\substack{k=1 \\ k \in \overline{S}}}^{q} x_k^2 \hat{\beta}_k^2 + x_j^2 \hat{\beta}_j^2 + 2 \sum_{\substack{1 \leq k < l \leq q \\ k,l \in S}} \hat{\beta}_k \hat{\beta}_l x_k x_l + 2 \sum_{\substack{1 \leq k < l \leq q \\ k,l \in \overline{S}}} \hat{\beta}_k \hat{\beta}_l x_k x_l \right] \\
& - \sigma_y^{-2} \left[2 \sum_{\substack{k=1 \\ k \in S}}^{q} \hat{\beta}_k \hat{\beta}_j x_k x_j + 2 \sum_{\substack{k=1 \\ k \in \overline{S}}}^{q} \hat{\beta}_k \hat{\beta}_j x_k x_j + 2 \sum_{\substack{k=1 \\ k \in S}}^{q} \sum_{\substack{l=1 \\ l \in \overline{S}}}^{q} \hat{\beta}_k \hat{\beta}_l x_k x_l \right] \\
& - \sigma_y^{-2} \left[-2 \sum_{\substack{k=1 \\ k \in S}}^{q} y x_k \hat{\beta}_k - 2 \sum_{\substack{k=1 \\ k \in \overline{S}}}^{q} y x_k \hat{\beta}_k - 2 y x_j \hat{\beta}_j \right].
\end{aligned}$$

Proof. Consider a linear regression model $\hat{f}(\mathbf{x}) = \sum_{j=1}^q \hat{\beta}_j x_j$ where we assume that the DGP of the test sample $S_n = \{\mathbf{x}_i, y_i, \hat{f}(\mathbf{x}_i)\}_{i=1}^n$ satisfies $\mathbb{E}(\mathbf{x}) = \mu_q$ and $\mathbb{V}(\mathbf{x}) = \Sigma$ a positive semi-definite matrix, with $\Sigma_{k,j} = \sigma_{x_k,x_j}$ the covariance between feature x_k and x_j.

Reminds that for a linear regression model $\hat{f}(\mathbf{x}) = \sum_{j=1}^q \hat{\beta}_j x_j$, the individual contribution to the R^2 is defined as (see Equation 2):

$$G(y;\mathbf{x}) = 1 - \frac{(y-\mathbf{x}\hat{\beta})^2}{\sigma_y^2}$$
$$= 1 - \sigma_y^{-2}\left[y^2 + \sum_{k=1}^{q} x_k^2\hat{\beta}_k^2 + 2\sum_{1\le k<l\le q}\hat{\beta}_k\hat{\beta}_l x_k x_l - 2\sum_{k=1}^{q} y x_k \hat{\beta}_k\right]. \quad \text{(A13)}$$

Considering a coalition $S \subseteq \mathcal{P}(\{\mathbf{x}\}\setminus\{x_j\})$ of features, the vector of features $\mathbf{x}$ is composed of three sub-vectors: $\mathbf{x}^S$ the vector of features in the coalition S, $\mathbf{x}^{\overline{S}}$ the vector of features apart from the coalition, and x_j the remaining feature of interest, such that $\mathbf{x} = (\mathbf{x}^S, \mathbf{x}^{\overline{S}}, x_j)$. Therefore, we can rewrite Equation A13 as:

$$G(y;\mathbf{x}) = 1 - \sigma_y^{-2}\left[y^2 + \sum_{\substack{l=1\\ l\in S}}^{q} x_l^2\hat{\beta}_l^2 + \sum_{\substack{k=1\\ k\in\overline{S}}}^{q} x_k^2\hat{\beta}_k^2 + x_j^2\hat{\beta}_j^2 + 2\sum_{\substack{1\le k<l\le q\\ k,l\in S}}\hat{\beta}_k\hat{\beta}_l x_k x_l + 2\sum_{\substack{1\le k<l\le q\\ k,l\in\overline{S}}}\hat{\beta}_k\hat{\beta}_l x_k x_l\right]$$
$$- \sigma_y^{-2}\left[2\sum_{\substack{k=1\\ k\in S}}^{q}\hat{\beta}_k\hat{\beta}_j x_k x_j + 2\sum_{\substack{k=1\\ k\in\overline{S}}}^{q}\hat{\beta}_k\hat{\beta}_j x_k x_j + 2\sum_{\substack{k=1\\ k\in S}}^{q}\sum_{\substack{l=1\\ l\in\overline{S}}}^{q}\hat{\beta}_k\hat{\beta}_l x_k x_l\right]$$
$$- \sigma_y^{-2}\left[-2\sum_{\substack{k=1\\ k\in S}}^{q} y x_k\hat{\beta}_k - 2\sum_{\substack{k=1\\ k\in\overline{S}}}^{q} y x_k\hat{\beta}_k - 2y x_j\hat{\beta}_j\right].$$

□

Lemma 3. *For a given set of features $\{\mathbf{x}\}\setminus\{x_j\}$, we have:*

$$2\sum_{S\subseteq\mathcal{P}(\{\mathbf{x}\}\setminus\{x_j\})}\omega_S\left(\sum_{\substack{k=1\\ k\in S}}^{q} h(x_k) + \sum_{\substack{k=1\\ k\in\overline{S}}}^{q} g(x_k)\right) = \sum_{\substack{k=1\\ k\neq j}}\left(h(x_k)+g(x_k)\right),$$

with $h(.)$ and $g(.)$ some unknown linear or non-linear functions.

Proof. Let consider a quantity A defined as follows:

$$A = 2\sum_{S\subseteq\mathcal{P}(\{\mathbf{x}\}\setminus\{x_j\})}\omega_S\left(\sum_{\substack{k=1\\ k\in S}}^{q} h(x_k) + \sum_{\substack{k=1\\ k\in\overline{S}}}^{q} g(x_k)\right), \quad \text{(A14)}$$

with $h(.)$ and $g(.)$ some unknown functions. As $\mathcal{P}(\{\mathbf{x}\} \setminus \{x_j\}) = \bigcup_{l=0}^{q-1} \mathcal{P}_l(\{\mathbf{x}\} \setminus \{x_j\})$, where $\mathcal{P}_l(\{\mathbf{x}\} \setminus \{x_j\})$ refers to the collection of all subsets of size l that can be formed from the powerset $\mathcal{P}(\{\mathbf{x}\} \setminus \{x_j\})$, we obtain from Equation A14:

$$A = 2 \sum_{S \subseteq \bigcup_{l=0}^{q-1} \mathcal{P}_l(\{\mathbf{x}\} \setminus \{x_j\})} \omega_S \left(\sum_{\substack{k=1 \\ k \in S}}^{q} h(x_k) + \sum_{\substack{k=1 \\ k \in \overline{S}}}^{q} g(x_k) \right). \tag{A15}$$

Note that for an even number of features q, we have:

$$\bigcup_{l=0}^{q-1} \mathcal{P}_l(\{\mathbf{x}\} \setminus \{x_j\}) = \bigcup_{l=0}^{(q-2)/2} \mathcal{P}_{q-1-l}(\{\mathbf{x}\} \setminus \{x_j\}) \cup \mathcal{P}_l(\{\mathbf{x}\} \setminus \{x_j\}), \tag{A16}$$

whereas for an odd number of features:

$$\bigcup_{l=0}^{q-1} \mathcal{P}_l(\{\mathbf{x}\} \setminus \{x_j\}) = \bigcup_{l=0}^{(q-1)/2} \mathcal{P}_{q-1-l}(\{\mathbf{x}\} \setminus \{x_j\}) \cup \mathcal{P}_l(\{\mathbf{x}\} \setminus \{x_j\}). \tag{A17}$$

Therefore, we now distinguish between the two cases to complete the proof.

When q is even, if we plug the expression of $\bigcup_{l=0}^{q-1} \mathcal{P}_l(\{\mathbf{x}\} \setminus \{x_j\})$ given in Equation A16 into Equation A15, then Equation A15 is equivalent to:

$$A = 2 \sum_{S \subseteq \bigcup_{l=0}^{(q-2)/2} \mathcal{P}_{q-1-l}(\{\mathbf{x}\} \setminus \{x_j\}) \cup \mathcal{P}_l(\{\mathbf{x}\} \setminus \{x_j\})} \omega_S \left(\sum_{\substack{k=1 \\ k \in S}}^{q} h(x_k) + \sum_{\substack{k=1 \\ k \in \overline{S}}}^{q} g(x_k) \right). \tag{A18}$$

As $\bigcap_{l=0}^{(q-2)/2} \mathcal{P}_{q-1-l}(\{\mathbf{x}\} \setminus \{x_j\}) \cup \mathcal{P}_l(\{\mathbf{x}\} \setminus \{x_j\}) = \emptyset$, we can rewrite Equation A18 as follows:

$$A = 2 \sum_{l=0}^{(q-2)/2} \sum_{S \subseteq \mathcal{P}_{q-1-l}(\{\mathbf{x}\} \setminus \{x_j\}) \cup \mathcal{P}_l(\{\mathbf{x}\} \setminus \{x_j\})} \omega_S \left(\sum_{\substack{k=1 \\ k \in S}}^{q} h(x_k) + \sum_{\substack{k=1 \\ k \in \overline{S}}}^{q} g(x_k) \right). \tag{A19}$$

Note that each coalition $S \subseteq \mathcal{P}_{q-1-l}(\{\mathbf{x}\} \setminus \{x_j\}) \cup \mathcal{P}_l(\{\mathbf{x}\} \setminus \{x_j\})$ is either composed of $|S| = q-1-l$ or $|S| = l$ elements obtained from the set of features $\{\mathbf{x}\} \setminus \{x_j\}$ of size $q-1$. Therefore, according to Lemma 4, all of the coalitions $S \subseteq \mathcal{P}_{q-1-l}(\{\mathbf{x}\} \setminus \{x_j\}) \cup \mathcal{P}_l(\{\mathbf{x}\} \setminus \{x_j\})$

have the same weight. We refer to this weight as ω_l. Thus, Equation A19 simplifies to:

$$A = 2 \sum_{l=0}^{(q-2)/2} \omega_l \sum_{S \subseteq \mathcal{P}_{q-1-l}(\{\mathbf{x}\} \setminus \{x_j\}) \cup \mathcal{P}_l(\{\mathbf{x}\} \setminus \{x_j\})} \left(\sum_{\substack{k=1 \\ k \in S}}^{q} h(x_k) + \sum_{\substack{k=1 \\ k \in \overline{S}}}^{q} g(x_k) \right). \tag{A20}$$

By construction, each feature $x_k \in \{\mathbf{x}\} \setminus \{x_j\}$ is included in half of the coalitions $S \subseteq \mathcal{P}_{q-1-l}(\{\mathbf{x}\} \setminus \{x_j\}) \cup \mathcal{P}_l(\{\mathbf{x}\} \setminus \{x_j\})$. Note that the set $\mathcal{P}_{q-1-l}(\{\mathbf{x}\} \setminus \{x_j\}) \cup \mathcal{P}_l(\{\mathbf{x}\} \setminus \{x_j\})$ is composed of $2 \times C_{q-1}^{q-1-l} = 2 \times C_{q-1}^{l}$ coalitions. Therefore, each feature x_k is included in C_{q-1}^{l} coalitions. Thus, we can write that:

$$\sum_{S \subseteq \mathcal{P}_{q-1-l}(\{\mathbf{x}\} \setminus \{x_j\}) \cup \mathcal{P}_l(\{\mathbf{x}\} \setminus \{x_j\})} \sum_{\substack{k=1 \\ k \in S}}^{q} h(x_k) = \sum_{\substack{k=1 \\ k \neq j}}^{q} C_{q-1}^{l} h(x_k). \tag{A21}$$

As each feature $x_k \in \{\mathbf{x}\} \setminus \{x_j\}$ is included in half of the coalitions $S \subseteq \mathcal{P}_{q-1-l}(\{\mathbf{x}\} \setminus \{x_j\}) \cup \mathcal{P}_l(\{\mathbf{x}\} \setminus \{x_j\})$, each of them is excluded from the coalitions S half of the time. Therefore, each feature x_k is included in C_{q-1}^{l} coalitions $\overline{S} \subseteq \mathcal{P}_{q-1-l}(\{\mathbf{x}\} \setminus \{x_j\}) \cup \mathcal{P}_l(\{\mathbf{x}\} \setminus \{x_j\})$, such that:

$$\sum_{S \subseteq \mathcal{P}_{q-1-l}(\{\mathbf{x}\} \setminus \{x_j\}) \cup \mathcal{P}_l(\{\mathbf{x}\} \setminus \{x_j\})} \sum_{\substack{k=1 \\ k \in \overline{S}}}^{q} g(x_k) = \sum_{\substack{k=1 \\ k \neq j}}^{q} C_{q-1}^{l} g(x_k). \tag{A22}$$

As a consequence, Equation A19 simplifies to:

$$A = 2 \sum_{l=0}^{(q-2)/2} \omega_l C_{q-1}^{l} \left(\sum_{\substack{k=1 \\ k \neq j}} \left(h(x_k) + g(x_k) \right) \right). \tag{A23}$$

Moreover, according to the definition of ω_l in Equation 4, we then have:

$$A = 2 \sum_{l=0}^{(q-2)/2} \frac{1}{q} \left(\sum_{\substack{k=1 \\ k \neq j}} \left(h(x_k) + g(x_k) \right) \right). \tag{A24}$$

Finally, for an even number of features q, we obtain:

$$2 \sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\} \setminus \{x_j\})} \omega_S \left(\sum_{\substack{k=1 \\ k \in S}}^{q} h(x_k) + \sum_{\substack{k=1 \\ k \in \overline{S}}}^{q} g(x_k) \right) = \sum_{\substack{k=1 \\ k \neq j}} \left(h(x_k) + g(x_k) \right). \tag{A25}$$

Similarly, when q is odd, we can write that:

$$A = 2 \sum_{l=0}^{(q-1)/2} \omega_l \sum_{S \subseteq \mathcal{P}_{q-1-l}(\{\mathbf{x}\} \setminus \{x_j\}) \cup \mathcal{P}_l(\{\mathbf{x}\} \setminus \{x_j\})} \left(\sum_{\substack{k=1 \\ k \in S}}^{q} h(x_k) + \sum_{\substack{k=1 \\ k \in \overline{S}}}^{q} g(x_k) \right). \tag{A26}$$

However, when $l = (q-1)/2$, the set $\mathcal{P}_{q-1-l}(\{\mathbf{x}\} \setminus \{x_j\}) \cup \mathcal{P}_l(\{\mathbf{x}\} \setminus \{x_j\})$ is only composed of $C_{q-1}^{q-1-l} = C_{q-1}^{l}$ coalitions as $\mathcal{P}_{(q-1)/2}(\{\mathbf{x}\} \setminus \{x_j\}) \cup \mathcal{P}_{(q-1)/2}(\{\mathbf{x}\} \setminus \{x_j\}) = \mathcal{P}_{(q-1)/2}(\{\mathbf{x}\} \setminus \{x_j\})$. Therefore, for $l = (q-1)/2$, each feature $x_k \in S \subseteq \mathcal{P}_{q-1-l}(\{\mathbf{x}\} \setminus \{x_j\}) \cup \mathcal{P}_l(\{\mathbf{x}\} \setminus \{x_j\})$ is included in $C_{q-1}^{l}/2$ coalitions. To take into account this specificity, we rewrite Equation A26 as follows:

$$A = 2 \sum_{l=0}^{(q-3)/2} \omega_l \sum_{S \subseteq \mathcal{P}_{q-1-l}(\{\mathbf{x}\} \setminus \{x_j\}) \cup \mathcal{P}_l(\{\mathbf{x}\} \setminus \{x_j\})} \left(\sum_{\substack{k=1 \\ k \in S}}^{q} h(x_k) + \sum_{\substack{k=1 \\ k \in \overline{S}}}^{q} g(x_k) \right) \tag{A27}$$

$$- 2\omega_{(q-1)/2} \sum_{S \subseteq \mathcal{P}_{(q-1)/2}(\{\mathbf{x}\} \setminus \{x_j\})} \left(\sum_{\substack{k=1 \\ k \in S}}^{q} h(x_k) + \sum_{\substack{k=1 \\ k \in \overline{S}}}^{q} g(x_k) \right), \tag{A28}$$

which is equal to:

$$A = 2 \sum_{l=0}^{(q-3)/2} \omega_l C_{q-1}^{l} \left(\sum_{\substack{k=1 \\ k \neq j}} \left(h(x_k) + g(x_k) \right) \right) - 2w_{|(q-1)/2|} \frac{C_{q-1}^{(q-1)/2}}{2} \left(\sum_{\substack{k=1 \\ k \neq j}} \left(h(x_k) + g(x_k) \right) \right). \tag{A29}$$

According to the definition of ω_l in Equation 4, we then have:

$$A = 2 \sum_{l=0}^{(q-3)/2} \frac{1}{q} \left(\sum_{\substack{k=1 \\ k \neq j}} \left(h(x_k) + g(x_k) \right) \right) - \left(\sum_{\substack{k=1 \\ k \neq j}} \left(h(x_k) + g(x_k) \right) \right). \tag{A30}$$

Finally, for an odd number of features q, we obtain:

$$A = 2 \sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\} \backslash \{x_j\})} \omega_S \left(\sum_{\substack{k=1 \\ k \in S}}^{q} h(x_k) + \sum_{\substack{k=1 \\ k \in \overline{S}}}^{q} g(x_k) \right) = \sum_{\substack{k=1 \\ k \neq j}} \left(h(x_k) + g(x_k) \right). \tag{A31}$$

As we obtain the same expression for A with an odd and an even number of features q, we conclude that for all q:

$$2 \sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\} \backslash \{x_j\})} \omega_S \left(\sum_{\substack{k=1 \\ k \in S}}^{q} h(x_k) + \sum_{\substack{k=1 \\ k \in \overline{S}}}^{q} g(x_k) \right) = \sum_{\substack{k=1 \\ k \neq j}} \left(h(x_k) + g(x_k) \right), \tag{A32}$$

with $h(.)$ and $g(.)$ some unknown functions. □

Proposition 1. *Consider a linear regression model $\hat{f}(\mathbf{x}) = \sum_{j=1}^{q} \hat{\beta}_j x_j$ where we assume that the DGP of the test sample $S_n = \{\mathbf{x}_i, y_i, \hat{f}(\mathbf{x}_i)\}_{i=1}^{n}$ satisfies $\mathbb{E}(\mathbf{x}) = \mu_q$ and $\mathbb{V}(\mathbf{x}) = \Sigma$ a positive semi-definite matrix, with $\Sigma_{k,j} = \sigma_{x_k,x_j}$ the covariance between feature x_k and x_j. Then, the XPER contribution ϕ_j of feature x_j to the R^2 is:*

$$\phi_j = \frac{2\hat{\beta}_j \sigma_{y,x_j}}{\sigma_y^2}, \qquad \forall j = 1, ..., q, \tag{A33}$$

with σ_y^2 the variance of the target variable and σ_{y,x_j} its covariance with feature x_j.

Proof. Consider a linear regression model $\hat{f}(\mathbf{x}) = \sum_{j=1}^{q} \hat{\beta}_j x_j$ where we assume that the DGP of the test sample $S_n = \{\mathbf{x}_i, y_i, \hat{f}(\mathbf{x}_i)\}_{i=1}^{n}$ satisfies $\mathbb{E}(\mathbf{x}) = \mu_q$ and $\mathbb{V}(\mathbf{x}) = \Sigma$ a positive semi-definite matrix, with $\Sigma_{k,j} = \sigma_{x_k,x_j}$ the covariance between feature x_k and x_j.

From Lemma 2 we can derive that, for a coalition $S \subseteq \mathcal{P}(\{\mathbf{x}\} \setminus \{x_j\})$, we have:

$$\begin{aligned} \mathbb{E}_{y,\mathbf{x}^S,x_j} \mathbb{E}_{\mathbf{x}^{\overline{S}}} \left(G(y; \mathbf{x}; \delta_0) \right) - \mathbb{E}_{y,\mathbf{x}^S} \mathbb{E}_{\mathbf{x}^{\overline{S}},x_j} \left(G(y; \mathbf{x}; \delta_0) \right) = \sigma_y^{-2} & \left[-2 \sum_{\substack{k=1 \\ k \in S}}^{q} \hat{\beta}_k \hat{\beta}_j \sigma_{x_k,x_j} + 2 \sum_{\substack{k=1 \\ k \in \overline{S}}}^{q} \hat{\beta}_k \hat{\beta}_j \sigma_{x_k,x_j} \right] \\ & + \sigma_y^{-2} \left[2\hat{\beta}_j \sigma_{y,x_j} \right], \end{aligned} \tag{A34}$$

with σ_y^2 the variance of the target variable and σ_{y,x_j} its covariance with feature x_j. Thus, according to Definition 4 and Equation A34, the XPER contribution ϕ_j to the R^2 is equal to:

$$\begin{aligned}
\phi_j &= \sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\}\backslash\{x_j\})} \omega_S \Big(\mathbb{E}_{y,\mathbf{x}^S,x_j} \mathbb{E}_{\mathbf{x}^{\overline{S}}} \left(G(y;\mathbf{x};\delta_0) \right) - \mathbb{E}_{y,\mathbf{x}^S} \mathbb{E}_{\mathbf{x}^{\overline{S}},x_j} \left(G(y;\mathbf{x};\delta_0) \right) \Big) \\
&= \sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\}\backslash\{x_j\})} \omega_S \left(\sigma_y^{-2} \left[-2 \sum_{\substack{k=1 \\ k \in S}}^{q} \hat{\beta}_k \hat{\beta}_j \sigma_{x_k,x_j} + 2 \sum_{\substack{k=1 \\ k \in \overline{S}}}^{q} \hat{\beta}_k \hat{\beta}_j \sigma_{x_k,x_j} \right] \right) \\
&+ \sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\}\backslash\{x_j\})} \omega_S \sigma_y^{-2} \left[2 \hat{\beta}_j \sigma_{y,x_j} \right]. \qquad \text{(A35)}
\end{aligned}$$

As according to Lemma 1 we know that $\sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\}\backslash\{x_j\})} \omega_S = 1$, we obtain:

$$\begin{aligned}
\phi_j &= \sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\}\backslash\{x_j\})} \omega_S \left(\sigma_y^{-2} \left[-2 \sum_{\substack{k=1 \\ k \in S}}^{q} \hat{\beta}_k \hat{\beta}_j \sigma_{x_k,x_j} + 2 \sum_{\substack{k=1 \\ k \in \overline{S}}}^{q} \hat{\beta}_k \hat{\beta}_j \sigma_{x_k,x_j} \right] \right) \\
&+ \sigma_y^{-2} \left[2 \hat{\beta}_j \sigma_{y,x_j} \right]. \qquad \text{(A36)}
\end{aligned}$$

According to Lemma 3, for $h(x_k) = -\hat{\beta}_k \hat{\beta}_j \sigma_{x_k,x_j}$ and $g(x_k) = \hat{\beta}_k \hat{\beta}_j \sigma_{x_k,x_j}$ we obtain:

$$2 \sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\}\backslash\{x_j\})} \omega_S \left(\sum_{\substack{k=1 \\ k \in S}}^{q} -\hat{\beta}_k \hat{\beta}_j \sigma_{x_k,x_j} + \sum_{\substack{k=1 \\ k \in \overline{S}}}^{q} \hat{\beta}_k \hat{\beta}_j \sigma_{x_k,x_j} \right) = \sum_{\substack{k=1 \\ k \neq j}} \Big(-\hat{\beta}_k \hat{\beta}_j \sigma_{x_k,x_j} + \hat{\beta}_k \hat{\beta}_j \sigma_{x_k,x_j} \Big) = 0 \qquad \text{(A37)}$$

Therefore, from Equation A36 and Equation A37 we deduce that the XPER contribution ϕ_j of feature x_j to the R^2 is:

$$\phi_j = \frac{2 \hat{\beta}_j \sigma_{y,x_j}}{\sigma_y^2}.$$

□

E.2 Proof of Equation 8

Lemma 4. *The weight associated to a coalition S built from a set of features of size $q-1$ is equal to the weight of the coalition $\tilde{S}$, where $|\tilde{S}| = q-1-|S|$, i.e., $\omega_S = \omega_{\tilde{S}}$.*

Proof. According to Equation 4, ω_S is defined as:

$$\omega_S = \frac{1}{q \times C_{q-1}^{|S|}} = \frac{1}{q \times \frac{(q-1)!}{|S|!(q-1-|S|)!}}.$$

Similarly, as $|\tilde{S}| = q-1-|S|$, $\omega_{\tilde{S}}$ is expressed as:

$$\omega_{\tilde{S}} = \frac{1}{q \times C_{q-1}^{|\tilde{S}|}} = \frac{1}{q \times C_{q-1}^{q-1-|S|}} = \frac{1}{q \times \frac{(q-1)!}{(q-1-|S|)!(q-1-(q-1-|S|))!}} = \frac{1}{q \times \frac{(q-1)!}{|S|!(q-1-|S|)!}} = \omega_S.$$

□

Proposition 2. *Consider a linear regression model $\hat{f}(\mathbf{x}) = \sum_{j=1}^{q} \hat{\beta}_j x_j$ where we assume that the DGP of the test sample $S_n = \{\mathbf{x}_i, y_i, \hat{f}(\mathbf{x}_i)\}_{i=1}^n$ satisfies $\mathbb{E}(\mathbf{x}) = \mu_q$ and $\mathbb{V}(\mathbf{x}) = \Sigma$ a positive semi-definite matrix, with $\Sigma_{k,j} = \sigma_{x_k,x_j}$ the covariance between feature x_k and x_j. The individual XPER contribution $\phi_{i,j}$ to the R^2 is:*

$$\phi_{i,j} = \sigma_y^{-2} \left[\hat{\beta}_j (x_{i,j} - \mathbb{E}(x_j)) A - \hat{\beta}_j^2 (x_{i,j}^2 - \mathbb{E}(x_j^2)) + \sum_{\substack{k=1 \\ k \neq j}}^{q} \hat{\beta}_k \hat{\beta}_j \sigma_{x_k,x_j} \right].$$

with $A = \left(2y_i - \sum_{\substack{k=1 \\ k \neq j}}^{q} \hat{\beta}_k (x_{i,k} + \mathbb{E}(x_k)) \right)$, σ_y^2 the variance of the target variable and σ_{x_k,x_j} the covariance between the feature x_k and x_j.

Proof. Consider a linear regression model $\hat{f}(\mathbf{x}) = \sum_{j=1}^{q} \hat{\beta}_j x_j$ where we assume that the DGP of the test sample $S_n = \{\mathbf{x}_i, y_i, \hat{f}(\mathbf{x}_i)\}_{i=1}^n$ satisfies $\mathbb{E}(\mathbf{x}) = \mu_q$ and $\mathbb{V}(\mathbf{x}) = \Sigma$ a positive semi-definite matrix, with $\Sigma_{k,j} = \sigma_{x_k,x_j}$ the covariance between feature x_k and x_j.

From Lemma 2, we can derive that for a coalition $S \subseteq \mathcal{P}(\{\mathbf{x}\} \setminus \{x_j\})$ and for an individual

i we have:

$$\mathbf{E}_{\mathbf{x}^{\overline{S}}}\left(G(y_i;\mathbf{x}_i;\delta_0)\right) - \mathbf{E}_{\mathbf{x}^{\overline{S}},x_j}\left(G(y_i;\mathbf{x}_i;\delta_0)\right) = \sigma_y^{-2}\left[2y_i\hat{\beta}_j(x_{i,j}-\mathbb{E}(x_j)) - \hat{\beta}_j^2(x_{i,j}^2-\mathbb{E}(x_j^2))\right] + \sigma_y^{-2}\left[2\sum_{\substack{k=1\\k\in\overline{S}}}\hat{\beta}_k\hat{\beta}_j\sigma_{x_k,x_j}\right] - \sigma_y^{-2}\left[2\hat{\beta}_j(x_{i,j}-\mathbb{E}(x_j))\left(\sum_{\substack{k=1\\k\in S}}\hat{\beta}_k x_{i,k} + \sum_{\substack{k=1\\k\in\overline{S}}}\hat{\beta}_k\mathbb{E}(x_k)\right)\right], \tag{A38}$$

with σ_y^2 the variance of the target variable and σ_{x_k,x_j} the covariance between the feature x_k and x_j. Thus, according to Definition 4 and Equation A38, the XPER contribution $\phi_{i,j}$ to the R^2 is equal to:

$$\begin{aligned}\phi_{i,j} &= \sum_{S\subseteq\mathcal{P}(\{\mathbf{x}\}\backslash\{x_j\})}\omega_S\left(\mathbf{E}_{\mathbf{x}^{\overline{S}}}\left(G(y_i;\mathbf{x}_i;\delta_0)\right) - \mathbf{E}_{\mathbf{x}^{\overline{S}},x_j}\left(G(y_i;\mathbf{x}_i;\delta_0)\right)\right)\\ &= \sum_{S\subseteq\mathcal{P}(\{\mathbf{x}\}\backslash\{x_j\})}\omega_S\sigma_y^{-2}\left[2y_i\hat{\beta}_j(x_{i,j}-\mathbb{E}(x_j)) - \hat{\beta}_j^2(x_{i,j}^2-\mathbb{E}(x_j^2)) + 2\sum_{\substack{k=1\\k\in\overline{S}}}\hat{\beta}_k\hat{\beta}_j\sigma_{x_k,x_j}\right]\\ &- \sum_{S\subseteq\mathcal{P}(\{\mathbf{x}\}\backslash\{x_j\})}\omega_S\sigma_y^{-2}\left[2\hat{\beta}_j(x_{i,j}-\mathbb{E}(x_j))\left(\sum_{\substack{k=1\\k\in S}}\hat{\beta}_k x_{i,k} + \sum_{\substack{k=1\\k\in\overline{S}}}\hat{\beta}_k\mathbb{E}(x_k)\right)\right].\end{aligned} \tag{A39}$$

As according to Lemma 1 we know that $\sum_{S\subseteq\mathcal{P}(\{\mathbf{x}\}\backslash\{x_j\})}\omega_S = 1$, we obtain:

$$\begin{aligned}\phi_{i,j} = \sigma_y^{-2}&\left[2y_i\hat{\beta}_j(x_{i,j}-\mathbb{E}(x_j)) - \hat{\beta}_j^2(x_{i,j}^2-\mathbb{E}(x_j^2)) + 2\sum_{S\subseteq\mathcal{P}(\{\mathbf{x}\}\backslash\{x_j\})}\omega_S\sum_{\substack{k=1\\k\in\overline{S}}}\hat{\beta}_k\hat{\beta}_j\sigma_{x_k,x_j}\right]\\ - \sigma_y^{-2}&\left[\hat{\beta}_j(x_{i,j}-\mathbb{E}(x_j))\times 2\sum_{S\subseteq\mathcal{P}(\{\mathbf{x}\}\backslash\{x_j\})}\omega_S\left(\sum_{\substack{k=1\\k\in S}}\hat{\beta}_k x_{i,k} + \sum_{\substack{k=1\\k\in\overline{S}}}\hat{\beta}_k\mathbb{E}(x_k)\right)\right].\end{aligned} \tag{A40}$$

According to Lemma 3, for $h(x_k) = 0$ and $g(x_k) = \hat{\beta}_k \hat{\beta}_j \sigma_{x_k,x_j}$ we find that:

$$2 \sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\} \setminus \{x_j\})} \omega_S \sum_{\substack{k=1 \\ k \in \overline{S}}} \hat{\beta}_k \hat{\beta}_j \sigma_{x_k,x_j} = \sum_{\substack{k=1 \\ k \neq j}}^{q} \hat{\beta}_k \hat{\beta}_j \sigma_{x_k,x_j} \tag{A41}$$

Similarly, for $h(x_k) = \hat{\beta}_k x_{i,k}$ and $g(x_k) = \hat{\beta}_k \mathbb{E}(x_k)$:

$$2 \sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\} \setminus \{x_j\})} \omega_S \left(\sum_{\substack{k=1 \\ k \in S}} \hat{\beta}_k x_{i,k} + \sum_{\substack{k=1 \\ k \in \overline{S}}} \hat{\beta}_k \mathbb{E}(x_{i,k}) \right) = \sum_{\substack{k=1 \\ k \neq j}}^{q} \hat{\beta}_k (x_{i,k} + \mathbb{E}(x_k)) \tag{A42}$$

From Equation A40, A41, A42, we obtain:

$$\begin{aligned} \phi_{i,j} = \sigma_y^{-2} & \left[2 y_i \hat{\beta}_j (x_{i,j} - \mathbb{E}(x_j)) - \hat{\beta}_j^2 (x_{i,j}^2 - \mathbb{E}(x_j^2)) + \sum_{\substack{k=1 \\ k \neq j}}^{q} \hat{\beta}_k \hat{\beta}_j \sigma_{x_k,x_j} \right] \\ - \sigma_y^{-2} & \left[\hat{\beta}_j (x_{i,j} - \mathbb{E}(x_j)) \sum_{\substack{k=1 \\ k \neq j}}^{q} \hat{\beta}_k (x_{i,k} + \mathbb{E}(x_k)) \right]. \end{aligned} \tag{A43}$$

Finally, rearranging the terms we obtain:

$$\phi_{i,j} = \sigma_y^{-2} \left[\hat{\beta}_j (x_{i,j} - \mathbb{E}(x_j)) A - \hat{\beta}_j^2 (x_{i,j}^2 - \mathbb{E}(x_j^2)) + \sum_{\substack{k=1 \\ k \neq j}}^{q} \hat{\beta}_k \hat{\beta}_j \sigma_{x_k,x_j} \right].$$

with $A = \left(2 y_i - \sum_{\substack{k=1 \\ k \neq j}}^{q} \hat{\beta}_k (x_{i,k} + \mathbb{E}(x_k)) \right)$, σ_y^2 the variance of the target variable and σ_{x_k,x_j} the covariance between the feature x_k and x_j. □

E.3 Proof of the MSE example in Table A1

Proposition 3. *Consider a linear regression model $\hat{f}(\mathbf{x}) = \sum_{j=1}^{q} \hat{\beta}_j x_j$, where $\mathbb{E}(\mathbf{x}) = 0_q$ and $\mathbb{V}(\mathbf{x}) = diag(\sigma_{x_j}^2), \forall j = 1, ..., q$, and $\mathbb{E}(y) = 0$. The contributions ϕ_j of features*

x_j to the (opposite of the) MSE satisfy the efficiency axiom such that:

$$\underbrace{2\sum_{j=1}^{q}\hat{\beta}_j\sigma_{y,x_j}-\sum_{j=1}^{q}\hat{\beta}_j^2\sigma_{x_j}^2-\sigma_y^2}_{\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0))}=\underbrace{-\sum_{j=1}^{q}\hat{\beta}_j^2\sigma_{x_j}^2-\sigma_y^2}_{\phi_0}+\sum_{j=1}^{q}\underbrace{2\hat{\beta}_j\sigma_{y,x_j}}_{\phi_j}. \tag{A44}$$

Proof. Consider a linear regression model $\hat{f}(\mathbf{x})=\sum_{j=1}^{q}\hat{\beta}_j x_j$, where $\mathbb{E}\left(\mathbf{x}\right)=0_q$ and $\mathbb{V}(\mathbf{x})=diag(\sigma_{x_j}^2),\forall j=1,...,q$. As shown in Table A1, the individual contribution to the (opposite) MSE is defined as $G(y;\mathbf{x};\delta_0))=-(y-\mathbf{x}\hat{\beta})^2$, where $\mathbf{x}\hat{\beta}=\hat{f}(\mathbf{x})$. Similarly, $G(y;\mathbf{x};\delta_0))$ can be expressed as:

$$G(y;\mathbf{x};\delta_0))=-\left[y^2+\sum_{j=1}^{q}x_j^2\hat{\beta}_j^2+2\sum_{1\leq j<l\leq q}\hat{\beta}_j\hat{\beta}_l x_j x_l-2\sum_{j=1}^{q}yx_j\hat{\beta}_j\right]. \tag{A45}$$

To complete the proof, we first start by proving that the (opposite) MSE can be written as:

$$\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0))=2\sum_{j=1}^{q}\hat{\beta}_j\sigma_{y,x_j}-\sum_{j=1}^{q}\hat{\beta}_j^2\sigma_{x_j}^2-\sigma_y^2. \tag{A46}$$

Indeed, by taking the expected value of Equation A45 with respect to the joint distribution of the target variable and the features, we obtain:

$$\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0))=-\mathbb{E}(y^2)-\sum_{j=1}^{q}\hat{\beta}_j^2\mathbb{E}(x_j^2)-2\sum_{1\leq j<l\leq q}\hat{\beta}_j\hat{\beta}_l\mathbb{E}(x_j x_l)+2\sum_{j=1}^{q}\hat{\beta}_j\mathbb{E}(yx_j). \tag{A47}$$

As $\mathbb{E}\left(\mathbf{x}\right)=0_q$ and $\mathbb{V}(\mathbf{x})=diag(\sigma_{x_j}^2),\forall j=1,...,q$ we have:

$$\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0))=-\mathbb{E}(y^2)-\sum_{j=1}^{q}\hat{\beta}_j^2\sigma_{x_j}^2+2\sum_{j=1}^{q}\hat{\beta}_j\sigma_{y,x_j}.$$

As we assume that $\mathbb{E}(y)=0$, we have:

$$\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0))=-\sigma_y^2-\sum_{j=1}^{q}\hat{\beta}_j^2\sigma_{x_j}^2+2\sum_{j=1}^{q}\hat{\beta}_j\sigma_{y,x_j}. \tag{A48}$$

Second, we prove that the benchmark value of the (opposite) MSE can be written as:

$$\phi_0 = \mathbb{E}_y \mathbb{E}_{\mathbf{x}}(G(y; \mathbf{x}; \delta_0)) = -\sigma_y^2 - \sum_{j=1}^{q} \hat{\beta}_j^2 \sigma_{x_j}^2. \tag{A49}$$

From Equation A45, by taking the expected value with respect to the target variable and the expected value with respect to the joint distribution of the features we obtain:

$$\phi_0 = \mathbb{E}_y \mathbb{E}_{\mathbf{x}}(G(y; \mathbf{x}; \delta_0)) = -\mathbb{E}(y^2) - \sum_{j=1}^{q} \hat{\beta}_j^2 \mathbb{E}(x_j^2) - 2 \sum_{1 \leq j < l \leq q} \hat{\beta}_j \hat{\beta}_l \mathbb{E}(x_j x_l) + 2 \sum_{j=1}^{q} \hat{\beta}_k \mathbb{E}(y) \mathbb{E}(x_j). \tag{A50}$$

As $\mathbb{E}(\mathbf{x}) = 0_q$ and $\mathbb{V}(\mathbf{x}) = diag(\sigma_{x_j}^2), \forall j = 1, ..., q$, we have:

$$\phi_0 = \mathbb{E}_y \mathbb{E}_{\mathbf{x}}(G(y; \mathbf{x}; \delta_0)) = -\sigma_y^2 - \sum_{j=1}^{q} \hat{\beta}_j^2 \sigma_{x_j}^2. \tag{A51}$$

Third, we show that the XPER value associated to the feature x_j for the (opposite) MSE can be expressed as:

$$\phi_j = 2 \hat{\beta}_j \sigma_{y, x_j}. \tag{A52}$$

Note that the individual contribution to the R^2, defined as $G^{R^2}(y; \mathbf{x}; \delta_0) = 1 - \sigma_y^{-2}(y - \mathbf{x}\hat{\beta})$ according to Equation 2, can be expressed as $G^{R^2}(y; \mathbf{x}; \delta_0) = 1 + \sigma_y^{-2} G(y; \mathbf{x}; \delta_0)$, with $G(y; \mathbf{x}; \delta_0)$ the individual contribution to the MSE. According to Definition 3, the XPER value associated to the feature x_j for the R^2:

$$\begin{aligned}
\phi_j^{R^2} &= \sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\} \backslash \{x_j\})} \omega_S \left[\mathbb{E}_{y, x_j, \mathbf{x}^S} \mathbb{E}_{\mathbf{x}^{\overline{S}}} \left(1 + \sigma_y^{-2} G(y; \mathbf{x}; \delta_0)\right) - \mathbb{E}_{y, \mathbf{x}^S} \mathbb{E}_{\mathbf{x}^{x_j, \overline{S}}} \left(1 + \sigma_y^{-2} G(y; \mathbf{x}; \delta_0)\right) \right] \\
\phi_j^{R^2} &= \sigma_y^{-2} \left[\sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\} \backslash \{x_j\})} \omega_S \left[\mathbb{E}_{y, x_j, \mathbf{x}^S} \mathbb{E}_{\mathbf{x}^{\overline{S}}} \left(G(y; \mathbf{x}; \delta_0)\right) - \mathbb{E}_{y, \mathbf{x}^S} \mathbb{E}_{\mathbf{x}^{x_j, \overline{S}}} \left(G(y; \mathbf{x}; \delta_0)\right) \right] \right] \\
\phi_j^{R^2} &= \frac{\phi_j}{\sigma_y^2}.
\end{aligned} \tag{A53}$$

Therefore, from Equation 6 and A53, we obtain:

$$\phi_j = 2\hat{\beta}_j \sigma_{y,x_j}. \tag{A54}$$

Finally, from Equation A48, A51, and A54, we conclude that:

$$\underbrace{2\sum_{j=1}^{q} \hat{\beta}_j \sigma_{y,x_j} - \sum_{j=1}^{q} \hat{\beta}_j^2 \sigma_{x_j}^2 - \sigma_y^2}_{\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0))} = \underbrace{-\sum_{j=1}^{q} \hat{\beta}_j^2 \sigma_{x_j}^2 - \sigma_y^2}_{\phi_0} + \sum_{j=1}^{q} \underbrace{2\hat{\beta}_j \sigma_{y,x_j}}_{\phi_j}. \tag{A55}$$

□

E.4 Proof of the R^2 example in Table A1

Proposition 4. *Consider a linear regression model* $\hat{y} = \hat{f}(\mathbf{x}) = \sum_{j=1}^{q} \hat{\beta}_j x_j$, *where* $\mathbb{E}(\mathbf{x}) = 0_q$ *and* $\mathbb{V}(\mathbf{x}) = diag(\sigma_{x_j}^2), \forall j = 1, ..., q$, *and the features are uncorrelated to the residuals* $\hat{\varepsilon} = y - \hat{y}$. *The contributions* ϕ_j *of features* x_j *to the* R^2 *satisfy the efficiency axiom such that:*

$$\underbrace{\frac{\sigma_{y,\hat{y}}}{\sigma_y^2}}_{\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0))} = \underbrace{-\frac{\sigma_{y,\hat{y}}}{\sigma_y^2}}_{\phi_0} + \sum_{j=1}^{q} \underbrace{\frac{2\hat{\beta}_j \sigma_{y,x_j}}{\sigma_y^2}}_{\phi_j}. \tag{A56}$$

Proof. Consider a linear regression model $\hat{f}(\mathbf{x}) = \sum_{j=1}^{q} \hat{\beta}_j x_j$, where $\mathbb{E}(\mathbf{x}) = 0_q$ and $\mathbb{V}(\mathbf{x}) = diag(\sigma_{x_j}^2), \forall j = 1, ..., q$. As shown in Equation 2, the individual contribution to the R^2 is defined as $G(y; \mathbf{x}; \delta_0)) = 1 - \sigma_y^2 (y - \mathbf{x}\hat{\beta})^2$, where $\mathbf{x}\hat{\beta} = \hat{f}(\mathbf{x})$. Similarly, $G(y; \mathbf{x}; \delta_0))$ can be expressed as:

$$G(y; \mathbf{x}; \delta_0)) = 1 - \sigma_y^{-2} \left[y^2 + \sum_{j=1}^{q} x_j^2 \hat{\beta}_j^2 + 2 \sum_{1 \le j < l \le q} \hat{\beta}_j \hat{\beta}_l x_j x_l - 2 \sum_{j=1}^{q} y x_j \hat{\beta}_j \right]. \tag{A57}$$

To complete the proof, we first start by proving that the (opposite) MSE can be written as:

$$\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0)) = \frac{\sigma_{y,\hat{y}}}{\sigma_y^2}. \tag{A58}$$

Indeed, by taking the expected value of Equation A57 with respect to the joint distribution of the target variable and the features, we obtain:

$$\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0)) = 1 - \sigma_y^{-2}\left[\mathbb{E}(y^2) + \sum_{j=1}^{q}\hat{\beta}_j^2\mathbb{E}(x_j^2) + 2\sum_{1\leq j<l\leq q}\hat{\beta}_j\hat{\beta}_l\mathbb{E}(x_jx_l) - 2\sum_{j=1}^{q}\hat{\beta}_j\mathbb{E}(yx_j)\right]. \tag{A59}$$

We assume that $\mathbb{E}\left(\mathbf{x}\right) = 0_q$ and $\mathbb{V}(\mathbf{x}) = diag(\sigma_{x_j}^2), \forall j = 1, ..., q$. Therefore, we have:

$$\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0)) = 1 - \sigma_y^{-2}\left[\mathbb{E}(y^2) + \sum_{j=1}^{q}\hat{\beta}_j^2\sigma_{x_j}^2 - 2\sum_{j=1}^{q}\hat{\beta}_j\sigma_{y,x_j}\right].$$

In a linear model without intercept, we know that $\mathbb{E}(y) = 0$, therefore:

$$\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0)) = -\sum_{j=1}^{q}\frac{\hat{\beta}_j^2\sigma_{x_j}^2}{\sigma_y^2} + \sum_{j=1}^{q}\frac{2\hat{\beta}_j\sigma_{y,x_j}}{\sigma_y^2}. \tag{A60}$$

Moreover, in a linear regression model, the target variable can be expressed as $y = \hat{y} + \hat{\varepsilon}$, with $\hat{y}$ the estimated model and $\hat{\varepsilon}$ the residuals. As the features are uncorrelated from each other, if we also assume that they are uncorrelated to the residuals we can show that:

$$\sigma_{y,\hat{y}} = \sum_{j=1}^{q}\hat{\beta}_j^2\sigma_{x_j}^2 = \sum_{j=1}^{q}\hat{\beta}_j\sigma_{y,x_j}, \tag{A61}$$

with $\sigma_{y,\hat{y}}$ the covariance between the target variable and its prediction. Therefore, the R^2 as expressed in Equation A60 is also written as:

$$\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0)) = \frac{\sigma_{y,\hat{y}}}{\sigma_y^2}. \tag{A62}$$

Second, we prove that the benchmark value of R^2 can be written as:

$$\phi_0 = \mathbb{E}_y \mathbb{E}_{\mathbf{x}}(G(y; \mathbf{x}; \delta_0)) = \frac{\sigma_{y,\hat{y}}}{\sigma_y^2}. \tag{A63}$$

From Equation A45, by taking the expected value with respect to the target variable and the expected value with respect to joint distribution of the features, we obtain:

$$\begin{aligned}\phi_0 = \mathbb{E}_y \mathbb{E}_{\mathbf{x}}(G(y; \mathbf{x}; \delta_0)) = 1 - \sigma_y^{-2} &\left[\mathbb{E}(y^2) + \sum_{j=1}^{q} \hat{\beta}_j^2 \mathbb{E}(x_j^2) + 2 \sum_{1 \leq j < l \leq q} \hat{\beta}_j \hat{\beta}_l \mathbb{E}(x_j x_l) \right] \\ - \sigma_y^{-2} &\left[-2 \sum_{j=1}^{q} \hat{\beta}_j \mathbb{E}(y) \mathbb{E}(x_j) \right]. \end{aligned} \tag{A64}$$

We assume that $\mathbb{E}(\mathbf{x}) = 0_q$ and $\mathbb{V}(\mathbf{x}) = diag(\sigma_{x_j}^2), \forall j = 1, ..., q$. Therefore, we have:

$$\phi_0 = \mathbb{E}_y \mathbb{E}_{\mathbf{x}}(G(y; \mathbf{x}; \delta_0)) = -\frac{\sum_{j=1}^{q} \hat{\beta}_j^2 \sigma_{x_j}^2}{\sigma_y^2}. \tag{A65}$$

From Equation A61, we can rewrite Equation A65 as:

$$\phi_0 = \mathbb{E}_y \mathbb{E}_{\mathbf{x}}(G(y; \mathbf{x}; \delta_0)) = -\frac{\sigma_{y,\hat{y}}}{\sigma_y^2}. \tag{A66}$$

Third, as stated in Proposition 1 the XPER value associated to the feature x_j for the R^2 can be expressed as:

$$\phi_j = \frac{2\hat{\beta}_j \sigma_{y,x_j}}{\sigma_y^2}. \tag{A67}$$

Finally, from Equation A62, A66, and A67, we conclude that:

$$\underbrace{\frac{\sigma_{y,\hat{y}}}{\sigma_y^2}}_{\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0))} = \underbrace{-\frac{\sigma_{y,\hat{y}}}{\sigma_y^2}}_{\phi_0} + \sum_{j=1}^{q} \underbrace{\frac{2\hat{\beta}_j \sigma_{y,x_j}}{\sigma_y^2}}_{\phi_j}. \tag{A68}$$

$\square$

E.5 Proof of the accuracy example in Table A1

Proposition 5. *Consider any binary classification model* $\hat{f}(\mathbf{x})$*, with* $\hat{P}(\mathbf{x}) = \hat{\mathbb{P}}(y = 1|\mathbf{x})$ *the estimated probability of belonging to class 1 (y=1). The contributions* ϕ_j *of features* x_j *to the accuracy satisfy the efficiency axiom such that:*

$$\underbrace{2\sigma_{y,\hat{f}(\mathbf{x})} + 2\mathbb{P}(y=1)\hat{P}(\mathbf{x}) + 1 - \mathbb{P}(y=1) - \hat{P}(\mathbf{x})}_{\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0))} = \underbrace{2\mathbb{P}(y=1)\hat{P}(\mathbf{x}) + 1 - \mathbb{P}(y=1) - \hat{P}(\mathbf{x})}_{\phi_0} + \underbrace{2\sigma_{y,\hat{f}(\mathbf{x})}}_{\sum_{j=1}^{q}\phi_j}, \tag{A69}$$

with $\hat{P}(\mathbf{x}) = \hat{\mathbb{P}}(y = 1|\mathbf{x})$ *and* $\sigma_{y,\hat{f}(\mathbf{x})}$ *the covariance between the target variable and the classification output.*

Proof. Consider any binary classification model $\hat{f}(\mathbf{x})$, with $\hat{P}(\mathbf{x}) = \hat{\mathbb{P}}(y = 1|\mathbf{x})$ the estimated probability of belonging to class 1 (y=1). As shown in Table A1, the individual contribution to the accuracy is defined as:

$$G(y;\mathbf{x};\delta_0)) = y\hat{f}(\mathbf{x}) + (1-y)(1-\hat{f}(\mathbf{x})). \tag{A70}$$

To complete the proof, we first start by proving that the accuracy can be written as:

$$\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0)) = 2\sigma_{y,\hat{f}(\mathbf{x})} + 2\mathbb{P}(y=1)\hat{P}(\mathbf{x}) + 1 - \mathbb{P}(y=1) - \hat{P}(\mathbf{x}). \tag{A71}$$

Indeed, by taking the expected value of Equation A70 with respect to the joint distribution of the target variable and the features, we obtain:

$$\begin{aligned}\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0)) &= 2\mathbb{E}\left(y\hat{f}(\mathbf{x})\right) + 1 - \mathbb{E}(y) - \mathbb{E}\left(\hat{f}(\mathbf{x})\right) \\ &= 2\sigma_{y,\hat{f}(\mathbf{x})} + 2\mathbb{E}(y)\mathbb{E}\left(\hat{f}(\mathbf{x})\right) + 1 - \mathbb{E}(y) - \mathbb{E}\left(\hat{f}(\mathbf{x})\right),\end{aligned} \tag{A72}$$

with $\sigma_{y,\hat{f}(\mathbf{x})}$ the covariance between the target variable and the classification output. As the target variable is binary, we know that $\mathbb{E}(y) = \mathbb{P}(y=1)$ and $\mathbb{E}\left(\hat{f}(\mathbf{x})\right) = \hat{\mathbb{P}}(y=1|\mathbf{x}) = \hat{P}(\mathbf{x})$. Therefore, we obtain:

$$\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0)) = 2\sigma_{y,\hat{f}(\mathbf{x})} + 2\mathbb{P}(y=1)\hat{P}(\mathbf{x}) + 1 - \mathbb{P}(y=1) - \hat{P}(\mathbf{x}). \tag{A73}$$

Second, from Equation A70, we can see that by taking the expected value with respect to the target variable and the expected value with respect to joint distribution of the features, the benchmark value of accuracy can be written as:

$$\begin{aligned}\phi_0 = \mathbb{E}_y\mathbb{E}_{\mathbf{x}}(G(y;\mathbf{x};\delta_0)) &= 2\mathbb{E}\left(y\right)\mathbb{E}\left(\hat{f}(\mathbf{x})\right) + 1 - \mathbb{E}(y) - \mathbb{E}\left(\hat{f}(\mathbf{x})\right) \\ &= 2\mathbb{P}(y=1)\hat{P}(\mathbf{x}) + 1 - \mathbb{P}(y=1) - \hat{P}(\mathbf{x}).\end{aligned} \tag{A74}$$

Third, according to the axiom 1, we deduce that the sum of the XPER values associated to the features x_j for the accuracy is equal to:

$$\sum_{j=1}^{q} \phi_j = 2\sigma_{y,\hat{f}(\mathbf{x})}. \tag{A75}$$

Finally, from Equation A73, A74 and A75, we conclude that:

$$\begin{aligned}\underbrace{2\sigma_{y,\hat{f}(\mathbf{x})} + 2\mathbb{P}(y=1)\hat{P}(\mathbf{x}) + 1 - \mathbb{P}(y=1) - \hat{P}(\mathbf{x})}_{\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0))} &= \underbrace{2\mathbb{P}(y=1)\hat{P}(\mathbf{x}) + 1 - \mathbb{P}(y=1) - \hat{P}(\mathbf{x})}_{\phi_0} \\ &+ \underbrace{2\sigma_{y,\hat{f}(\mathbf{x})}}_{\sum_{j=1}^{q}\phi_j}.\end{aligned} \tag{A76}$$

□

E.6 Proof of Equation A10

Proposition 6. *Consider a regression model $\hat{f}(\mathbf{x})$ including only one feature x_1 such as $\mathbf{x} = x_1$. We can show that for the performance metric $G(y_i; \hat{f}(\mathbf{x}_i); \delta_0) = -(y_i - \hat{f}(\mathbf{x}_i))^2$, if we assume that $\mathbb{E}(\hat{f}(x_{i,1})) = 0$, then the corresponding individual XPER value $\phi_{i,1}$ is equal to:*

$$\phi_{i,1} = 2\hat{\varepsilon}_i \phi_{i,1}^{SHAP} + \left(\phi_{i,1}^{SHAP}\right)^2 + \mathbb{V}\left(\hat{f}(x_{i,1})\right), \tag{A77}$$

where $\hat{\varepsilon}_i = y_i - \hat{f}(x_{i,1})$ is the prediction error of the model for individual i, and $\phi_{i,1}^{SHAP}$ refers to the SHAP value of the feature x_1 for this individual.

Proof. Consider a regression model $\hat{f}(\mathbf{x})$ including only one feature x_1 such as $\mathbf{x} = x_1$, and the performance metric $G(y_i; \hat{f}(\mathbf{x}_i); \delta_0) = -(y_i - \hat{f}(\mathbf{x}_i))^2$.

According to Equation 7, the XPER value $\phi_{i,1}$ decomposes $G(y_i; \hat{f}(\mathbf{x}_i); \delta_0)$ such as:

$$G(y_i; \hat{f}(\mathbf{x}_i); \delta_0) = \phi_{i,0} + \phi_{i,1}, \tag{A78}$$

with $\phi_{i,0}$ the benchmark value of the performance metric. Therefore, if we replace $G(y_i; \hat{f}(\mathbf{x}_i); \delta_0)$ by its expression in the previous equation, we can see that:

$$\phi_{i,1} = -y_i^2 - \hat{f}(x_{i,1})^2 + 2y_i\hat{f}(x_{i,1}) - \phi_{i,0}. \tag{A79}$$

Moreover, the benchmark value $\phi_{i,0}$ is equal to:

$$\phi_{i,0} = \mathbb{E}_{x_1}\left(G(y_i; \hat{f}(\mathbf{x}_i); \delta_0)\right) = -y_i^2 - \mathbb{E}\left(\hat{f}(x_{i,1})^2\right) + 2y_i\mathbb{E}\left(\hat{f}(x_{i,1})\right). \tag{A80}$$

As we assume that $\mathbb{E}\left(\hat{f}(x_{i,1})\right) = 0$, we obtain:

$$\phi_{i,0} = -y_i^2 - \mathbb{V}\left(\hat{f}(x_{i,1})\right). \tag{A81}$$

Replacing the expression of $\phi_{i,0}$ in Equation A79 we obtain:

$$\phi_{i,1} = 2y_i\hat{f}(x_{i,1}) + \mathbb{V}\left(\hat{f}(x_{i,1})\right) - \hat{f}(x_{i,1})^2. \tag{A82}$$

As the prediction error ε_i can be expressed as the difference between the target variable y_i and the prediction $\hat{f}(x_{i,1})$, i.e., $\hat{\varepsilon}_i = y_i - \hat{f}(x_{i,1})$, $\phi_{i,1}$ is then equal to:

$$\phi_{i,1} = 2\hat{\varepsilon}_i\hat{f}(x_{i,1}) + \hat{f}(x_{i,1})^2 + \mathbb{V}\left(\hat{f}(x_{i,1})\right). \tag{A83}$$

Now, according to Equation A8 and A9, as $\mathbb{E}\left(\hat{f}(x_{i,1})\right) = 0$, we can see that:

$$\hat{f}(x_{i,1}) = \phi_{i,1}^{SHAP}. \tag{A84}$$

with $\phi_{i,1}^{SHAP}$ the SHAP value associated to feature x_1 for individual i.

Finally, from Equation A83 and A84, we obtain:

$$\phi_{i,1} = 2\hat{\varepsilon}_i\phi_{i,1}^{SHAP} + \left(\phi_{i,1}^{SHAP}\right)^2 + \mathbb{V}\left(\hat{f}(x_{i,1})\right). \tag{A85}$$

□

E.7 Proof of Proposition 1

Proposition 1. *SHAP is a particular case of XPER where the individual contribution to the performance metric is equal to the predicted value of the model,* $G\left(y_i; \mathbf{x}_i; \delta_0\right) = \hat{f}(\mathbf{x}_i)$.

Proof. According to Definition 4, the individual XPER value ϕ_j associated to the performance metric $G\left(y_i; \mathbf{x}_i; \delta_0\right) = \hat{f}(\mathbf{x}_i)$ is equal to:

$$\phi_{i,j} = \sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\} \backslash \{x_j\})} w_S \left[\mathbb{E}_{\mathbf{x}^{\overline{S}}}\left(\hat{f}\left(x_{i,j}, \mathbf{x}_i^S, \mathbf{x}^{\overline{S}}\right)\right) - \mathbb{E}_{x_j, \mathbf{x}^{\overline{S}}}\left(\hat{f}\left(x_j, \mathbf{x}_i^S, \mathbf{x}^{\overline{S}}\right)\right)\right]. \tag{A86}$$

Moreover, according Equation A7, the SHAP value ϕ_j^{SHAP} associated to $\hat{f}(\mathbf{x}_i)$ is equal to:

$$\phi_{i,j}^{SHAP} = \sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\}\backslash\{x_j\})} w_S \left[\mathbb{E}_{\mathbf{x}^{\overline{S}}} \left(\hat{f}(x_{i,j}, \mathbf{x}_i^S, \mathbf{x}^{\overline{S}}) \right) - \mathbb{E}_{x_j, \mathbf{x}^{\overline{S}}} \left(\hat{f}(x_j, \mathbf{x}_i^S, \mathbf{x}^{\overline{S}}) \right) \right]. \tag{A87}$$

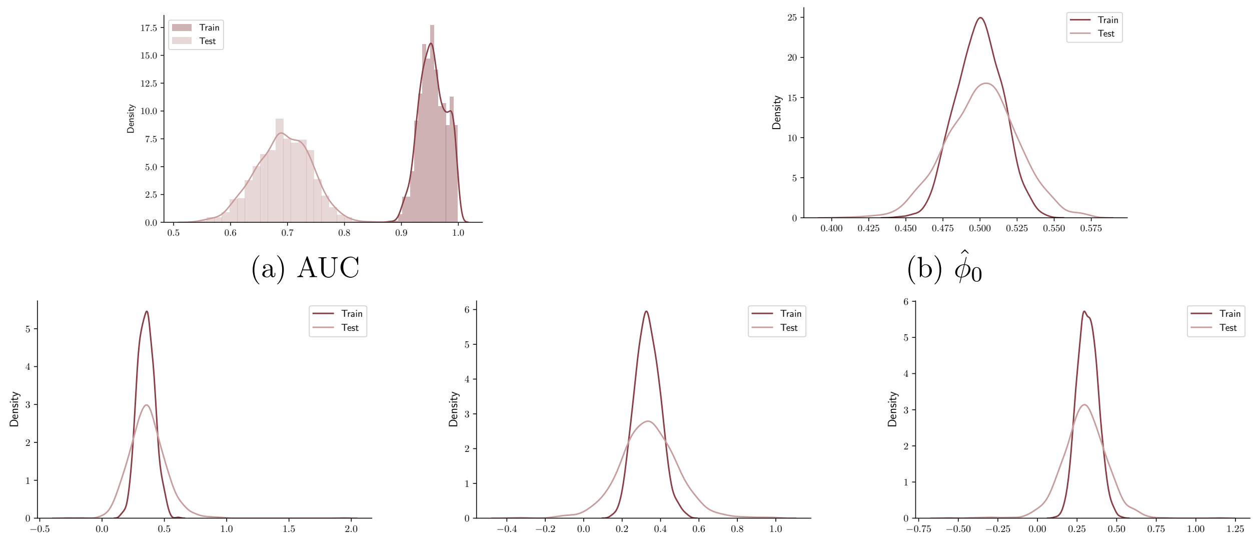
Therefore, in the particular case where $G\left(y_i; \mathbf{x}_i; \delta_0\right) = \hat{f}(\mathbf{x}_i)$, according to Equation A86 and A87, the individual XPER value ϕ_j is equal to the SHAP value ϕ_j^{SHAP}. □

F Simulations: Explaining Overfitting

In Appendix F.1 and F.2, we propose two additional Monte Carlo simulation experiments illustrating how XPER values can be used to detect the origin of overfitting. The latter can arise for at least two reasons: (1) an improper control of the bias-variance trade-off through model hyperparameters, or (2) a shift of the feature distributions between the training and test samples. We illustrate each of these two cases in the following subsections.

F.1 Case 1: Improper control of the bias-variance trade-off

Consider a DGP given by $y_i = \mathbf{1}(y_i^* > 0)$ with $y_i^* = \omega_i \beta + \varepsilon_i$ a latent variable, $\omega_i = (1 : \mathbf{x}_i^{'})$, and ε_i an i.i.d. error term with $\varepsilon_i \sim \mathcal{N}(0, 1)$. We consider three independent features such that $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ with $diag(\Sigma) = (1.3, 1.2, 1.1)$. The true vector of parameters is $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^{'} = (0.05, 0.5, 0.5, 0.5)^{'}$ with β_0 the intercept. We generate $K = 5,000$ pseudo-samples $\{y_i^s, \mathbf{x}_i^s\}_{i=1}^{T+n}$ of size $1,000$ using this DGP. Then, we estimate a decision tree using 5-fold cross validation on the first $T = 700$ observations of each pseudo-sample and we use the remaining $n = 300$ observations as a test sample. In order to intentionally generate overfitting, we impose a minimum tree-depth of 6 nodes for only three features in the model. For each trained model, we implement XPER to decompose the effect of the features on the AUC of the training and the test samples. We display in Figure A1a the empirical distributions of the AUC. As expected, the trained tree models are overfitting the data, illustrated by the relatively low AUC values obtained on the test samples compared to the training samples. The empirical distributions of the XPER values reported in other panels of Figure A1 show that this drop in performance does not come from a particular feature. Indeed, the XPER contributions to the AUC are relatively close between the training and the test sample for all features. Thus, when overfitting is due to an improper control of the bias-variance trade-off, we observe a large decrease of the performance metric along with a stability of XPER values between the training and the test sample. Therefore, XPER can be used as a reverse engineering tool to detect wrong settings of hyperparameters.



(a) AUC (b) $\hat{\phi}_0$

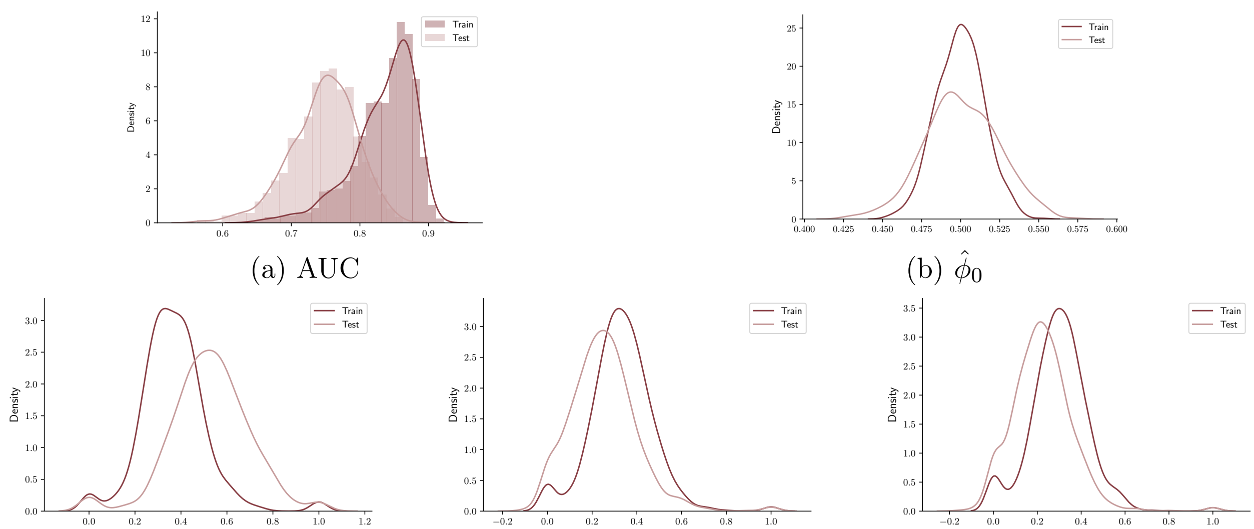
(c) $\hat{\phi}_1$ (d) $\hat{\phi}_2$ (e) $\hat{\phi}_3$

Figure A1: Empirical distributions of AUC and XPER values in case of overfitting due to improper control of the bias-variance trade-off

Note: This figure displays the empirical distributions of the AUC and XPER values on the training (dark color) and test (light color) sample according to the framework detailed in Illustration 2, case 1. XPER values are divided by the difference between the AUC of the model and the benchmark value to be comparable between the training and the test sample. The solid lines refer to kernel density estimations.

F.2 Case 2: Shift of the feature distribution

Overfitting can also arise from a shift of the feature distributions between the training and the test sample. To illustrate this origin of overfitting, we consider two distinct DGPs for the training and the test sample. For the former, we keep the same DGP as in the first case. For the test sample, we assume an increase in the variance of the first feature while keeping other parameters unchanged, such that $diag(\tilde{\boldsymbol{\Sigma}}) = (3, 1.2, 1.1)$. In the context of time series, such shift in the variance can come from a structural change. See Perron and Yamamoto (2021) on how to detect forecasting performance changes with structural change tests. As in case 1, we generate $K = 5,000$ pseudo-samples $\{y_i^s, \mathbf{x}_i^s\}_{i=1}^{T+n}$ of size 1,000 ($T = 700$ and $n = 300$). For each pseudo-sample, we estimate a decision tree with a depth between 1 to 5 using 5-fold cross validation. Setting a relatively low tree depth avoids overfitting due to an improper control of the bias-variance trade-off.



(a) AUC (b) $\hat{\phi}_0$ (c) $\hat{\phi}_1$ (d) $\hat{\phi}_2$ (e) $\hat{\phi}_3$

Figure A2: Empirical distributions of AUC and XPER values in case of overfitting due to a shift of the distribution of the features

Note: This figure displays the empirical distributions of the AUC and XPER values on the training (dark color) and test (light color) sample according to the framework detailed in Illustration 2, case 2. XPER values are divided by the difference between the AUC of the model and the benchmark value to be comparable between the training and the test sample. The solid lines refer to kernel density estimations.

In Figure A2a, we observe a decrease in AUC between the training and test samples. Contrary to the previous case, this decrease is due to the shift of the distribution of x_1 which has also an impact on XPER values. More precisely, in Figure A2, we observe that the contribution of feature x_1 to the AUC increases from the training to the test sample whereas the contribution of the other features decreases. Thus, observing both a drop in the performance of the model and some variations in the XPER values from the training to the test sample can indicate a change in the data structure. Such change is not captured by the model and not related to hyperparameter settings.

G Empirical Application: Some additional results

G.1 Summary statistics and features distribution

Table A4: Summary Statistics

	Count	Mean	Std.	Minimum	25%	50%	75%	Maximum
Job tenure	7,440	9.3298	9.9787	0	2	5	15	58
Age	7,440	45.1691	14.7965	18	33	46	55	89
Car price	7,440	12,935	6,204	546	8,149	11,950	16,500	47,051
Funding amount	7,440	11,461	6,019	546	6,846	10,382	15,000	30,000
Loan duration	7,440	56.2176	19.3833	6	48	60	72	96
Monthly payment	7,440	0.1051	0.0611	0.0051	0.0690	0.0947	0.1304	2.6300
Downpayment	7,440	0.0897		0				1
Credit event	7,440	0.0220		0				1
Married	7,440	0.5347		0				1
Homeowner	7,440	0.3848		0				1
Default	7,440	0.2000		0				1

Note: This table displays summary statistics for each feature used in the XGBoost model as well as the target variable. For each categorical feature the standard deviation (Std.) and the quartiles (25%, 50% and 75%) are not displayed.

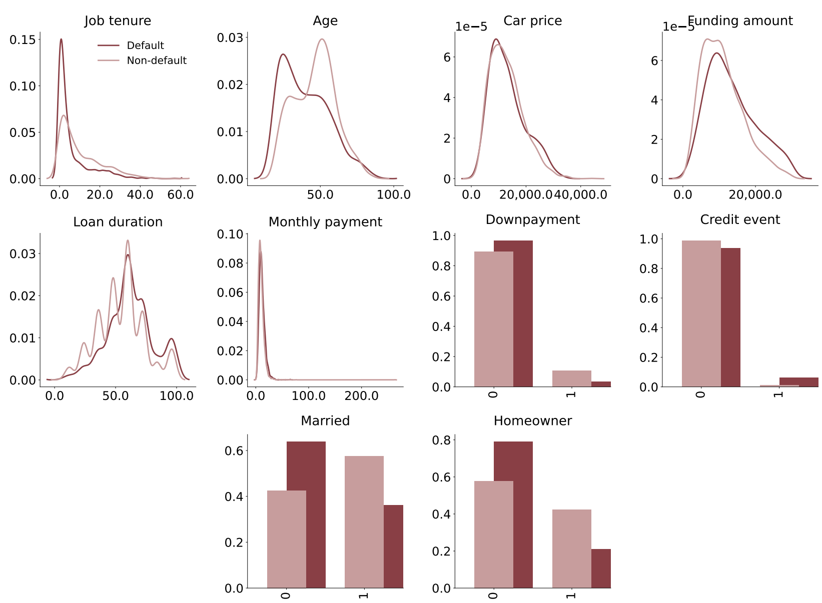


Figure A3: Features distribution by default class

Note: This figure displays the distribution of the features by default class on the training sample, using kernel density estimation for continuous features. Dark red refers to defaulting borrowers and light red to non-defaulting borrowers.

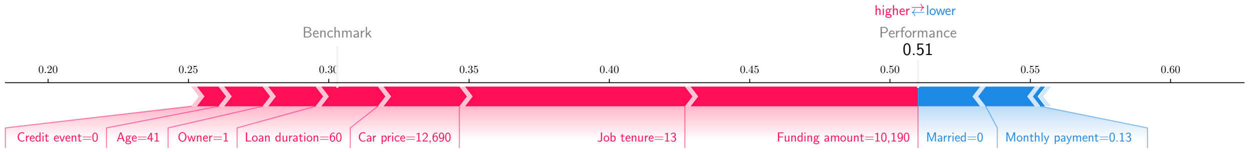
G.2 XPER decomposition

Figure A4: XPER decomposition of the AUC

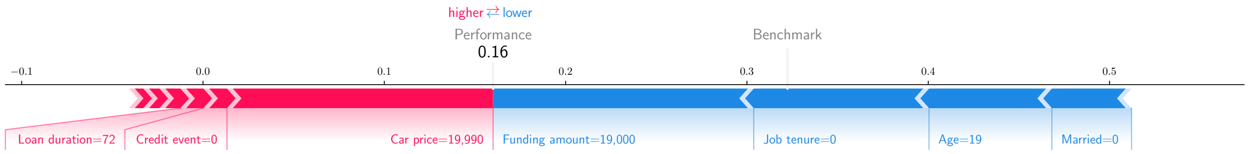
Note: This figure displays the XPER values of the AUC of the XGBoost model estimated on the training sample.

G.3 Individual XPER decomposition and data visualization

We analyse the impact of the various features on the performance metric but we now do it for each borrower individually. We start by analysing in Figure A5 the XPER decomposition for two sample borrowers. These force plots enable us to decompose the individual performance of each borrower, as defined in Equation 7. By doing so, they allow us to understand why some individuals contribute more to the AUC of the model than others. In each panel of Figure A5, *Performance* refers to the contribution of the borrower to the AUC of the model and *Benchmark* to their benchmark value, i.e., $\phi_{i,0}$ in Equation 7. For each borrower, the features increasing (respectively decreasing) the performance appear in red (blue). Borrower #3 has a relatively high individual AUC compared to borrower #28 (both have theoretically the same benchmark). The over-performance of borrower #3 is mainly due to the large positive XPER values for *funding amount*, *job tenure*, and *car price*. It also comes from the small negative XPER values for the marital status (*married*) and the share of the monthly payment in the borrower's income (*monthly payment*).



(a) Borrower #3



(b) Borrower #28

Figure A5: Force plots of individual XPER values

Note: This figure displays the XPER decomposition (for the AUC) of two loan borrowers (see Equation 4). Borrower #3 did not default on his loan and has a probability of default of 8% according to the XGBoost ((Panel (a)). Borrower #28 did not default on his loan and has a probability of default of 57% according to the XGBoost (Panel (b)). *Performance* refers to the individual level of the AUC whereas *Benchmark* represents the individual contribution to the AUC associated to a population where the target variable y_i is independent from the features $\mathbf{x}_i$. The red color refers to positive XPER values, i.e., features increasing performance. The blue color refers to negative XPER values, i.e., features decreasing performance.

To better understand the relative influence of each feature for the two borrowers, we analyse their risk-profiles and probabilities of default predicted by the model. Let us start with borrower #3. He is 41 years old, homeowner, has a stable job, and applied for a loan to buy a moderately-priced car. He provided a down payment greater than 50% of the car value and experienced no past credit event. Intuitively, we would naturally classify this borrower as low-risk and this is confirmed by the 8% default probability estimated by the XGBoost model. Thus, as borrower #3 eventually did not default on his loan, his contribution to the AUC is high. The situation of borrower #28 is quite different as he exhibits a higher risk profile (young, jobless, not married, relatively large credit amount, no down payment). Yet, the model remains quite undecided about his capacity to pay back the loan with a 57% estimated default probability. As the AUC measures the discriminatory ability of the model, this uncertainty leads to a low individual contribution, and even lower than the benchmark value.

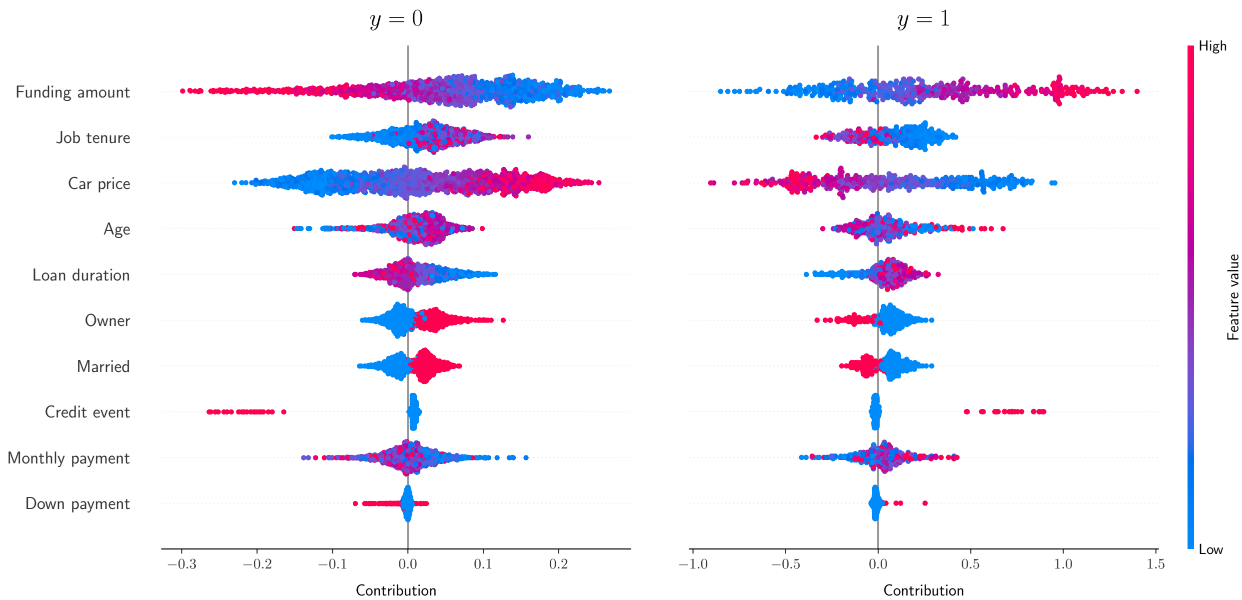
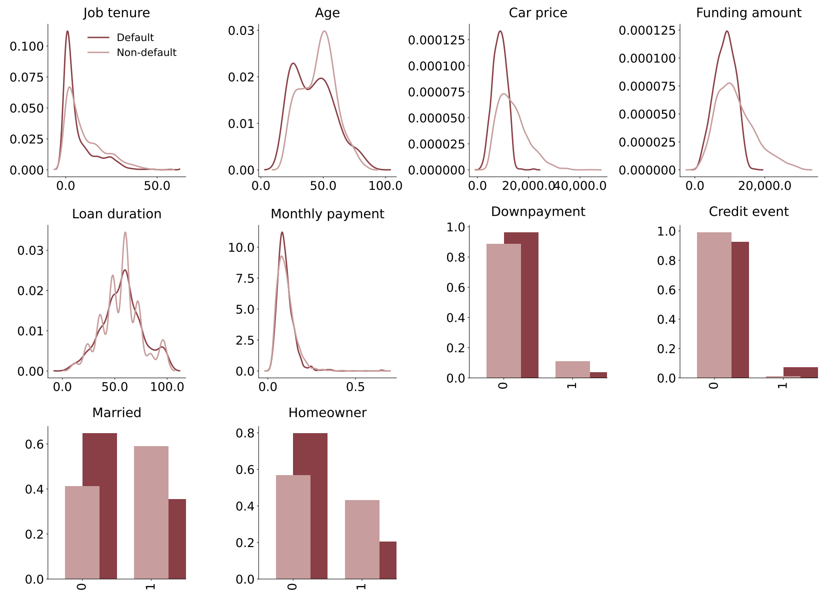


Figure A6: Summary plots of individual XPER values

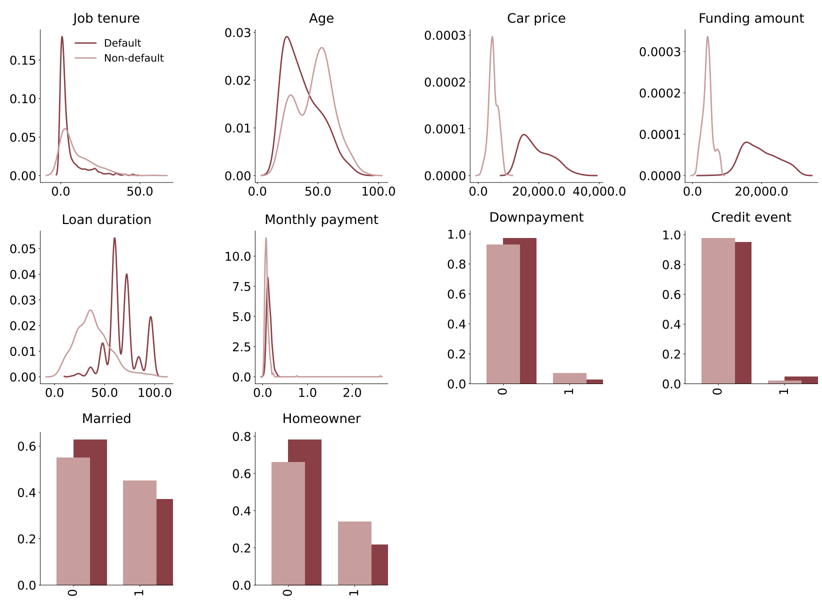
Note: This figure displays the individual XPER values for each feature used in the XGBoost model. Each dot represents the value for a given borrower (see Equation 4). We display the results for the borrowers who paid back their loans (left graphic) and those who do not (right graphic).

We then consider the entire sample of borrowers. In Figure A6, we display the XPER values for each feature as a function of the feature value. We analyse these results according to two types of borrowers: non-defaulting borrowers (y=0) and defaulting borrowers (y=1). We clearly see that depending on the value of the feature and the type of borrower, we know if this feature contributes to increase or decrease the performance of the model. For instance, for a non-defaulting borrower (left panel), a relatively high job tenure is associated to a positive XPER value. This result is due to the fact that a relatively long job tenure tends to lower the probability of default in the model. Hence, this increases the ability of the model to distinguish him from the defaulting borrowers and boosts the XPER value. On the opposite, for a defaulting borrower (right panel), a relatively high job tenure leads to a negative XPER value and thus decreases his contribution to the AUC of the model.

G.4 Boosting model performance



(a) Features distribution by default class in group 1



(b) Features distribution by default class in group 2

Figure A7: Features distribution by default class for each group

Note: This figure displays the distribution of the features by default class on the training sample, for the first (Panel (a)) and second group (Panel (b)) created from individual XPER values using the KMedoids methodology. For continuous features, we use a kernel density estimation. Dark red refers to defaulting borrowers and light red to non-defaulting borrowers.

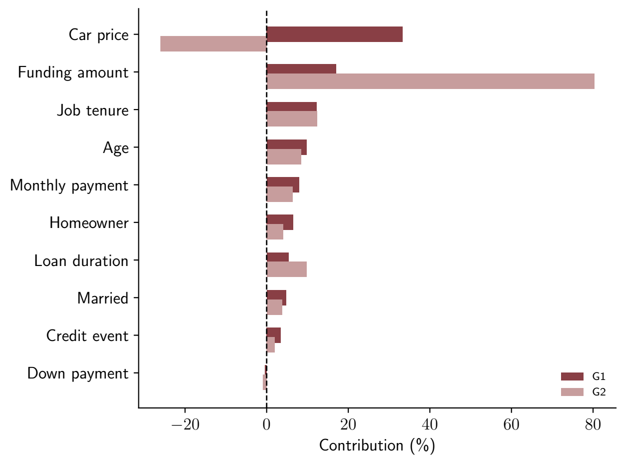


Figure A8: XPER decomposition of the AUC by group

Note: This figure displays the XPER values of the AUC of the XGBoost model estimated on the training sample by group creating with the KMedoid method.