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

Sullivan Hué, Christophe Hurlin, Christophe Pérignon, Sébastien Saurin

A Performance metric decompositions using Shapley values

In this appendix, we provide a literature review including numerous recent studies from computer science and statistics that specifically use Shapley values to measure the individual contributions of features to *model performance*. In contrast, our review does not include articles that aim to decompose model predictions, $\hat{y}$, as this represents a different objective from ours. To facilitate the comparison of our approach with those proposed in the surveyed articles, we present a comparison in Table A1 across several important dimensions.

This comprehensive literature review permits to clearly identify three main contributions of the XPER methodology:

- XPER stands out for its versatility, enabling the analysis of the drivers of *any* performance metric for *any* estimated model (see paragraph starting by "First" below).
- XPER uniquely addresses heterogeneity issues by identifying groups of individuals for which features exhibit similar effects on model performance (see paragraph starting by "Second" below).
- XPER offers a meaningful decomposition of the performance metric, achieving this without relying on strong assumptions, unlike many other methods in the literature (see paragraph starting by "Third" and the following ones below).

First, we contrast papers according to the *metrics they decompose*. Six out of the twelve papers considered in this survey focus exclusively on the decomposition of a single performance metric, which is often the R^2 for the linear regression model (Bell et al., 2024; Israeli, 2007; Lipovetsky and Conklin, 2001). Among the remaining papers, half of them decompose loss functions, which means they

are not designed to decompose accuracy metrics (e.g., AUC, Gini, accuracy), goodness of fit (R^2), information criterion (AIC, BIC), or economic performance metric (profit-and-loss function). Among these, Zhang et al. (2023) propose an original approach that links changes in model performance to shifts in features distributions, relying on the definition of a causal graph. In contrast, XPER enables the analysis of feature contributions for *any* given performance metric. Unlike XPER, four surveyed papers are model-specific, as shown in Table A1.

Second, we distinguish existing methodologies by their *level of analysis*, specifically whether they provide a global and/or a local analysis of the features influencing model performance. All but one of the approaches presented in these studies only deliver a global analysis. Indeed, Sutera et al. (2021) is the only considered article that proposes both a local and a global analysis. However, they only do so for tree-ensembles (a single model) and only for the Shannon entropy impurity measure (a single performance metric). As a result, XPER appears to be the only method able to conduct both local or global analyses for any model and any performance metric. The local analysis enabled by XPER is particularly valuable as it addresses heterogeneity issues by identifying groups of individuals for whom the features have similar effects on performance. This capability provides deeper insights and a more nuanced understanding of feature impacts across different subpopulations.

Third, we classify the approaches based on how they handle features excluded from the feature coalitions used to compute the Shapley values. We identify two main strategies in the literature: (i) re-estimating the model without the excluded features and (ii) marginalizing over the excluded features. While intuitive, the *re-estimating strategy* can result in model specification errors, such as omitted variable bias in linear regression. This concern is particularly relevant here because most of these methods are applied within a linear regression framework to analyze the drivers of R^2. A closely related approach is employed by Moehle et al. (2021) as they use a simulator to model the investment process under scenarios where certain drivers are excluded from the decision-making. A limitation of their approach is that the reliability of the derived feature contributions depends heavily on the accuracy of the simulator.

The *marginalization strategy* involves marginalizing over the features excluded from the model $f(\cdot)$ to avoid re-estimating it. This marginalization can be carried out in different ways, depending on the feature distribution considered (joint or conditional) and whether the performance metric $G(\cdot)$ is evaluated at the expected value of the model's predictions, $G(\mathbb{E}(f(x)))$, or the expected value of the performance metric itself, $\mathbb{E}(G(f(x)))$. These distinctions are not merely technical; they have significant implications for the interpretation of the Shapley decomposition. For instance, Covert et al. (2020) and Borup et al. (2022) evaluate the performance metric at the expected value of the model's predictions. This expected value is computed by integrating over the distribution of the excluded features while treating the included features as fixed.[15] In this case, the benchmark is defined as the performance metric, as the MSE for instance, calculated for a restricted model that predicts a *homogeneous* value, $\mathbb{E}_{\mathbf{x}}(\hat{f}(\mathbf{x}))$, for all instances. The contributions of the features, $\phi_1, \ldots, \phi_q$, then explain the difference between the model's performance and this specific benchmark. For similar reasons, Fryer et al. (2021) examine the use of the SAGE (Covert et al., 2020) and SHAP (Lundberg and Lee, 2017) approaches, concluding that these methods are not well-suited for explaining performance.

In our paper, we designed the marginalization approach to end up with a meaningful benchmark. Specifically, the XPER values explain the difference between the model performance and the performance of a model that would be obtained if the features were independent of the target variable.[16] Our benchmark is therefore defined as the value of the performance metric in the case where the model is completely misspecified. Our reasoning is analogous to a global Fisher test in standard linear regression, where one compares the MSE of the model to its value for a purely misspecified

[15] Using our notations, they compute $\mathbb{E}_{\mathbf{x}^s,y}\left(\widetilde{G}\left(y;\mathbb{E}_{\mathbf{x}^{\overline{S}}}(\hat{f}(\mathbf{x}));\delta_0\right)\right)$, where $\mathbf{x}^S$ ($\mathbf{x}^{\overline{S}}$) represents the vector of features included in (excluded from) coalition S, y is the target variable, δ_0 is a nuisance parameter, and $\widetilde{G}(.)$ is the performance metric comparing the model's prediction $\hat{f}(\mathbf{x})$ to its target value y. In this case, the benchmark ϕ_0 is defined as $\phi_0 = \mathbb{E}_y\left(\widetilde{G}\left(y;\mathbb{E}_{\mathbf{x}}(\hat{f}(\mathbf{x}));\delta_0\right)\right)$, which corresponds to the performance that would be obtained if the model were predicting the same constant value for everyone, $\mathbb{E}_{\mathbf{x}}(\hat{f}(\mathbf{x}))$.

[16] Formally, XPER computes $\mathbb{E}_{\mathbf{x}^s,y}\mathbb{E}_{\mathbf{x}^{\overline{S}}}\left(\widetilde{G}\left(y;\hat{f}(\mathbf{x});\delta_0\right)\right)$. As a consequence, the XPER values $\phi_1, \ldots, \phi_q$ decompose the difference between the model's performance $\mathbb{E}_{\mathbf{x},y}\left(\widetilde{G}\left(y;\hat{f}(\mathbf{x});\delta_0\right)\right)$ and the benchmark value $\phi_0 = \mathbb{E}_y\mathbb{E}_{\mathbf{x}}\left(\widetilde{G}\left(y;\hat{f}(\mathbf{x});\delta_0\right)\right)$.

model in which all variables are insignificant. While we are not conducting a formal test here, the underlying idea is the same. For example, when using the AUC criterion for classification, our benchmark value is 0.5, which corresponds to the value achieved by purely random classification. Thus, our XPER values are meaningful because they measure the improvement in predictive performance attributable to the features relative to this well-founded benchmark.

To the best of our knowledge, the Shapley Feature IMPortance (SFIMP) method proposed by Casalicchio et al. (2019) is the only approach in the literature that performs marginalization in the same manner as we do. However, there are several important differences between SFIMP and XPER. First, XPER enables both local and global analyses of feature contributions to performance, whereas SFIMP can only perform global analyses. Second, while SFIMP decomposes loss functions, XPER is designed to decompose 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). Third, SFIMP does not specify any assumptions about the set of loss functions it can decompose. In contrast, XPER is built on several assumptions about the performance metric, such as the additivity assumption (Assumption 2). This assumption ensures that XPER can decompose any performance that can be expressed as an average of individual contributions to the performance. Crucially, the introduction of the nuisance parameter δ_0, unique to XPER, extends this framework to handle even complex metrics like the AUC. By reformulating such metrics into an additive structure, XPER significantly expands its scope, allowing it to handle a broader range of performance measures than SFIMP. Fourth, as opposed to SFIMP, XPER puts emphasis on the definition of the benchmark ϕ_0 as it plays a key role to explain the meaning of the feature contributions derived from the Shapley value decomposition. Our benchmark has a meaningful interpretation: it corresponds to the performance obtained on a hypothetical sample where the target variable y is independent of all model features $\mathbf{x}$—that is, a scenario in which the model $\hat{f}(\mathbf{x})$ is fully misspecified.

Table A1: Literature review

References	Metrics to be decomposed	Model-agnostic	Global	Local	Data	Applications	Management of excluded features from the model
Bell et al. (2024)	R^2		✓		Synthetic	Numerical experiments	Re-estimation
Borup et al. (2022)	Loss functions	✓	✓		Real	Forecasting inflation	marginalization
Casalicchio et al. (2019)	Loss functions	✓	✓		Real	Housing price prediction	marginalization
Covert et al. (2020)	Loss functions	✓	✓		Real	Many (text classification, credit scoring, etc.)	marginalization
Fryer et al. (2021)	Evaluation functions	✓	✓		Synthetic	Numerical experiments	marginalization
Ghorbani and Zou (2019)	Performance metrics	✓	✓		Real	Many (disease prediction, spam classification, etc.)	Re-estimation
Israeli (2007)	R^2		✓		Real	Income prediction	Re-estimation
Lipovetsky and Conklin (2001)	R^2		✓		Real	Customer satisfaction	Re-estimation
Moehle et al. (2021)	Performance metrics	✓	✓		Real	Portfolio management	Other
Owen and Prieur (2017)	Explained variance in ANOVA	✓	✓		None	None	Re-estimation
Sutera et al. (2021)	Mean Decrease of Impurity		✓	✓	Real	Led and digits problems	Other
Verdinelli and Wasserman (2024)	Mean Squared Error	✓	✓		Synthetic	Numerical experiments	Re-estimation
Zhang et al. (2023)	Loss functions	✓	✓		Real	Mortality and Tumor prediction	Other
Our paper	Performance metrics	✓	✓	✓	Real	Credit Scoring	marginalization

B Examples of performance metrics

Table A2: 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	$-\frac{1}{n}\sum_{i=1}^{n}\left\|y_i - \hat{f}(\mathbf{x}_i)\right\|$	$-\left\|y_i - \hat{f}(\mathbf{x}_i)\right\|$	$\emptyset$
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	$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i\hat{\delta}_{n_1}(\mathbf{x}_i) + (1-y_i)\hat{\delta}_{n_2}(\mathbf{x}_i)}{\hat{\delta}_{n_3}}\right)$	$\left(y_i\hat{\delta}_{n_1}(\mathbf{x}_i) + (1-y_i)\hat{\delta}_{n_2}(\mathbf{x}_i)\right)\hat{\delta}_{n_3}^{-1}$	$\hat{\delta}_{n_1}(\mathbf{x}_i) = \frac{1}{n}\sum_{j=1}^{n}(1-y_j)I(\hat{P}(\mathbf{x}_i) > \hat{P}(\mathbf{x}_j))$ $\hat{\delta}_{n_2}(\mathbf{x}_i) = \frac{1}{n}\sum_{j=1}^{n}y_jI(\hat{P}(\mathbf{x}_i) < \hat{P}(\mathbf{x}_j))$ $\hat{\delta}_{n_3} = \frac{2}{n^2}\sum_{j=1}^{n}y_j\sum_{j=1}^{n}(1-y_j)$

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 with regression models (Panel A) and those related to classification models (Panel B).

C XPER decomposition of the R^2 in a linear regression model

C.1 Interpretation of the benchmark

One of the main advantages of the XPER decomposition is the intuitive interpretation of the benchmark ϕ_0. This value represents the expected performance metric in a hypothetical scenario where the model is applied (without re-estimation) to a dataset in which the target variable y is independent of all the features $\mathbf{x}$.

When using the R^2 performance metric for a linear regression model, this benchmark value corresponds to $\phi_0 = -R^2$. This result arises from the fact that the model $\hat{f}(\mathbf{x})$ is treated as fixed and pre-estimated on a training sample, while the XPER decomposition is applied to the test sample. If we were to *estimate* the model $\hat{f}(\mathbf{x})$ on a hypothetical sample where y is independent of $\mathbf{x}$, the benchmark R^2 would be zero asymptotically. However, in our case, the model is estimated on the training set where y and $\mathbf{x}$ are correlated. We then apply this *pre-trained* model to a hypothetical test set with spurious regressors (i.e., where y is independent of $\mathbf{x}$), without re-estimating the model. This procedure yields a negative R^2, which reflects the opposite of the R^2 obtained on the test sample with real (correlated) data.

To illustrate this point, consider the linear regression model given by $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i$, where the parameters are set to $\{\beta_0, \beta_1, \beta_2\} = \{0, 1, 3\}$. The features $\mathbf{x}_i = (x_{1,i}, x_{2,i})$ are i.i.d. with $\mathbb{V}(x_j) = 1$, and the errors ε_i are i.i.d. following a normal distribution $\mathcal{N}(0, \sigma_\varepsilon^2)$ with $\sigma_\varepsilon^2 = 10$. We simulate a sample of size $n = 4,000$, using the first $3,000$ observations as the training sample and the remaining ones as the test sample. We estimate the model $y_i = \mathbf{x}_i\beta + \varepsilon_i$ using the training sample, resulting in the following parameter estimates $\{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2\} = \{-0.0500, 0.9767, 3.0166\}$. The corresponding R^2 value for the training sample is $R^2_{\text{train}} = 0.4994$.

Next, we use the estimated model to make predictions $\hat{y}_i$ on the test sample. We obtain an R^2 equal to $R^2_{\text{test}} = 0.5544$. Then, we compute the XPER values on the test sample, yielding the following

decomposition:

$$\underbrace{0.5544}_{R^2} = \underbrace{-0.4761}_{\hat{\phi}_0} + \underbrace{0.0948}_{\hat{\phi}_1} + \underbrace{0.9357}_{\hat{\phi}_2}.$$

To understand the interpretation of the benchmark value $\hat{\phi}_0$, we propose the following experiment. We simulate two spurious regressors z_1 and z_2, with $\mathbb{V}(z_j) = 1$, that are uncorrelated with the target variable y, and with the features x_1 and x_2.

1. We *estimate* a linear regression model $y_i = \gamma_0 + \gamma_1 z_{1,i} + \gamma_2 z_{2,i} + \mu_i$, where y_i corresponds to the simulated variable generated from the true DGP $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i$, on a sample of size $n = 3000$. We then apply this model on the test set of size $n = 1000$ and compute the R^2. As expected, we obtain a value close to 0, specifically $R^2_{\text{test}} = -0.0030$. The corresponding parameter estimates are close to zero with $\{\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2\} = \{-0.0202, -0.0424, 0.0939\}$.

2. We *apply* the estimated model obtained from the training sample, i.e., $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i}$, with $\{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2\} = \{-0.0500, 0.9767, 3.0166\}$ to a hypothetical test set where the model features are uncorrelated with the target. Formally, we set $x_1 = z_1$ and $x_2 = z_2$ and compute the fitted values $\hat{y}_i = -0.0500 + 0.9767 z_{1,i} + 3.0166 z_{2,i}$ using the parameters estimated from the training sample with the "true" regressors x_1 and x_2. It is as if the variables used to estimate the model in the training sample had suddenly become irrelevant in the test sample. The corresponding R^2 value is -0.4538, which is close to the benchmark estimate $\hat{\phi}_0$, and nearly the opposite of the test R^2 initially obtained on the test set. Indeed, the fact that the constant term in the regression is not re-estimated induces a negative R^2.

C.2 Interpretation of the individual XPER values

Consider a linear regression model and the R^2 as sample performance metric such that $y_i = \mathbf{x}_i^T \beta + \varepsilon_i$, with ε_i i.i.d $\mathcal{N}(0, 4)$. We consider three i.i.d. features such that $\mathbf{x}_i^T = (x_{i,1}, x_{i,2}, x_{i,3})$ and $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ with $diag(\mathbf{\Sigma}) = (8, 4, 1)$. The true vector of parameters is $\{\beta_0, \beta_1, \beta_2, \beta_3\} = \{0.2, 1, 1, 1\}$ with β_0 the intercept.

We simulate a sample of size $2,000$. We use the first $T = 1,000$ observations to estimate the model

Table A3: Illustration of R^2 XPER values in a three-fold standard linear model

	$G(y_i; \mathbf{x}_i; \hat{\delta}_n)$	$\hat{\phi}_0$	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$	y_i	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$
i = 1	0.7875	-0.3388	0.9287	-0.2303	0.4279	3.2871	1.3815	3.6312
i = 2	0.0151	0.2491	-0.2697	0.0454	-0.0097	-0.3773	-4.4793	16.8266
i = 3	0.5367	0.0830	-0.0161	0.1953	0.2745	-1.6622	1.1513	7.9157
i = 4	0.9992	-1.0697	1.8839	-0.0317	0.2166	4.8534	4.7354	0.0139
i = 5	0.8478	0.1834	0.4011	0.0691	0.1942	1.2401	-0.3725	2.6007
...	...	...	...	...	...	...	...	...
i = 996	0.4523	0.2511	0.1402	0.0947	-0.0336	-0.3395	-3.3984	9.3570
i = 997	0.8812	0.2219	0.2834	0.2266	0.1494	-0.7394	0.6851	2.0290
i = 998	0.9627	-1.1944	0.9645	0.6049	0.5876	-4.9032	-4.1044	0.6380
i = 999	0.9607	0.2520	0.3705	0.2684	0.0698	-0.3202	0.4997	0.6722
i = 1,000	0.9203	0.2449	0.5788	0.0221	0.0745	0.6174	-0.5495	1.3617
	0.7629	-0.7385	0.9649	0.4202	0.1163	0.1138	0.0843	4.0504

and the remaining ones as test sample S_n. Consider a simulation for which the estimated parameters are equal to $\{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3\} = \{0.1395, 1.0208, 0.9743, 1.0434\}$. The objective is to evaluate the contribution of each feature to the R^2 of the model $\hat{f}(\mathbf{x}_i) = \mathbf{x}_i^T \hat{\beta}$, computed on the test sample S_n. The estimated R^2 and feature contributions are the following:

$$\underbrace{0.7629}_{R^2} = \underbrace{-0.7385}_{\hat{\phi}_0} + \underbrace{0.9649}_{\hat{\phi}_1} + \underbrace{0.4202}_{\hat{\phi}_2} + \underbrace{0.1163}_{\hat{\phi}_3}.$$

As expected, the estimated benchmark $\hat{\phi}_0$ is negative. The difference between the sample R^2 and the benchmark is explained by the contribution of the features. We verify that the features contributing the most to the R^2 are, in decreasing order, x_1, x_2, and x_3. Given Equation (6), differences in ϕ_j values across features come from either $\hat{\beta}_j$ or σ_{y,x_j}. Here, the ranking of XPER values is the same as the ranking of the covariances of the feature with the target variable as $\sigma_{y,x_1} > \sigma_{y,x_2} > \sigma_{y,x_3}$.

Local analysis of feature contributions to the R^2 is detailed in Table A3. In the second column, we report the individual contributions to the R^2 on the test sample. By definition, the average of individual contributions corresponds to the (test) sample R^2 equal to 0.7629. Individuals for which the contribution $G(y_i; \mathbf{x}_i; \hat{\delta}_n)$ is larger than the R^2 are those for which the squared residual (reported in last column) is lower than MSE. For instance, individual $i = 1$ has a contribution to the R^2 (0.7875) larger than the sample R^2 (0.7629) as well as a squared residual (3.6312) lower than

the MSE (4.0504). Individual $i = 4$ contributes more to the R^2 than individual $i = 5$ ($0.9992 > 0.8478$) because its squared residual is smaller than the latter ($0.0139 < 2.6007$). Thus, contribution $G(y_i; \mathbf{x}_i; \hat{\delta}_n)$ measures the ability of the model (as measured by the R^2) to predict the target variable y for a given individual on the test sample.

Individual feature contributions to the R^2 are reported in columns 4, 5, and 6. A positive contribution $\phi_{i,j}$ means that the feature x_j tends to improve the predictive ability of the model for individual i, as measured by the R^2. For instance, for individual $i = 4$ feature values $x_{4,1}$ and $x_{4,3}$ (respectively 4.8040 and 0.5165) contributes to reduce its squared residual, so $\phi_{4,1}$ and $\phi_{4,3}$ are positive (respectively 1.8839 and 0.2166). On the contrary, the second feature value $x_{4,2}$ (-0.8694) reduces the contribution to the R^2 of this individual, i.e., $\phi_{4,2} = -0.0317 < 0$. The intuition behind this result is as follows: (i) The realization of feature x_2 is inferior to the average ($-0.8694 < 0.0100$) and its coefficient $\hat{\beta}_2$ in the model is positive. (ii) Hence, x_2 tends to decrease the prediction $\hat{y}_4$ compared to the average whereas we observe that the target value y_4 (4.8534) is larger than the average (0.1138). (iii) Therefore, feature x_2 increases prediction error and so hinders model performance.

One advantage of this local analysis is to reveal heterogeneity of the model predictive performance. The model better predict the target variable y for some individuals than for others. Our methodology allows to understand why the model is more or less suited for individuals in the test sample, through individual feature contributions $\phi_{i,j}$. The scatter plot in Figure A1 illustrates this heterogeneous analysis. We report the realizations of feature x_1 on the x-axis and target values on the y-axis. Blue (red) dots refer to positive (negative) individual feature contributions. We verify that the closer is the characteristic x_1 to its average, the less it contributes to the R^2, as indicated by the fading color of the dots. On the contrary, when the characteristic x_1 moves away from its expected value, individual feature contributions are either positive or negative. In order to understand the intuition behind this result let us assume that the target variable represents the consumption and the first feature corresponds to wages. In the estimated model, the consumption is positively correlated to wages as $\hat{\beta}_1 = 0.1395 > 0$. Therefore, according to the model, individuals with large wages (compared to the

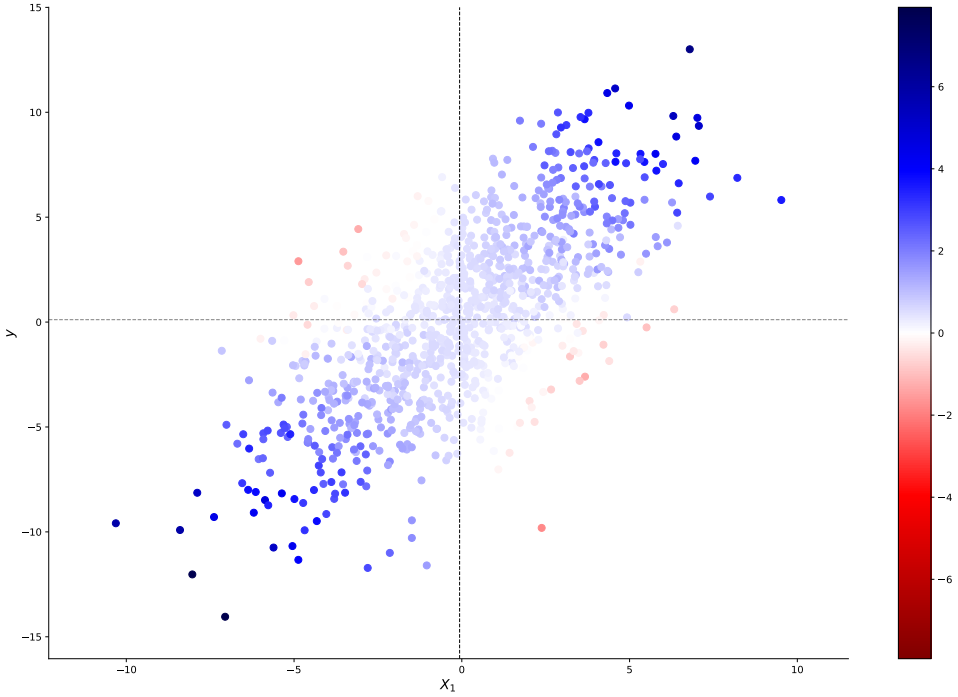


Figure A1: R^2 XPER values $\phi_{i,1}$ in a three-fold model

Note: This figure displays XPER values $\phi_{i,1}$ as a function of feature x_1 (x-axis) and target variable (y-axis). The vertical dotted line refers to the expected value of feature x_1 and the horizontal dotted line the one of the target variable.

average) are likely to have large consumption (compared to the average). Consider an individual with a large (low) wage and a large (low) consumption, i.e., an individual belonging to the top right (bottom left) panel. In this case, the wage contribution to the R^2 is positive (blue dots) as the observed consumption is consistent with the model prediction. On the contrary, if an individual has an above-average salary (top left panel), then this feature misleads the model and thus the corresponding XPER value of the wage turns negative (red dots).

Finally, the larger is the feature deviation from the average, the larger is its contribution to $G(y_i; \mathbf{x}_i; \hat{\delta}_n)$. This result is highlighted in Figure A1 by the darkest dots associated with individuals with both, large deviation from the average and large contribution to $G(y_i; \mathbf{x}_i; \hat{\delta}_n)$. Therefore, the larger the variance of the feature, the larger is its contribution, in absolute value, to the R^2. As illustrated by Figures A1 and A2, given that $\mathbb{V}(x_1) > \mathbb{V}(x_2)$, feature x_1 contributions are larger than feature x_2. The corresponding weights are reported in Table 1 in the paper.

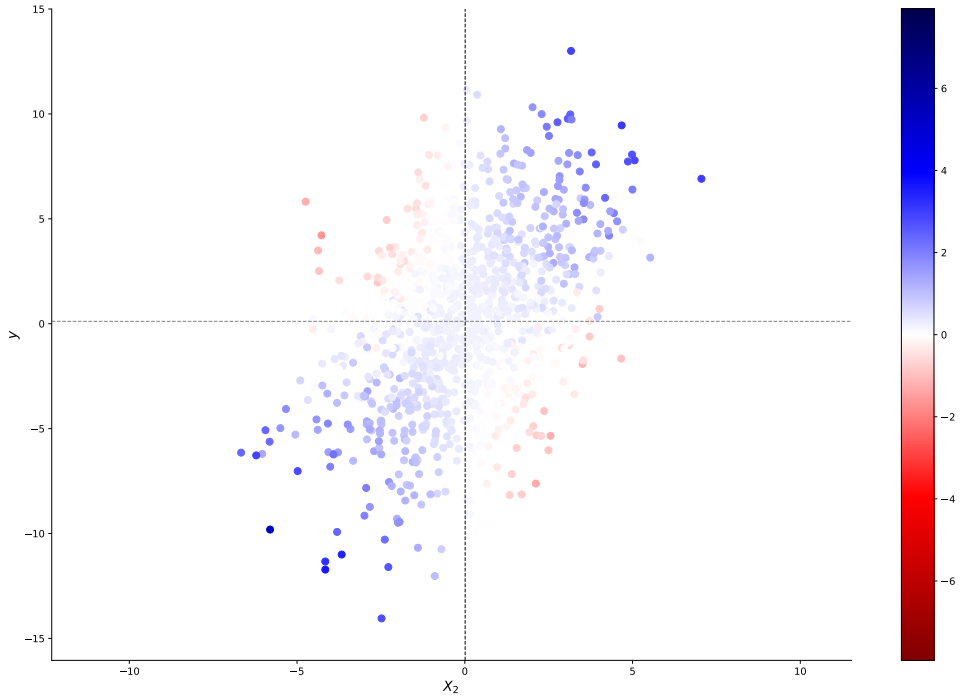


Figure A2: R^2 XPER values $\phi_{i,2}$ in a three-fold model

Note: This figure displays XPER values $\phi_{i,1}$ as a function of feature x_1 (x-axis) and target variable (y-axis). The vertical dotted line refers to the expected value of feature x_1 and the horizontal dotted line the one of the target variable.

D Examples of XPER values decomposition

We provide several examples in Table A4 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 $\mathbb{E}(\mathbf{x}) = 0_q$ and $\mathbb{V}(\mathbf{x}) = diag(\sigma_{x_j}^2)\ \forall j = 1, ..., q$, and $\mathbb{E}(y) = 0$. We denote by σ_y^2 the variance of the target variable and by σ_{y,x_j} the covariance between the feature x_j and the target variable. 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 A4: 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	$2\sum_{j=1}^{q}\hat{\beta}_j\sigma_{y,x_j}$ $-\sum_{j=1}^{q}\hat{\beta}_j^2\sigma_{x_j}^2 - \sigma_y^2$	$-\sum_{j=1}^{q}\hat{\beta}_j^2\sigma_{x_j}^2 - \sigma_y^2$	$2\hat{\beta}_j\sigma_{y,x_j}$
R^2	$\frac{\sigma_{y,\hat{y}}}{\sigma_y^2}$	$-\frac{\sigma_{y,\hat{y}}}{\sigma_y^2}$	$\frac{2\hat{\beta}_j\sigma_{y,x_j}}{\sigma_y^2}$

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 for 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 with regression models (Panel A) and those associated with classification models (Panel B). See Appendices K.3, K.4, and K.5 for the proofs related to the MSE, R^2, and accuracy.

E Estimation with a large number of features

E.1 A kernel-based approach

Computing the individual XPER value $\phi_{i,j}$, as defined in Definition 4, for a given feature x_j requires evaluating all possible coalitions of the other features of the model, $S \subseteq \mathcal{P}(\{\mathbf{x}\} \setminus \{x_j\})$. For a model with q features, this involves $2^{(q-1)}$ coalitions to compute $\phi_{i,j}$. To calculate the individual XPER values for all features $(\phi_{i,1}, \ldots, \phi_{i,q})$, a total of $q \times 2^{(q-1)}$ coalitions must be evaluated. As the number of coalitions grows exponentially with the number of features (e.g., $5,120$ for $q = 10$ and $10,485,760$ for $q = 20$), computing individual XPER values rapidly becomes computationally prohibitive. This scalability issue is not unique to our methodology but is a fundamental challenge for any approach based on Shapley values.

To address this issue, we have adapted the Kernel SHAP methodology proposed by Lundberg and Lee (2017) for estimating SHAP values. A key advantage of Kernel SHAP is its model-agnostic nature, making it applicable to any predictive model, unlike model-specific methods such as DeepSHAP or TreeSHAP (Lundberg et al., 2018). This approach significantly reduces computational complexity by sampling only a subset K of all possible feature coalitions. The approximate XPER values are then derived using a regression-based approximation, which ensures both practical feasibility and accurate estimation of feature contributions.

Specifically, we adapt this methodology to estimate the individual XPER values, $\phi_{i,0}, \phi_{i,1}, \ldots, \phi_{i,q}$, as defined in Definition 4. Recall that the individual XPER value $\phi_{i,j}$, for $j = 1, \ldots, q$, quantifies the contribution of model feature j for individual i to the performance metric, while $\phi_{i,0}$ represents the benchmark contribution for individual i. We approximate these values using a system of K equations derived from the performance metrics associated with the K selected coalitions of features. The estimation process is structured as a linear regression problem, where the individual XPER values are the parameters to be estimated:

$$G_k\left(y_i; \mathbf{x}_i\right) = \phi_{i,0} + \sum_{j=1}^{q} \phi_{i,j} z_{k,j} + \epsilon_k, \quad \text{for } k = 1, \ldots, K, \tag{A1}$$

where ϵ_k is an error term and:

$$G_k\left(y_i; \mathbf{x}_i\right) = \frac{1}{n}\sum_{u=1}^{n} G\left(y_i; \mathbf{x}_i^{S_k}, \mathbf{x}_u^{\overline{S}_k}; \hat{\delta}_n\right), \tag{A2}$$

represents the individual contribution to the sample performance metric associated with coalition S_k, and $z_{k,j}$ are binary variables indicating the presence ($z_{k,j} = 1$) or absence ($z_{k,j} = 0$) of feature x_j in coalition k. For instance, consider a regression model $f(x)$ with $q = 5$ features indexed by j, denoted as $x_1, \ldots, x_5$, for which we aim to estimate the XPER values associated with the mean squared error (MSE). Let i index an individual in the sample, represented by $(y_i, x_{i,1}, \ldots, x_{i,5})$. Now, consider a specific coalition of features, randomly chosen and indexed by $k = 1$, such that:

$$S_1 = \{x_{i,1}, x_{i,3}, x_{i,5}\}, \tag{A3}$$

$$z_{1,1} = 1, \quad z_{1,2} = 0, \quad z_{1,3} = 1, \quad z_{1,4} = 0, \quad z_{1,5} = 1. \tag{A4}$$

The set of features excluded from this coalition is $\overline{S}_1 = \{x_{i,2}, x_{i,4}\}$, and the individual contribution of observation i to the MSE for this coalition is computed as:

$$G_1(y_i, \mathbf{x}_i) = \frac{1}{n}\sum_{u=1}^{n} \left(y_i - f(x_{i,1}, x_{u,2}, x_{i,3}, x_{u,4}, x_{i,5})\right)^2. \tag{A5}$$

For a second coalition, $S_2 = \{x_{i,2}, x_{i,3}\}$, we have $z_{2,2} = 1$ and $z_{2,3} = 1$, with all other dummy variables set to 0. The individual contribution for this coalition is computed as:

$$G_2(y_i, \mathbf{x}_i) = \frac{1}{n}\sum_{u=1}^{n} \left(y_i - f(x_{u,1}, x_{i,2}, x_{i,3}, x_{u,4}, x_{u,5})\right)^2. \tag{A6}$$

By repeating this process for K coalitions, we obtain a series of individual contributions, $G_1(y_i, \mathbf{x}_i)$, $\ldots$, $G_K(y_i, \mathbf{x}_i)$, along with q series of dummy variables $\{z_{1,1}, \ldots, z_{K,1}\}$, $\ldots$, $\{z_{1,q}, \ldots, z_{K,q}\}$.

In a vectorial form, the system in Equation (A1) can be written as:

$$\mathbf{G}\left(y_i; \mathbf{x}_i\right) = \mathbf{Z}\phi_i + \epsilon, \tag{A7}$$

$$\underset{(q+1)\times 1}{\phi_i} = \begin{pmatrix} \phi_{i,0} \\ \phi_{i,1} \\ \vdots \\ \phi_{i,q} \end{pmatrix}, \quad \underset{K\times(q+1)}{\mathbf{Z}} = \begin{pmatrix} 1 & z_{1,1} & . . & z_{1,q} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & z_{K,1} & . . & z_{K,q} \end{pmatrix}, \quad \underset{K\times 1}{\mathbf{G}\left(\mathbf{y_i};\mathbf{x_i}\right)} = \begin{pmatrix} G_1\left(y_i;\mathbf{x}_i\right) \\ \vdots \\ G_K\left(y_i;\mathbf{x}_i\right) \end{pmatrix}. \tag{A8}$$

The vector of individual XPER values, ϕ_i, can be estimated using a least squares estimator. However, to accurately approximate the individual XPER values, it is necessary to account for (1) the weights of each coalition of features and (2) the fact that the estimation relies on a subset of coalitions of size K. Hence, we estimate Equation (A1) using a Weighted Least Squares (WLS) estimator as follows:

$$\hat{\boldsymbol{\phi}}_i = \underset{\phi_i}{\operatorname{argmin}} \sum_{k=1}^{K} \omega_{S_k} \left(G_k\left(y_i;\mathbf{x}_i\right) - \left(\phi_{i,0} + \sum_{j=1}^{q} \phi_{i,j} z_{k,j} \right) \right)^2,$$

where the weights ω_{S_k} are defined as:

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

with $|S_k|$ denoting the number of features included in coalition S_k. The estimates of the individual XPER values are then computed as:

$$\hat{\boldsymbol{\phi}}_i = \left(\mathbf{Z}'\boldsymbol{\Omega}\mathbf{Z}\right)^{-1}\mathbf{Z}'\boldsymbol{\Omega}\mathbf{G}\left(y_i;\mathbf{x}_i\right), \tag{A9}$$

where

$$\underset{(K\times K)}{\boldsymbol{\Omega}} = \begin{pmatrix} \omega_{S_1} & 0 & . . & 0 \\ 0 & \omega_{S_2} & . . & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & . . & \omega_{S_K} \end{pmatrix}.$$

Estimating the individual XPER values using WLS significantly reduces computational time by limiting the number of coalitions considered and allowing for the simultaneous estimation of contributions for all features. The steps for implementing this approximation are detailed in Algorithm 1.

Algorithm 1 Pseudo-code to estimate individual XPER values with a large number of features

1: Train the model $f(.)$ using the training sample $\{\mathbf{x}_i, y_i\}_{i=1}^{T}$.
2: Randomly select a coalition S_k from the set of all coalitions $\widetilde{\mathcal{P}}(\{\mathbf{x}\})$, and determine its complement $\overline{S}_k$.
3: Store the dummy variables $z_{k,1}, \ldots, z_{k,q}$.
4: Select an observation $i = 1$ from the test sample $\{\mathbf{x}_i, y_i, \hat{f}(\mathbf{x}_i)\}_{i=1}^{n}$.
5: Compute the individual contribution of observation i to the performance metric $G_k(y_i; \mathbf{x}_i)$ using S_k and $\overline{S}_k$.
6: Calculate the weight ω_{S_k} for the selected coalition S_k.
7: Repeat steps 2 through 6 K times to obtain K values for the individual contribution $G_k(y_i; \mathbf{x}_i)$ and the corresponding weights ω_{S_k}, for $k = 1, \ldots, K$.
8: Estimate the individual XPER values $\phi_{i,j}$ for $j = 1, \ldots, q$ for observation i using Weighted Least Squares.
9: Repeat steps 3 through 8 for all observations $i = 2, \ldots, n$.

Finally, the global XPER value $\hat{\phi}_j$ for feature x_j, as defined in Definition 3, is estimated as the empirical mean of the individual XPER values $\hat{\phi}_{i,j}$, as shown in Equation (15).

E.2 How similar are the exact and approximated XPER values?

In this section, we investigate the *accuracy* of the kernel-based method in approximating the true XPER values and the associated *computational time*. For that purpose, we design a simulation experiment in which we vary the number of model features. We consider a regression model with non-linear effects (specifically, quadratic terms) and non-normal, correlated features. To ensure comparability as the number of primary variables increases, we construct a DGP where the true normalized XPER value of each primary variable (expressed as a percentage) is set to $1/q$, regardless of the total number of primary variables q. Thus, we can compare the approximated XPER values obtained using the kernel-based method to this theoretical benchmark of $1/q$ across all models.

The DGP is given by a regression model with quadratic effects:

$$y_i = \beta_0 + \sum_{j=1}^{q} x_{i,j}\beta_j + \sum_{j=1}^{q} x_{i,j}^2\omega_j + \varepsilon_i, \quad i = 1, \ldots, T+n, \tag{A10}$$

where ε_i is an i.i.d. error term, with $\varepsilon_i \sim \mathcal{N}(0,1)$. We consider q correlated primary variables $x_{i,1}, \ldots, x_{i,q}$, with $x_{i,j} \sim Beta(2,5)$, $\forall j = 1, \ldots, q$, $\forall i = 1, \ldots, T+n$. The dependence structure is

modeled through a Gaussian copula with mean vector zero and correlation matrix $\boldsymbol{\Sigma}$ defined as:

$$\underset{(q,q)}{\boldsymbol{\Sigma}} = \begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{pmatrix},$$

with $\rho = 0.3$. We assume that each pair of parameters (β_j, ω_j) is identical across all primary variables, setting $\beta_j = a\theta$ and $\omega_j = \theta$ for $j = 1, \ldots, q$, where a is a constant. In particular, we choose $\theta = 1$ and $a = 2$. The main advantage of this design is that the normalized XPER value associated with each primary variable x_j is equal to $1/q$ (see the proof in Appendix K.6):

$$\tilde{\phi}_j = \frac{1}{q}, \quad \forall j = 1, \ldots, q. \tag{A11}$$

Let K denote the number of coalitions selected to estimate the kernel-based approximation of the XPER values, with $K < q \times 2^{(q-1)}$, where $q \times 2^{(q-1)}$ represents the total number of possible coalitions required to compute the exact XPER values. Two alternative approaches can be used to set the number of coalitions K. The first consists of defining K as a fixed proportion of the total number of coalitions, for example, 25% or 50%. However, this approach quickly becomes impractical as the number of primary variables q increases, typically for $q > 15$, due to the exponential growth in the number of coalitions and the resulting computational burden. The second approach consists of fixing K independently of the total number of coalitions. For instance, Lundberg and Lee (2017) propose the rule $K = 2048 + 2q$ in their SHAP package available on GitHub. This rule has the clear advantage of keeping the computational cost at a reasonable level, regardless of the model dimension.

Thus, to assess the accuracy of the kernel-based approximation, we compare the estimated XPER values under different choices for K: 25%, 50%, and the fixed rule proposed by Lundberg and Lee (2017). For a model with q primary variables, we draw a subset of K coalitions and compute the XPER values using the kernel method on a test sample. Since the accuracy of the approximation

depends on the specific coalitions selected, we repeat the process $S = 100$ times. We denote by $\hat{\phi}_j^{q,s}(K)$ the estimated XPER values of feature x_j in the model including q primary variables x_j, and for the specific s-th random draw of K coalitions. The quality of the kernel-based approximation is assessed by comparing the estimated XPER values $\hat{\phi}_j^{q,s}(K)$ to the theoretical XPER values $\tilde{\phi}_j = 1/q$, using the Mean Absolute Prediction Error (MAPE) defined as:[17]

$$\text{MAPE}(K) = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{q}\sum_{j=1}^{q}\left|\frac{\hat{\phi}_j^{q,s}(K) - 1/q}{1/q}\right|. \tag{A12}$$

Intuitively, as the number of coalitions K increases, the MAPE decreases, reflecting a more accurate approximation, but at the expense of increased computational time.

To limit approximation errors beyond those specifically induced by the kernel-based method, we assume that the parameters β_j and ω_j of the regression model (A10) are known. Hence, we remove the source of variability related to the estimation or training of the model. We consider models with a number of primary variables q ranging from 7 to 20. Then, for each model, we simulate a single test sample $\{y_i^q, \mathbf{x}_i^q\}_{i=1}^n$ of size $n = 2,000$. This test sample is used to evaluate the model's performance through the R^2 and to compute the XPER values, which are estimated using the kernel-based method.[18]

Figures A3 and A4 report the average MAPE and the computational time (in minutes) associated with the kernel-based XPER estimates.[19] Several key findings emerge. First, the approximation errors, measured by the MAPE, remain relatively small across all tested values of K, ranging from 2.9% to 5.7%. This suggests that the statistical instability due to the kernel-based approximation does not compromise the stability of the XPER estimates, at least in this experimental setting.

[17] Alternatively, one could compute empirical XPER estimates based on the full set of feature coalitions, as defined in Equation (13). However, we rely on the theoretical true values $\tilde{\phi}_j = 1/q$ as a benchmark, since estimating XPER values using all $q \times 2^{(q-1)}$ coalitions for each observation in the test set becomes computationally prohibitive for $q > 10$.

[18] Since the estimation of the performance is performed on a finite test sample, it introduces approximation errors. However, these errors remain relatively small with a test sample of size $n = 2,000$. Consequently, although finite test sample effects cannot be entirely eliminated, their impact is expected to be limited in the context of our evaluation of the kernel-based method's accuracy.

[19] Computational times were obtained by running the Python scripts on a Windows 11 Education laptop equipped with an Intel(R) Xeon(R) w9-3475x 2.21 GHz processor and 256 GB of RAM.

These results are also consistent with Lundberg and Lee (2017), who demonstrate the accuracy of the kernel-based approximation for SHAP values.

Second, as expected, for a fixed number of primary variables q, increasing the number of sampled coalitions reduces the approximation error. For instance, with $q = 7$ (implying a true XPER value of $1/q \approx 0.143$), using 25% of the coalitions results in a MAPE of approximately 5.7%, corresponding to an average absolute deviation of 0.0081. Increasing the proportion of sampled coalitions to 50% lowers the MAPE to 4.5%, or an average absolute deviation of 0.0064, demonstrating the anticipated improvement in accuracy as more coalitions are considered.

Third, increasing the number of sampled coalitions reduces the approximation error but it comes at a significant computational cost. The results clearly illustrate the exponential growth in computational time when K is set as a fixed proportion of all possible coalitions. For instance, at $q = 15$, the computational time exceeds 100 minutes under the 50% strategy, which can become prohibitive in practical applications. By contrast, the linear rule $K = 2048 + 2q$ keeps computation times stable and below 5 minutes, even for $q = 20$ primary variables, demonstrating its clear advantage for high-dimensional models.

Finally, the linear rule achieves a favorable trade-off between approximation accuracy and computational cost. The approximation error remains around 4.5%, even for large values of q, while computational time stays reasonable. Notably, with $q = 20$, the linear rule relies on only about 2% of all possible coalitions yet achieves a MAPE as low as 3.5%.[20] Based on these results, we recommend this rule of thumb for setting the number of sampled coalitions K.

[20] The fluctuations observed in the MAPE are likely due to finite-sample effects when estimating model performance on a single test set.

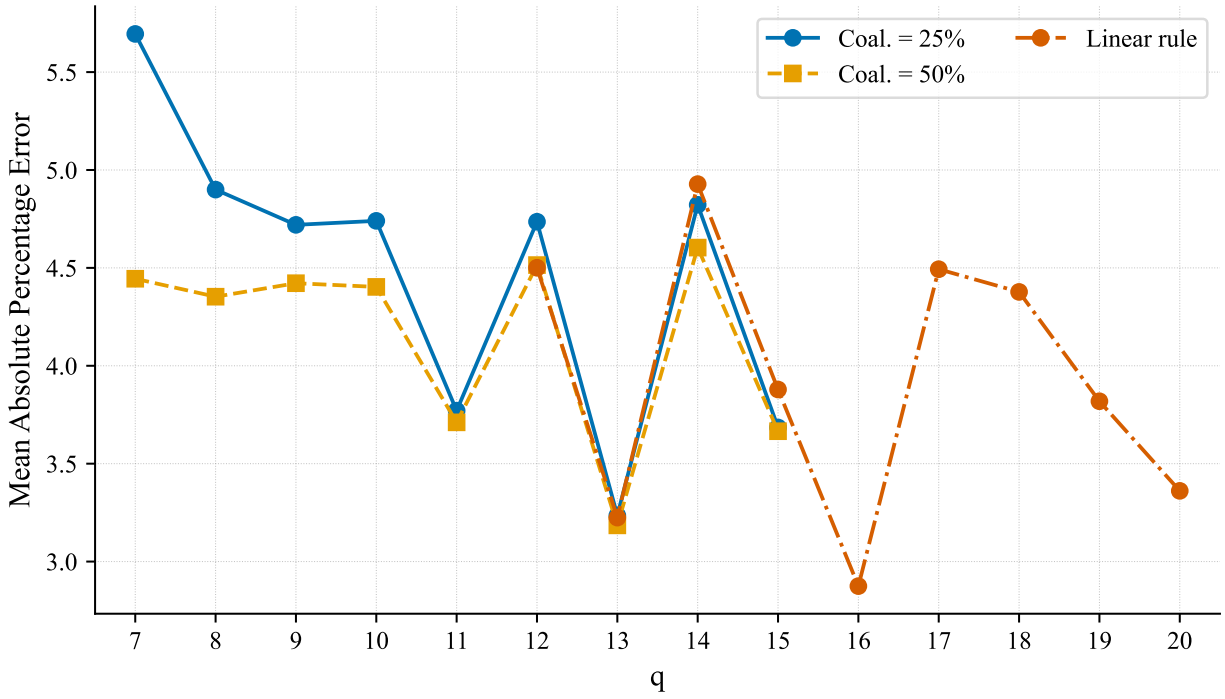


Figure A3: Mean Absolute Percentage Error

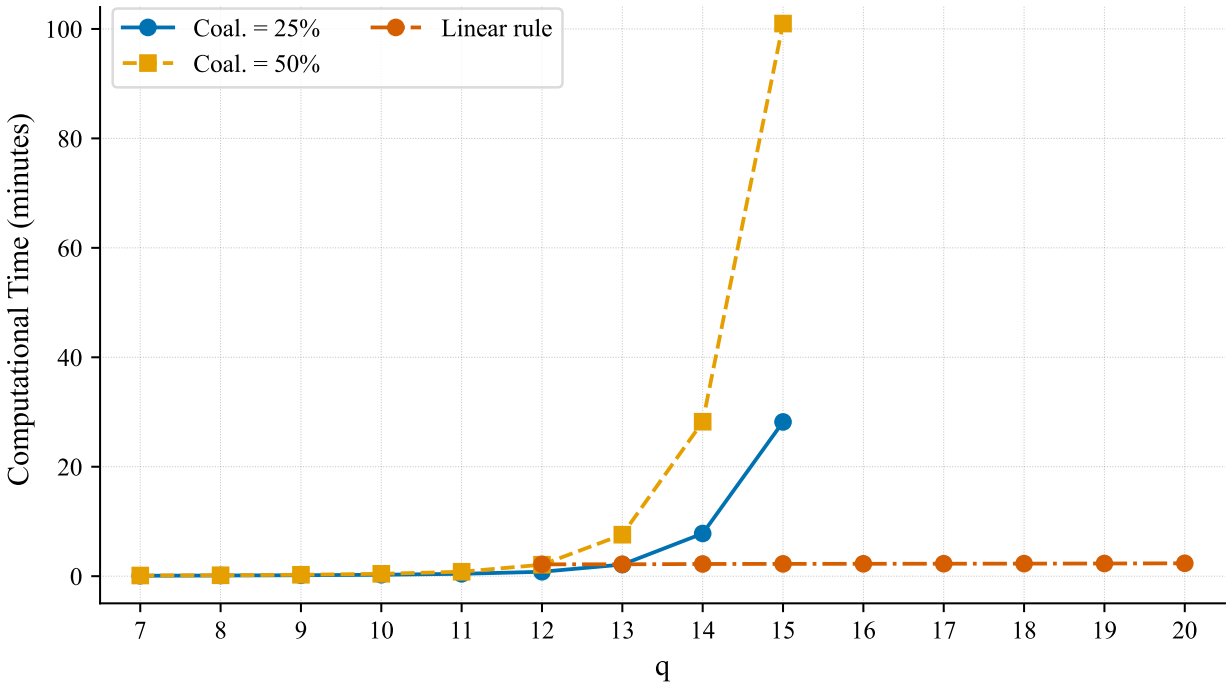


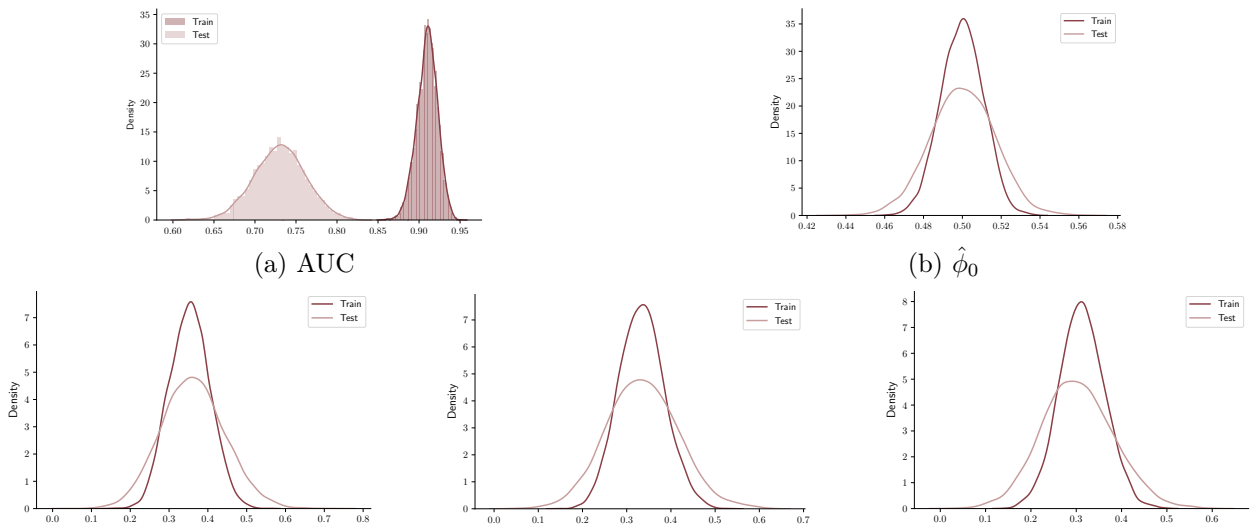
Figure A4: Computational time

F Simulations: Explaining overfitting

In Appendices 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 test samples. We display in Figure A5a 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 A5 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) $\hat{\phi}_3$

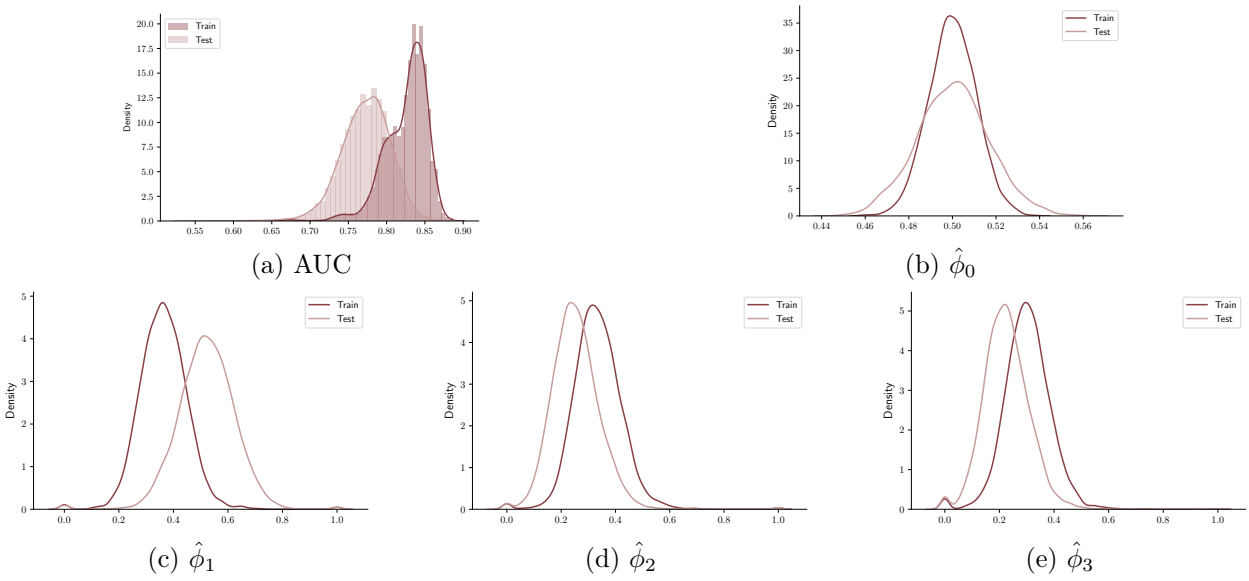
Figure A5: Empirical distributions of AUC and XPER values in case of overfitting due to improper control of the bias-variance trade-off

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 $diag(\tilde{\mathbf{\Sigma}}) = (3, 1.2, 1.1)$. 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.

In Figure A6a, 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



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

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

Figure A6: Empirical distributions of AUC and XPER values in case of overfitting due to a shift of the distribution of the features

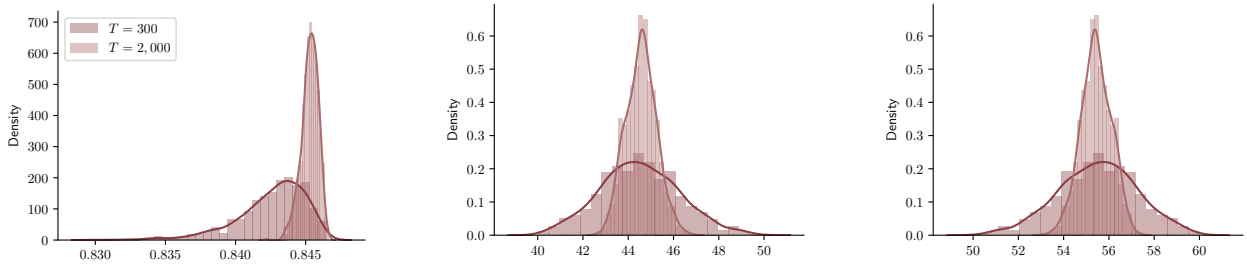
Note: This figure displays the empirical distributions of the AUC and XPER values on the training (dark color) and test (light color) sample according to the framework detailed in Illustration 2, case 2. XPER values are divided by the difference between the AUC of the model and the benchmark value to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

on XPER values. More precisely, in Figure A6, 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 Simulations: Stability of the XPER decomposition

As discussed in Section 6.2, when estimating the XPER values, *three main sources* of noise can arise, each of which may potentially affect the statistical stability of the XPER decomposition: (a) the estimation of the model on the training dataset, (b) the estimation of the performance metric on the test dataset, (c) the approximation of the XPER value by the kernel-SHAP approach. In this section, we study the effect of the first source of noise (source (a)) on the XPER decomposition. For that purpose, we keep the same DGP as the one described in Section 6.2.

In this experiment, we re-estimate the model on multiple simulated *training datasets* and decompose the R^2 computed on the same fixed test sample to assess the variability of the XPER values caused by source (a). To avoid cumulating multiple sources of variability, we design the simulation study as follows. The model is re-estimated on $S = 1{,}000$ independent training samples $\{y_i^s, \mathbf{x}_i^s\}_{i=1}^T$ of size $T \in \{300,\ 2{,}000\}$, and used to forecast the values of y on a unique test sample $\{y_i, \mathbf{x}_i\}_{i=1}^n$ of size $n = 700$, thereby mitigating error (b). For each replication s, we compute the R^2 on this fixed test sample, along with the corresponding exact XPER values based on all possible coalitions, which eliminates approximation error (c). Figure A7 displays the empirical distributions of the R^2 and the XPER values associated with the primary variables.



(a) R^2 (b) $\hat{\phi}_1$ (%) (c) $\hat{\phi}_2$ (%)

Figure A7: Empirical distributions of R^2 and XPER values associated to primary variables

Note: This figure displays the empirical distributions of the R^2 and XPER values for the two primary variables x_1 and x_2. These values are estimated from a single pseudo test sample of size $n = 700$. The solid red lines refer to kernel density estimates.

As expected, increasing the training sample size reduces the variability of the model parameter estimates, which in turn also reduces the variability of the empirical R^2. As T tends to infinity, these estimates converge to their theoretical counterparts. Consequently, the variability of the XPER estimates decreases as well. As in the experiment detailed in Section 6.2, the noise arising from the training of the model does not appear to compromise the stability of the XPER decomposition for the primary variables. For instance, our results show that the XPER value for the second primary variable consistently exceeds that of the first. This suggests that the ranking of the variables in terms of their contribution to the performance metric remains unaffected by the noise introduced by the training sample.

H Empirical application: Some additional results

H.1 Summary statistics and features distribution

Table A5: 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	10,383	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.

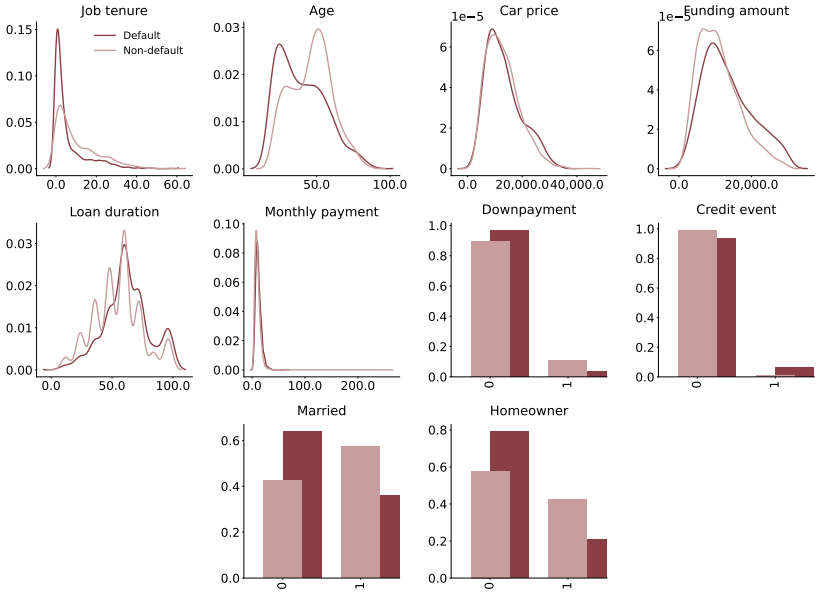


Figure A8: Features distribution by default class

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.

H.2 XPER decomposition for various predictive performance metrics

In this section, we show that we can easily calculate and compare XPER values across different performance metrics. To illustrate this, we have applied XPER to decompose all the classification performance metrics listed in Table A2: AUC, Brier score, accuracy, balanced accuracy, sensitivity, specificity, and precision. Figure A9 displays the XPER values for these seven metrics, highlighting several key insights.

First, some model features have contrasting effects on different metrics. For instance, *Car price* negatively impacts sensitivity on the test set, as indicated by its negative XPER value. This suggests it reduces the model's ability to correctly identify true credit defaults. However, this feature improves precision, which measures the proportion of true positive predictions out of all positive predictions. This example demonstrates how XPER values offer a nuanced understanding of how features affect performance across various metrics.

Second, the XPER values (expressed as percentages) for accuracy, balanced accuracy, sensitivity, and specificity are exactly equal. This equivalence can be attributed to the shared underlying structure of these metrics. As detailed in Table A4, the differences between these metrics and their benchmarks depends on the covariance between the target variable and the predictions, $Cov(y, \hat{f}(x))$, along with a normalization factor independent of the features. For example, this normalization factor is $1/\mathbb{P}(Y = 1)$ for sensitivity and $1/\mathbb{P}(Y = 0)$ for specificity (see Appendix K.7 for more details). Thus, XPER measures the contribution of features to this normalized covariance, explaining why the values are identical when expressed as percentages, as the normalization factor cancels out.

Lastly, regardless of the metric considered, two features —*Funding amount* and *Job tenure*— account for a significant portion of the model's predictive performance, explaining at least 56% of the difference between the performance and its benchmark value. Beyond these top-ranked features, the ranking of other features varies across metrics. For example, *Car price* ranks as third and fourth for AUC and precision, respectively, but is either the last or second-to-last for other metrics. Similarly, *Age* ranks fourth for AUC and ninth for precision, highlighting the diverse influence features can

have across different metrics.

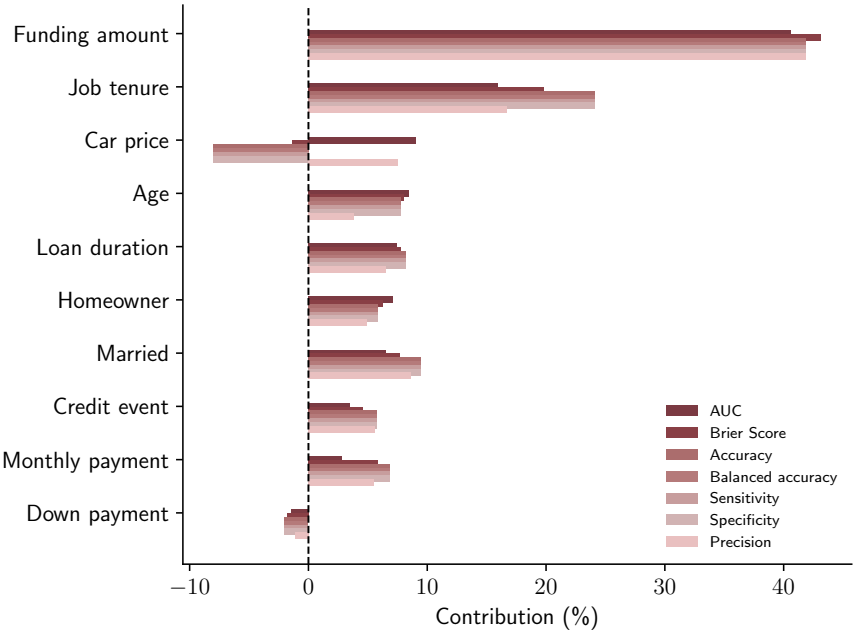


Figure A9: XPER decompositions for seven predictive performance metrics

Note: This figure displays the XPER values of seven classification performance metrics for the XGBoost model: (1) AUC, (2) Brier Score, (3) Accuracy, (4) Balanced accuracy, (5) Sensitivity, (6) Specificity, and (7) Precision. The results are obtained on the test sample.

H.3 XPER decomposition for imbalanced datasets

The issue of imbalanced data is well known in machine learning, particularly in credit scoring, as it affects both the learning capacity of models (predictions tend to favor the majority class) and the validity of certain evaluation metrics (e.g., accuracy, sensitivity, specificity). However, these are not the only consequences. A recent study by Chen et al. (2024) highlights another issue: the instability of interpretability methods such as LIME and SHAP under extreme class imbalance. Specifically, they demonstrate that feature importance rankings generated by these methods become unstable as class imbalance increases, and they also observe greater variability in the absolute SHAP values for the same feature in low-default scenarios.

Recognizing the importance of this issue, we now show how XPER can be used and behaves for imbalanced datasets. First, we illustrate how the model's performance and its XPER decomposition vary when the average default rate changes from 5% to 30% by applying undersampling to the

initial dataset. Second, we apply XPER decomposition to sensitivity and specificity, and compare the results with those obtained for the AUC (see Figure 3a).

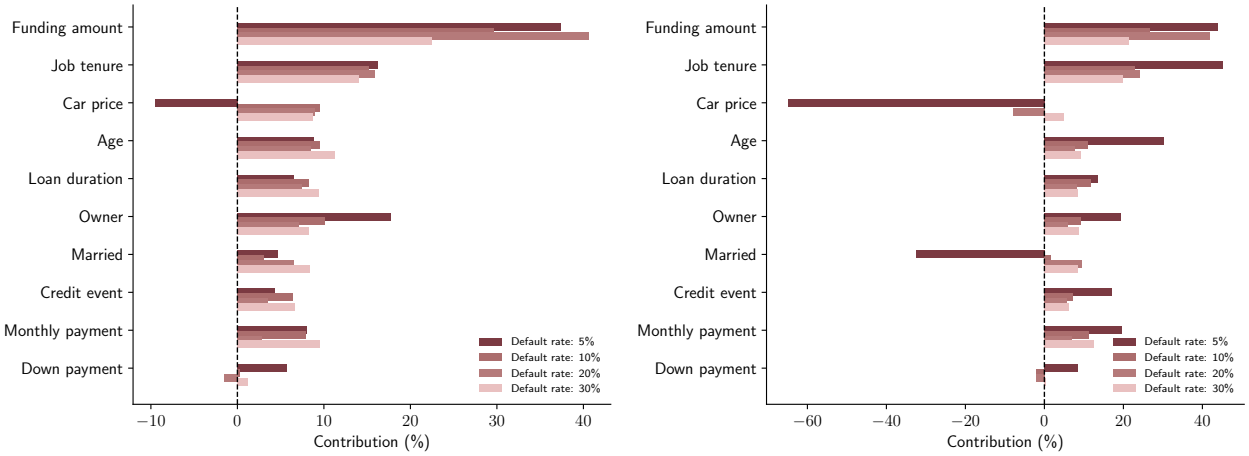
Varying the default rate. We report in Table A6 and Figure A10a, the AUC and the corresponding XPER decomposition obtained for our credit scoring empirical application when the average default rate is equal to 30%, 20%, 10%, and 5%. To decrease the default rates from the initial sample (20%), we implement random undersampling, which involves removing some default observations from the initial sample at random. Conversely, to increase the default rate to 30%, we remove non-default observations from the sample. Our findings are as follows: (i) the AUC varies between 0.7154 and 0.7884, (ii) significant variations in XPER values are observed when the default rate is as low as 5%, compared to the other rates. Interestingly, even when the default rates are set to 5% and 30%, producing similar AUC values (0.7174 vs. 0.7154), the XPER decompositions differ significantly. This discrepancy arises because two distinct models can achieve the same AUC: (i) one that accurately predicts most defaults but performs poorly on non-defaults, and (ii) one that performs well on non-defaults but poorly on defaults. This demonstrates why it is essential to also evaluate the model's sensitivity and specificity.

Table A6: Model performances for different default rates

Default Rate	AUC	Brier Score	Accuracy	BA	Sensitivity	Specificity
5%	0.7174	0.0479	94.89	50.45	1.06	99.83
10%	0.7884	0.0771	90.53	61.66	25.50	97.82
20%	0.7521	0.1433	79.53	58.69	23.99	93.39
30%	0.7154	0.195	70.83	61.99	39.91	84.07

Note: This table displays the performance metrics of an XGBoost model trained on datasets with default rates varying from 5% to 30%. BA stands for Balanced Accuracy.

Sensitivity and specificity. We also decompose the sensitivity and specificity with XPER across different imbalance rates, ranging from 5% to 30%. We compare these results with the decomposition obtained for the AUC and report the performance metrics in Table A6. As expected, the model's sensitivity significantly decreases as the dataset becomes more imbalanced, dropping from 25.50% at a 10% default rate to just 1.06% at a 5% default rate. Conversely, the model's specificity increases



(a) AUC

(b) Sensitivity/Specificity

Figure A10: XPER decomposition of the AUC and Sensitivity/Specificity for multiple default rates

Note: This figure displays the XPER values of the AUC (Panel (a)) and Sensitivity/Specificity (Panel (b)) considering four different test samples with default rates of 5%, 10%, 20%, and 30%. The results for the sample with a 20% default rate for the AUC correspond to those shown in Figure 2a. Note that the results in Panel (b) correspond to both Sensitivity and Specificity, as we obtain identical XPER values (in %) for these two performance metrics.

as the default rate decreases, rising from 84.07% at a 30% default rate to 99.83% at a 5% default rate. This stark variation in sensitivity and specificity suggests that the underlying models trained on datasets with different default rates are fundamentally distinct. Consequently, we anticipate significant variations in XPER values for sensitivity and specificity, particularly at the lower default rate of 5%.

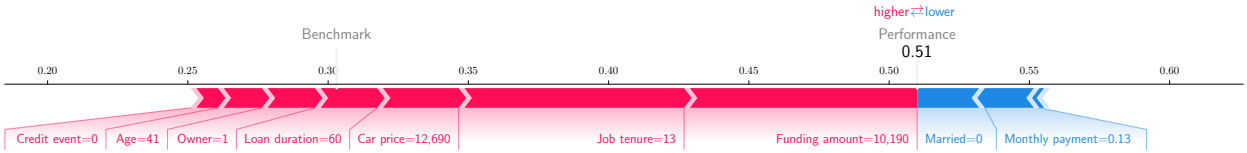
We present the XPER values (in %) for sensitivity and specificity in Figure A10b. Since these values are identical for both metrics, we display only one XPER value for each feature (see Appendix K.7 for further details). As expected, at a default rate of 5%, the XPER values differ substantially from those at higher default rates. For instance, at a 5% default rate, the *Car price* feature drastically decreases the model's sensitivity/specificity ($\hat{\phi} < -60\%$), whereas its contribution is relatively negligible ($\hat{\phi} < -5\%$) for the other default rates. It is worth noting that while sensitivity and specificity vary across default rates of 10%, 20%, and 30%, these variations are much less pronounced compared to the significant differences observed at a default rate of 5%.

H.4 Individual XPER decomposition and data visualization

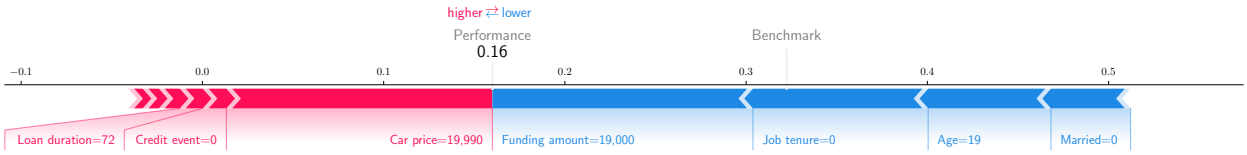
We analyze the impact of the various features on the performance metric but we now do it for each borrower individually. We start by analyzing in Figure A11 the XPER decomposition for two sample borrowers. These force plots enable us to decompose the individual performance of each borrower, as defined in Equation (11). 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 A11, *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 (11). 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*).

To better understand the relative influence of each feature for the two borrowers, we analyze 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.

We then consider the entire sample of borrowers. In Figure A12, we display the XPER values



(a) Borrower #3



(b) Borrower #28

Figure A11: 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 model (Panel (a)). Borrower #28 did not default on his loan and has a probability of default of 57% according to the XGBoost model (Panel (b)). *Performance* refers to the individual level of the AUC whereas *Benchmark* represents the individual contribution to the AUC associated with 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.

for each feature as a function of the feature value. We analyze 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 with 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.

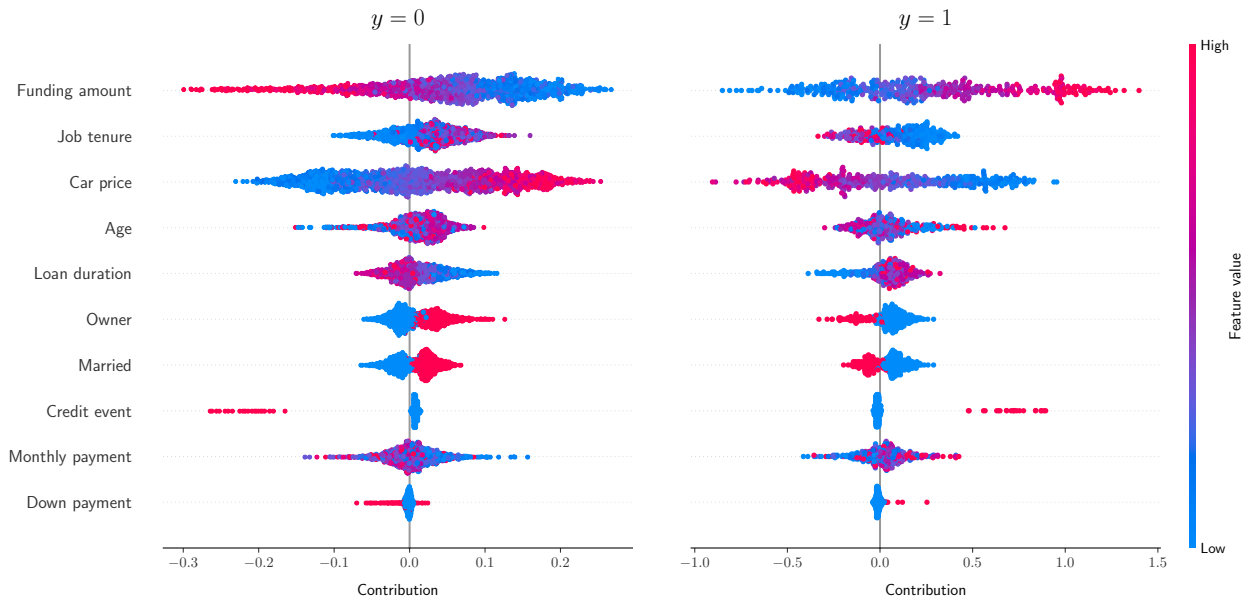


Figure A12: Summary plots of individual XPER values

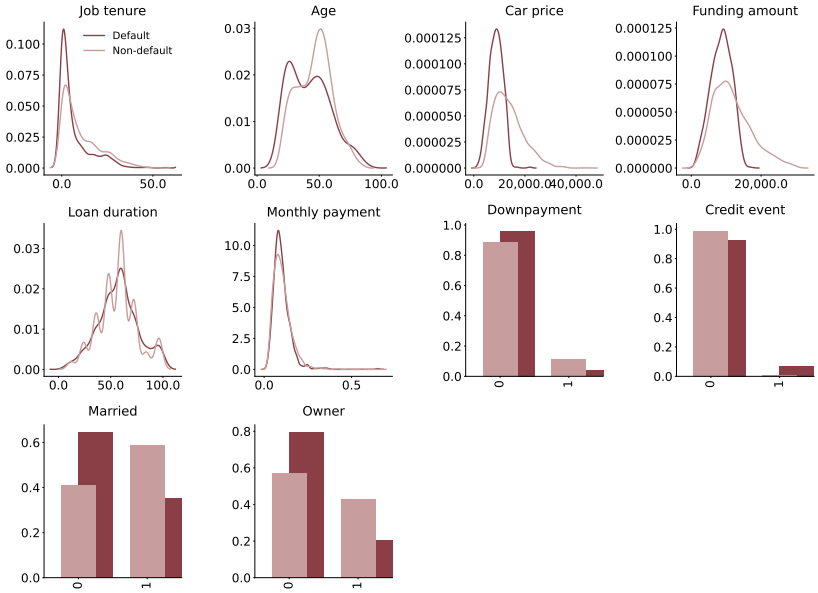
Note: This figure displays the individual XPER values for each feature used in the XGBoost model. Each dot represents the value for a given borrower (see Equation (4)). We display the results for the borrowers who paid back their loans (left panel) and those who do not (right panel).

H.5 Boosting model performance

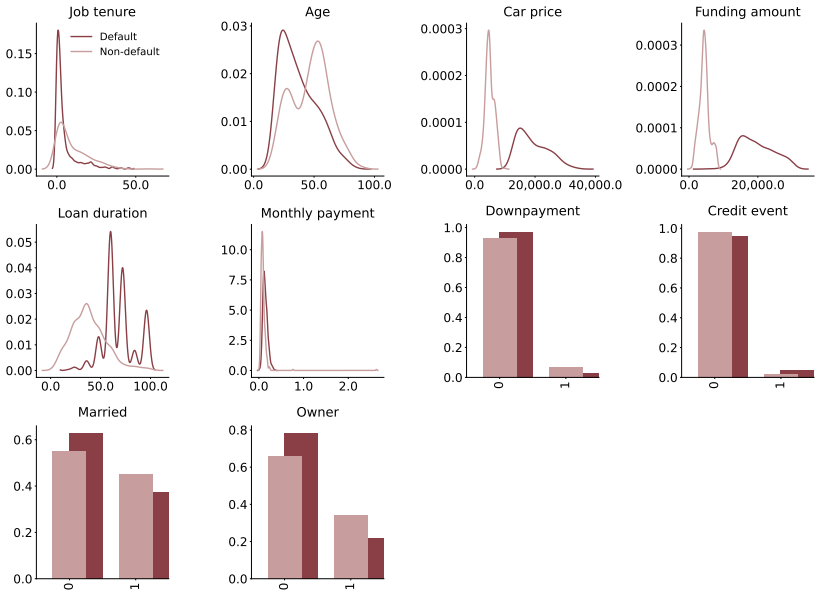
Table A7: Silhouette scores for the clustering based on XPER values and features

Number of Clusters	XPER clustering	Feature clustering
2	0.2037	0.2129
3	-0.0244	0.1846
4	-0.0608	0.1718
5	-0.1858	0.1609
6	-0.0624	0.1589
7	-0.0734	0.1595
8	-0.0596	0.1501
9	-0.1027	0.1514
10	-0.0665	0.1471

Note: This table presents the silhouette scores obtained for different numbers of clusters when clustering based on XPER values or feature values. The silhouette score ranges from -1 (worst) to 1 (best).



(a) Features distribution by default class in group 1



(b) Features distribution by default class in group 2

Figure A13: 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 K-Medoids methodology. For continuous features, we use a kernel density estimation. Dark red refers to defaulting borrowers and light red to non-defaulting borrowers.

Figure A14: 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 K-Medoids method.

I XPER vs. Permutation Importance

In this section, we compare the XPER method with the Permutation Importance (PI) method (Breiman, 2001; Fisher et al., 2019). Both methods aim to measure the effect of the features on model performance. Therefore, it is important to choose between PI and XPER. In our opinion, there are two key differences between these approaches. First, XPER satisfies the axioms of the Shapley value, unlike PI. These axioms, such as efficiency and the null effects axiom, simplify the interpretation of XPER and ensure its relevance. For example, the sum of the contributions obtained from PI lacks a specific meaning, except in very restricted cases. Second, in the performance context, PI is designed only at the global level and does not account for individual-level analysis, thus excluding the possibility of conducting heterogeneity analysis as done in Section 7.3.

More fundamentally, XPER encompasses PI. Specifically, it can be shown that XPER values can be expressed as the sum of two terms, one of which is the (normalized) PI. To derive this result, let us formally define PI. Initially developed by Breiman (2001) and later generalized by Fisher et al. (2019), PI corresponds to the difference between the model's performance with the original feature values ("original" performance) and its performance when the feature values are randomly permuted ("permuted" performance). Formally, the "original" performance of the model is given by the population performance metric $PM = \mathbb{E}_{y,\mathbf{x}}\left(G(y,\mathbf{x};\delta_0)\right)$. The "permuted" performance metric associated with a feature x_j is $PM_{switch,j} = \mathbb{E}_{y,\mathbf{x}_{-j}}\mathbb{E}_{x_j}\left(G(y,\mathbf{x};\delta_0)\right)$, with $\mathbf{x}_{-j}$ the vector of features excluding the feature x_j. This metric represents the expected performance of the model when x_j is replaced with random noise, rendering x_j uninformative about y, while keeping the marginal distribution of x_j unchanged (Fisher et al., 2019). By taking the difference between the two terms, we obtain the PI value for feature x_j:

$$PI_j \equiv PM - PM_{switch} = \mathbb{E}_{y,\mathbf{x}}\left(G(y,\mathbf{x};\delta_0)\right) - \mathbb{E}_{y,x^S,x^{\overline{S}}}\mathbb{E}_{x_j}\left(G(y,\mathbf{x};\delta_0)\right). \tag{A13}$$

Given this definition, it is possible to show that the XPER value encompasses the PI value, as we

have (see Appendix K.8 for the proof):

$$\phi_j = \sum_{S \subseteq \mathcal{P}(\{\mathbf{x}_{-j}\}) \setminus \{\mathbf{x}_{-j}\}} \omega_S \left[\mathbb{E}_{y,x_j,\mathbf{x}^S} \mathbb{E}_{\mathbf{x}^{\overline{S}}} \left(G\left(y; \mathbf{x}; \delta_0\right)\right) - \mathbb{E}_{y,\mathbf{x}^S} \mathbb{E}_{x_j,\mathbf{x}^{\overline{S}}} \left(G\left(y; \mathbf{x}; \delta_0\right)\right) \right] + \frac{PI_j}{q}, \quad \text{(A14)}$$

where q denotes the number of model features, and $\mathcal{P}(\{\mathbf{x}_{-j}\}) \setminus \{\mathbf{x}_{-j}\}$ represents the set of all possible feature coalitions excluding the feature x_j and discarding the coalition with all features except x_j. For example, in a model with three variables, and for $j = 1$, this set includes the coalitions $\{\emptyset\}$, $\{x_2\}$, and $\{x_3\}$. The PI corresponds to the gain in performance associated with adding feature x_j to the coalition of all the model features (excluding x_j), corresponding to $\{x_2, x_3\}$ in this example. Equation (A14) illustrates the similarity between PI and XPER values. Both metrics assess the impact of adding a given feature x_j to a pre-existing coalition of features on the model's performance metric. The key difference is that PI evaluates this impact for only *one specific coalition* that includes all features except x_j. In contrast, XPER takes into account all possible feature coalitions, satisfying the axioms of a Shapley value. As a result, PI and XPER coincide in a linear regression model, but generally differ in nonlinear models, as illustrated in Figure 3b of the empirical application.

To illustrate this simple case where XPER and PI deliver the same results, we conduct a Monte Carlo experiment. The DGP is defined by $y_i = x_{i,1}\beta_1 + x_{i,2}^3\beta_2 + \varepsilon_i$, with ε_i is an i.i.d. error term, and $\mathbf{x}_i = (x_{i,1}, x_{i,2})'$ two i.i.d. features. We assume that $\varepsilon_i \sim \mathcal{N}(0,1)$, $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ with $diag(\Sigma) = (1,1)$, and $\beta = (2,1)'$. We simulate $K = 1,000$ pseudo-samples $\{y_i^s, \mathbf{x}_i^s\}_{i=1}^{T+n}$ of size $4,000$ for $s = 1, \ldots, K$. For each pseudo-sample, we use the first $T = 2,000$ observations to estimate a linear regression model involving $x_{i,1}$ and $x_{i,2}^3$, and the remaining $n = 2,000$ observations to compute the R^2 and the corresponding XPER and PI values. Over the $1,000$ replications, the average R^2 obtained is equal to 0.9492, which is extremely close to its theoretical value (0.95). A similar result is obtained for estimated coefficients as the average values are equal to $\{\hat{\beta}_1, \hat{\beta}_2\} = \{2.0010, 0.9999\}$. On average, we obtain the following XPER values:

$$\phi_0 = -0.9496, \ \phi_1 = 0.4059, \ \phi_2 = 1.4930.$$

We obtain similar contributions with the PI, as on average:

$$PI_1 = 0.4053,\ PI_2 = 1.4925.$$

J 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:

$$\phi_{i,j}^{SHAP} = \sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\} \setminus \{x_j\})} w_S \left[\mathbb{E}_{\mathbf{x}^{\overline{S}}} \left(\hat{f}(x_{i,j}, \mathbf{x}_i^S, \mathbf{x}^{\overline{S}}) \right) - \mathbb{E}_{x_j, \mathbf{x}^{\overline{S}}} \left(\hat{f}(x_j, \mathbf{x}_i^S, \mathbf{x}^{\overline{S}}) \right) \right], \tag{A15}$$

$$\hat{f}(\mathbf{x}_i) = \phi_{i,0}^{SHAP} + \sum_{j=1}^{q} \phi_{i,j}^{SHAP}, \tag{A16}$$

$$\phi_{i,0}^{SHAP} = \mathbb{E}\left(\hat{f}(\mathbf{x}_i) \right). \tag{A17}$$

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,* $G\left(y_i; \mathbf{x}_i; \delta_0\right) = \hat{f}(\mathbf{x}_i)$.

See Appendix K.9 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 with 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:

$$\phi_{i,1} = 2\hat{\varepsilon}_i \phi_{i,1}^{SHAP} + \left(\phi_{i,1}^{SHAP}\right)^2 + \mathbb{V}\left(\hat{f}(x_{i,1})\right), \tag{A18}$$

where $\hat{\varepsilon}_i = y_i - \hat{f}(x_{i,1})$ is the prediction error of the model for individual i. See Appendix K.10 for the proof. As we can see, a positive XPER value can be associated with 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, since the individual performance and the model predictions do not share the same domain, except in the case where $G\left(y_i; \mathbf{x}_i; \delta_0\right) = \hat{f}(\mathbf{x}_i)$, the magnitude of SHAP and XPER values can differ significantly.

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:

$$\phi_j^{SHAP} = \mathbb{E}\left(\left|\phi_{i,j}^{SHAP}\right|\right), \tag{A19}$$

where ϕ_j^{SHAP} refers to the Mean Absolute SHAP values for 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 the 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 Equations (A16) and (A17), it would not be appropriate to take the mean of the SHAP values without taking the absolute value, as we

would end up decomposing a zero:

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

$$\Leftrightarrow \sum_{j=1}^{q} \mathbb{E}\left(\phi_{i,j}^{SHAP}\right) = \mathbb{E}\left(\hat{f}(\mathbf{x}_i)\right) - \mathbb{E}\left(\hat{f}(\mathbf{x}_i)\right) = 0. \tag{A20}$$

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 study in Section 7, we confirm that the differences between XPER and SHAP can be substantial in practice.

K Proofs

K.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:

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

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)):

$$
\begin{aligned}
G(y;\mathbf{x}) &= 1 - \frac{(y - \mathbf{x}\hat{\beta})^2}{\sigma_y^2} \\
&= 1 - \sigma_y^{-2}\left[y^2 + \sum_{k=1}^{q} x_k^2\hat{\beta}_k^2 + 2\sum_{1\leq k<l\leq q} \hat{\beta}_k\hat{\beta}_l x_k x_l - 2\sum_{k=1}^{q} y x_k\hat{\beta}_k \right]. \qquad \text{(A21)}
\end{aligned}
$$

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 (A21) as:

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

□

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

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

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

Proof. Let consider a quantity A defined as follows:

$$A = 2\sum_{S\subseteq\mathcal{P}(\{\mathbf{x}\}\setminus\{x_j\})}\omega_S\left(\sum_{\substack{k=1\\k\in S}}^{q}h(x_k) + \sum_{\substack{k=1\\k\in\overline{S}}}^{q}g(x_k)\right), \tag{A22}$$

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 (A22):

$$A = 2\sum_{S\subseteq\bigcup_{l=0}^{q-1}\mathcal{P}_l(\{\mathbf{x}\}\setminus\{x_j\})}\omega_S\left(\sum_{\substack{k=1\\k\in S}}^{q}h(x_k) + \sum_{\substack{k=1\\k\in\overline{S}}}^{q}g(x_k)\right). \tag{A23}$$

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

$$\bigcup_{l=0}^{q-1}\mathcal{P}_l(\{\mathbf{x}\}\setminus\{x_j\}) = \bigcup_{l=0}^{(q-2)/2}\mathcal{P}_{q-1-l}(\{\mathbf{x}\}\setminus\{x_j\})\cup\mathcal{P}_l(\{\mathbf{x}\}\setminus\{x_j\}), \tag{A24}$$

whereas for an odd number of features, we end up with:

$$\bigcup_{l=0}^{q-1} \mathcal{P}_l(\{\mathbf{x}\} \setminus \{x_j\}) = \bigcup_{l=0}^{(q-1)/2} \mathcal{P}_{q-1-l}(\{\mathbf{x}\} \setminus \{x_j\}) \cup \mathcal{P}_l(\{\mathbf{x}\} \setminus \{x_j\}). \tag{A25}$$

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

When q is even, substituting the expression of $\bigcup_{l=0}^{q-1} \mathcal{P}_l(\{\mathbf{x}\}\setminus\{x_j\})$ from Equation (A24) into Equation (A23) renders Equation (A23) equivalent to:

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

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 (A26) as follows:

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

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 (A27) simplifies to:

$$A = 2 \sum_{l=0}^{(q-2)/2} \omega_l \sum_{S \subseteq \mathcal{P}_{q-1-l}(\{\mathbf{x}\}\setminus\{x_j\}) \cup \mathcal{P}_l(\{\mathbf{x}\}\setminus\{x_j\})} \left(\sum_{\substack{k=1 \\ k \in S}}^{q} h(x_k) + \sum_{\substack{k=1 \\ k \in \overline{S}}}^{q} g(x_k) \right). \tag{A28}$$

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:

$$\sum_{S \subseteq \mathcal{P}_{q-1-l}(\{\mathbf{x}\}\setminus\{x_j\}) \cup \mathcal{P}_l(\{\mathbf{x}\}\setminus\{x_j\})} \sum_{\substack{k=1 \\ k \in S}}^{q} h(x_k) = \sum_{\substack{k=1 \\ k \neq j}}^{q} C_{q-1}^{l} h(x_k). \tag{A29}$$

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:

$$\sum_{S \subseteq \mathcal{P}_{q-1-l}(\{\mathbf{x}\} \setminus \{x_j\}) \cup \mathcal{P}_l(\{\mathbf{x}\} \setminus \{x_j\})} \sum_{\substack{k=1 \\ k \in \overline{S}}}^{q} g(x_k) = \sum_{\substack{k=1 \\ k \neq j}}^{q} C_{q-1}^l g(x_k). \tag{A30}$$

As a consequence, Equation (A27) simplifies to:

$$A = 2 \sum_{l=0}^{(q-2)/2} \omega_l C_{q-1}^l \left(\sum_{\substack{k=1 \\ k \neq j}} (h(x_k) + g(x_k)) \right). \tag{A31}$$

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

$$A = 2 \sum_{l=0}^{(q-2)/2} \frac{1}{q} \left(\sum_{\substack{k=1 \\ k \neq j}} (h(x_k) + g(x_k)) \right). \tag{A32}$$

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

$$2 \sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\} \setminus \{x_j\})} \omega_S \left(\sum_{\substack{k=1 \\ k \in S}}^{q} h(x_k) + \sum_{\substack{k=1 \\ k \in \overline{S}}}^{q} g(x_k) \right) = \sum_{\substack{k=1 \\ k \neq j}} (h(x_k) + g(x_k)). \tag{A33}$$

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

$$A = 2 \sum_{l=0}^{(q-1)/2} \omega_l \sum_{S \subseteq \mathcal{P}_{q-1-l}(\{\mathbf{x}\} \setminus \{x_j\}) \cup \mathcal{P}_l(\{\mathbf{x}\} \setminus \{x_j\})} \left(\sum_{\substack{k=1 \\ k \in S}}^{q} h(x_k) + \sum_{\substack{k=1 \\ k \in \overline{S}}}^{q} g(x_k) \right). \tag{A34}$$

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 (A34) as follows:

$$A = 2 \sum_{l=0}^{(q-3)/2} \omega_l \sum_{S \subseteq \mathcal{P}_{q-1-l}(\{\mathbf{x}\}\setminus\{x_j\}) \cup \mathcal{P}_l(\{\mathbf{x}\}\setminus\{x_j\})} \left(\sum_{\substack{k=1 \\ k\in S}}^{q} h(x_k) + \sum_{\substack{k=1 \\ k\in \overline{S}}}^{q} g(x_k) \right) \tag{A35}$$

$$- 2\omega_{(q-1)/2} \sum_{S \subseteq \mathcal{P}_{(q-1)/2}(\{\mathbf{x}\}\setminus\{x_j\})} \left(\sum_{\substack{k=1 \\ k\in S}}^{q} h(x_k) + \sum_{\substack{k=1 \\ k\in \overline{S}}}^{q} g(x_k) \right), \tag{A36}$$

which is equal to:

$$A = 2 \sum_{l=0}^{(q-3)/2} \omega_l C_{q-1}^{l} \left(\sum_{\substack{k=1 \\ k\neq j}} \left(h(x_k) + g(x_k) \right) \right) - 2w_{|(q-1)/2|} \frac{C_{q-1}^{(q-1)/2}}{2} \left(\sum_{\substack{k=1 \\ k\neq j}} \left(h(x_k) + g(x_k) \right) \right). \tag{A37}$$

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

$$A = 2 \sum_{l=0}^{(q-3)/2} \frac{1}{q} \left(\sum_{\substack{k=1 \\ k\neq j}} \left(h(x_k) + g(x_k) \right) \right) - \left(\sum_{\substack{k=1 \\ k\neq j}} \left(h(x_k) + g(x_k) \right) \right). \tag{A38}$$

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

$$A = 2 \sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\}\setminus\{x_j\})} \omega_S \left(\sum_{\substack{k=1 \\ k\in S}}^{q} h(x_k) + \sum_{\substack{k=1 \\ k\in \overline{S}}}^{q} g(x_k) \right) = \sum_{\substack{k=1 \\ k\neq j}} \left(h(x_k) + g(x_k) \right). \tag{A39}$$

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

$$2 \sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\}\setminus\{x_j\})} \omega_S \left(\sum_{\substack{k=1 \\ k\in S}}^{q} h(x_k) + \sum_{\substack{k=1 \\ k\in \overline{S}}}^{q} g(x_k) \right) = \sum_{\substack{k=1 \\ k\neq j}} \left(h(x_k) + g(x_k) \right), \tag{A40}$$

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:

$$\phi_j = \frac{2\hat{\beta}_j \sigma_{y,x_j}}{\sigma_y^2}, \qquad \forall j = 1, ..., q, \tag{A41}$$

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:

$$\begin{aligned}\mathbb{E}_{y,\mathbf{x}^S,x_j}\mathbb{E}_{\mathbf{x}^{\overline{S}}}\left(G(y;\mathbf{x};\delta_0)\right) - \mathbb{E}_{y,\mathbf{x}^S}\mathbb{E}_{\mathbf{x}^{\overline{S}},x_j}\left(G(y;\mathbf{x};\delta_0)\right) &= \sigma_y^{-2}\left[-2\sum_{\substack{k=1\\k\in S}}^{q}\hat{\beta}_k\hat{\beta}_j\sigma_{x_k,x_j} + 2\sum_{\substack{k=1\\k\in\overline{S}}}^{q}\hat{\beta}_k\hat{\beta}_j\sigma_{x_k,x_j}\right] \\ &\quad + \sigma_y^{-2}\left[2\hat{\beta}_j\sigma_{y,x_j}\right],\end{aligned} \tag{A42}$$

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 (A42), the XPER contribution ϕ_j to the R^2 is equal to:

$$\begin{aligned}\phi_j &= \sum_{S\subseteq\mathcal{P}(\{\mathbf{x}\}\setminus\{x_j\})} \omega_S \left(\mathbb{E}_{y,\mathbf{x}^S,x_j}\mathbb{E}_{\mathbf{x}^{\overline{S}}}\left(G(y;\mathbf{x};\delta_0)\right) - \mathbb{E}_{y,\mathbf{x}^S}\mathbb{E}_{\mathbf{x}^{\overline{S}},x_j}\left(G(y;\mathbf{x};\delta_0)\right)\right) \\ &= \sum_{S\subseteq\mathcal{P}(\{\mathbf{x}\}\setminus\{x_j\})} \omega_S \left(\sigma_y^{-2}\left[-2\sum_{\substack{k=1\\k\in S}}^{q}\hat{\beta}_k\hat{\beta}_j\sigma_{x_k,x_j} + 2\sum_{\substack{k=1\\k\in\overline{S}}}^{q}\hat{\beta}_k\hat{\beta}_j\sigma_{x_k,x_j}\right]\right) \\ &\quad + \sum_{S\subseteq\mathcal{P}(\{\mathbf{x}\}\setminus\{x_j\})} \omega_S\sigma_y^{-2}\left[2\hat{\beta}_j\sigma_{y,x_j}\right].\end{aligned} \tag{A43}$$

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

$$\begin{aligned}\phi_j &= \sum_{S\subseteq\mathcal{P}(\{\mathbf{x}\}\setminus\{x_j\})} \omega_S \left(\sigma_y^{-2}\left[-2\sum_{\substack{k=1\\k\in S}}^{q}\hat{\beta}_k\hat{\beta}_j\sigma_{x_k,x_j} + 2\sum_{\substack{k=1\\k\in\overline{S}}}^{q}\hat{\beta}_k\hat{\beta}_j\sigma_{x_k,x_j}\right]\right) \\ &\quad + \sigma_y^{-2}\left[2\hat{\beta}_j\sigma_{y,x_j}\right].\end{aligned} \tag{A44}$$

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:

$$2\sum_{S\subseteq\mathcal{P}(\{\mathbf{x}\}\setminus\{x_j\})}\omega_S\left(\sum_{\substack{k=1\\k\in S}}^{q}-\hat{\beta}_k\hat{\beta}_j\sigma_{x_k,x_j}+\sum_{\substack{k=1\\k\in\overline{S}}}^{q}\hat{\beta}_k\hat{\beta}_j\sigma_{x_k,x_j}\right)=\sum_{\substack{k=1\\k\neq j}}\left(-\hat{\beta}_k\hat{\beta}_j\sigma_{x_k,x_j}+\hat{\beta}_k\hat{\beta}_j\sigma_{x_k,x_j}\right)=0 \tag{A45}$$

Therefore, from Equations (A44) and (A45), we deduce that the XPER contribution ϕ_j of feature x_j to the R^2 is:

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

□

K.2 Proof of Equation (12)

Lemma 4. *The weight associated with 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:

$$\begin{aligned}\mathbf{E}_{\mathbf{x}^{\overline{S}}}\left(G(y_i;\mathbf{x}_i;\delta_0)\right) - \mathbf{E}_{\mathbf{x}^{\overline{S},x_j}}\left(G(y_i;\mathbf{x}_i;\delta_0)\right) &= \sigma_y^{-2}\left[2y_i\hat{\beta}_j(x_{i,j} - \mathbb{E}(x_j)) - \hat{\beta}_j^2(x_{i,j}^2 - \mathbb{E}(x_j^2))\right] \\ &\quad + \sigma_y^{-2}\left[2\sum_{\substack{k=1 \\ k\in\overline{S}}} \hat{\beta}_k\hat{\beta}_j\sigma_{x_k,x_j}\right] \\ &\quad - \sigma_y^{-2}\left[2\hat{\beta}_j(x_{i,j} - \mathbb{E}(x_j))\left(\sum_{\substack{k=1 \\ k\in S}} \hat{\beta}_k x_{i,k} + \sum_{\substack{k=1 \\ k\in\overline{S}}} \hat{\beta}_k\mathbb{E}(x_k)\right)\right],\end{aligned} \tag{A46}$$

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 (A46), the XPER contribution $\phi_{i,j}$ to the R^2 is equal to:

$$\begin{aligned}\phi_{i,j} &= \sum_{S\subseteq\mathcal{P}(\{\mathbf{x}\}\setminus\{x_j\})} \omega_S\left(\mathbf{E}_{\mathbf{x}^{\overline{S}}}\left(G(y_i;\mathbf{x}_i;\delta_0)\right) - \mathbf{E}_{\mathbf{x}^{\overline{S},x_j}}\left(G(y_i;\mathbf{x}_i;\delta_0)\right)\right) \\ &= \sum_{S\subseteq\mathcal{P}(\{\mathbf{x}\}\setminus\{x_j\})} \omega_S\sigma_y^{-2}\left[2y_i\hat{\beta}_j(x_{i,j} - \mathbb{E}(x_j)) - \hat{\beta}_j^2(x_{i,j}^2 - \mathbb{E}(x_j^2)) + 2\sum_{\substack{k=1 \\ k\in\overline{S}}} \hat{\beta}_k\hat{\beta}_j\sigma_{x_k,x_j}\right] \\ &\quad - \sum_{S\subseteq\mathcal{P}(\{\mathbf{x}\}\setminus\{x_j\})} \omega_S\sigma_y^{-2}\left[2\hat{\beta}_j(x_{i,j} - \mathbb{E}(x_j))\left(\sum_{\substack{k=1 \\ k\in S}} \hat{\beta}_k x_{i,k} + \sum_{\substack{k=1 \\ k\in\overline{S}}} \hat{\beta}_k\mathbb{E}(x_k)\right)\right].\end{aligned} \tag{A47}$$

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

$$\phi_{i,j} = \sigma_y^{-2}\left[2y_i\hat{\beta}_j(x_{i,j}-\mathbb{E}(x_j)) - \hat{\beta}_j^2(x_{i,j}^2-\mathbb{E}(x_j^2)) + 2\sum_{S\subseteq\mathcal{P}(\{\mathbf{x}\}\setminus\{x_j\})}\omega_S\sum_{\substack{k=1\\k\in S}}\hat{\beta}_k\hat{\beta}_j\sigma_{x_k,x_j}\right]$$
$$-\sigma_y^{-2}\left[\hat{\beta}_j(x_{i,j}-\mathbb{E}(x_j))\times 2\sum_{S\subseteq\mathcal{P}(\{\mathbf{x}\}\setminus\{x_j\})}\omega_S\left(\sum_{\substack{k=1\\k\in S}}\hat{\beta}_k x_{i,k} + \sum_{\substack{k=1\\k\in \overline{S}}}\hat{\beta}_k\mathbb{E}(x_k)\right)\right]. \tag{A48}$$

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:

$$2\sum_{S\subseteq\mathcal{P}(\{\mathbf{x}\}\setminus\{x_j\})}\omega_S\sum_{\substack{k=1\\k\in S}}\hat{\beta}_k\hat{\beta}_j\sigma_{x_k,x_j} = \sum_{\substack{k=1\\k\neq j}}^{q}\hat{\beta}_k\hat{\beta}_j\sigma_{x_k,x_j} \tag{A49}$$

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

$$2\sum_{S\subseteq\mathcal{P}(\{\mathbf{x}\}\setminus\{x_j\})}\omega_S\left(\sum_{\substack{k=1\\k\in S}}\hat{\beta}_k x_{i,k} + \sum_{\substack{k=1\\k\in \overline{S}}}\hat{\beta}_k\mathbb{E}(x_{i,k})\right) = \sum_{\substack{k=1\\k\neq j}}^{q}\hat{\beta}_k(x_{i,k}+\mathbb{E}(x_k)) \tag{A50}$$

From Equations (A48), (A49), and (A50), we obtain:

$$\phi_{i,j} = \sigma_y^{-2}\left[2y_i\hat{\beta}_j(x_{i,j}-\mathbb{E}(x_j)) - \hat{\beta}_j^2(x_{i,j}^2-\mathbb{E}(x_j^2)) + \sum_{\substack{k=1\\k\neq j}}^{q}\hat{\beta}_k\hat{\beta}_j\sigma_{x_k,x_j}\right]$$
$$-\sigma_y^{-2}\left[\hat{\beta}_j(x_{i,j}-\mathbb{E}(x_j))\sum_{\substack{k=1\\k\neq j}}^{q}\hat{\beta}_k(x_{i,k}+\mathbb{E}(x_k))\right]. \tag{A51}$$

Finally, after 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 $A = \left(2y_i - \sum_{\substack{k=1\\k\neq j}}^{q}\hat{\beta}_k(x_{i,k}+\mathbb{E}(x_k))\right)$, 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. □

K.3 Proof of the MSE example in Table A4

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* $\mathbb{V}(\mathbf{x}) = diag(\sigma^2_{x_j}), \forall j = 1, ..., q$*, 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:*

$$\underbrace{2\sum_{j=1}^{q} \hat{\beta}_j \sigma_{y,x_j} - \sum_{j=1}^{q} \hat{\beta}_j^2 \sigma^2_{x_j} - \sigma_y^2}_{\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0))} = \underbrace{-\sum_{j=1}^{q} \hat{\beta}_j^2 \sigma^2_{x_j} - \sigma_y^2}_{\phi_0} + \sum_{j=1}^{q} \underbrace{2\hat{\beta}_j \sigma_{y,x_j}}_{\phi_j}. \tag{A52}$$

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 $\mathbb{V}(\mathbf{x}) = diag(\sigma^2_{x_j}), \forall j = 1, ..., q$. As shown in Table A2, 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:

$$G(y; \mathbf{x}; \delta_0)) = -\left[y^2 + \sum_{j=1}^{q} x_j^2 \hat{\beta}_j^2 + 2 \sum_{1 \leq j < l \leq q} \hat{\beta}_j \hat{\beta}_l x_j x_l - 2 \sum_{j=1}^{q} y x_j \hat{\beta}_j \right]. \tag{A53}$$

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

$$\mathbb{E}_{y,\mathbf{x}}(G(y; \mathbf{x}; \delta_0)) = 2\sum_{j=1}^{q} \hat{\beta}_j \sigma_{y,x_j} - \sum_{j=1}^{q} \hat{\beta}_j^2 \sigma^2_{x_j} - \sigma_y^2. \tag{A54}$$

Indeed, by taking the expected value of Equation (A53) with respect to the joint distribution of the target variable and the features, we obtain:

$$\mathbb{E}_{y,\mathbf{x}}(G(y; \mathbf{x}; \delta_0)) = -\mathbb{E}(y^2) - \sum_{j=1}^{q} \hat{\beta}_j^2 \mathbb{E}(x_j^2) - 2 \sum_{1 \leq j < l \leq q} \hat{\beta}_j \hat{\beta}_l \mathbb{E}(x_j x_l) + 2 \sum_{j=1}^{q} \hat{\beta}_j \mathbb{E}(y x_j). \tag{A55}$$

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

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

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

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

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

$$\phi_0 = \mathbb{E}_y \mathbb{E}_{\mathbf{x}}(G(y;\mathbf{x};\delta_0)) = -\sigma_y^2 - \sum_{j=1}^{q} \hat{\beta}_j^2 \sigma_{x_j}^2. \tag{A57}$$

From Equation (A53), 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:

$$\phi_0 = \mathbb{E}_y \mathbb{E}_{\mathbf{x}}(G(y;\mathbf{x};\delta_0)) = -\mathbb{E}(y^2) - \sum_{j=1}^{q} \hat{\beta}_j^2 \mathbb{E}(x_j^2) - 2 \sum_{1 \leq j < l \leq q} \hat{\beta}_j \hat{\beta}_l \mathbb{E}(x_j x_l) + 2 \sum_{j=1}^{q} \hat{\beta}_k \mathbb{E}(y) \mathbb{E}(x_j). \tag{A58}$$

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

$$\phi_0 = \mathbb{E}_y \mathbb{E}_{\mathbf{x}}(G(y;\mathbf{x};\delta_0)) = -\sigma_y^2 - \sum_{j=1}^{q} \hat{\beta}_j^2 \sigma_{x_j}^2. \tag{A59}$$

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

$$\phi_j = 2\hat{\beta}_j \sigma_{y,x_j}. \tag{A60}$$

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 with the feature x_j for the R^2:

$$\begin{aligned}
\phi_j^{R^2} &= \sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\} \setminus \{x_j\})} \omega_S \left[\mathbb{E}_{y,x_j,\mathbf{x}^S} \mathbb{E}_{\mathbf{x}^{\overline{S}}} \left(1 + \sigma_y^{-2} G(y;\mathbf{x};\delta_0)\right) - \mathbb{E}_{y,\mathbf{x}^S} \mathbb{E}_{\mathbf{x}^{x_j,\overline{S}}} \left(1 + \sigma_y^{-2} G(y;\mathbf{x};\delta_0)\right) \right] \\
\phi_j^{R^2} &= \sigma_y^{-2} \left[\sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\} \setminus \{x_j\})} \omega_S \left[\mathbb{E}_{y,x_j,\mathbf{x}^S} \mathbb{E}_{\mathbf{x}^{\overline{S}}} \left(G(y;\mathbf{x};\delta_0)\right) - \mathbb{E}_{y,\mathbf{x}^S} \mathbb{E}_{\mathbf{x}^{x_j,\overline{S}}} \left(G(y;\mathbf{x};\delta_0)\right) \right] \right] \\
\phi_j^{R^2} &= \frac{\phi_j}{\sigma_y^2}.
\end{aligned} \tag{A61}$$

Therefore, from Equations (6) and (A61), we obtain:

$$\phi_j = 2\hat{\beta}_j \sigma_{y,x_j}. \tag{A62}$$

Finally, from Equations (A56), (A59), and (A62), we conclude that:

$$\underbrace{2\sum_{j=1}^{q}\hat{\beta}_j\sigma_{y,x_j} - \sum_{j=1}^{q}\hat{\beta}_j^2\sigma_{x_j}^2 - \sigma_y^2}_{\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0))} = \underbrace{-\sum_{j=1}^{q}\hat{\beta}_j^2\sigma_{x_j}^2 - \sigma_y^2}_{\phi_0} + \sum_{j=1}^{q}\underbrace{2\hat{\beta}_j\sigma_{y,x_j}}_{\phi_j}. \tag{A63}$$

□

K.4 Proof of the R^2 example in Table A4

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:*

$$\underbrace{\frac{\sigma_{y,\hat{y}}}{\sigma_y^2}}_{\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0))} = \underbrace{-\frac{\sigma_{y,\hat{y}}}{\sigma_y^2}}_{\phi_0} + \sum_{j=1}^{q}\underbrace{\frac{2\hat{\beta}_j\sigma_{y,x_j}}{\sigma_y^2}}_{\phi_j}. \tag{A64}$$

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:

$$G(y;\mathbf{x};\delta_0)) = 1 - \sigma_y^{-2}\left[y^2 + \sum_{j=1}^{q}x_j^2\hat{\beta}_j^2 + 2\sum_{1\leq j<l\leq q}\hat{\beta}_j\hat{\beta}_l x_j x_l - 2\sum_{j=1}^{q}yx_j\hat{\beta}_j\right]. \tag{A65}$$

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

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

Indeed, by taking the expected value of Equation (A65) with respect to the joint distribution of the target variable and the features, we obtain:

$$\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0)) = 1 - \sigma_y^{-2}\left[\mathbb{E}(y^2) + \sum_{j=1}^{q}\hat{\beta}_j^2\mathbb{E}(x_j^2) + 2\sum_{1\leq j<l\leq q}\hat{\beta}_j\hat{\beta}_l\mathbb{E}(x_j x_l) - 2\sum_{j=1}^{q}\hat{\beta}_j\mathbb{E}(yx_j)\right]. \tag{A67}$$

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

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

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

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

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:

$$\sigma_{y,\hat{y}} = \sum_{j=1}^{q}\hat{\beta}_j^2\sigma^2_{x_j} = \sum_{j=1}^{q}\hat{\beta}_j\sigma_{y,x_j}, \tag{A69}$$

with $\sigma_{y,\hat{y}}$ the covariance between the target variable and its prediction. Therefore, the R^2 as expressed in Equation (A68) is also written as:

$$\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0)) = \frac{\sigma_{y,\hat{y}}}{\sigma_y^2}. \tag{A70}$$

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

$$\phi_0 = \mathbb{E}_y\mathbb{E}_{\mathbf{x}}(G(y;\mathbf{x};\delta_0)) = \frac{\sigma_{y,\hat{y}}}{\sigma_y^2}. \tag{A71}$$

From Equation (A53), 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:

$$\begin{aligned}\phi_0 = \mathbb{E}_y\mathbb{E}_{\mathbf{x}}(G(y;\mathbf{x};\delta_0)) = 1 - \sigma_y^{-2}&\left[\mathbb{E}(y^2) + \sum_{j=1}^{q}\hat{\beta}_j^2\mathbb{E}(x_j^2) + 2\sum_{1\leq j<l\leq q}\hat{\beta}_j\hat{\beta}_l\mathbb{E}(x_jx_l)\right] \\ &- \sigma_y^{-2}\left[-2\sum_{j=1}^{q}\hat{\beta}_j\mathbb{E}(y)\mathbb{E}(x_j)\right].\end{aligned} \tag{A72}$$

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

$$\phi_0 = \mathbb{E}_y \mathbb{E}_{\mathbf{x}}(G(y; \mathbf{x}; \delta_0)) = -\frac{\sum_{j=1}^{q} \hat{\beta}_j^2 \sigma^2_{x_j}}{\sigma^2_y}. \tag{A73}$$

From Equation (A69), we can rewrite Equation (A73) as:

$$\phi_0 = \mathbb{E}_y \mathbb{E}_{\mathbf{x}}(G(y; \mathbf{x}; \delta_0)) = -\frac{\sigma_{y,\hat{y}}}{\sigma^2_y}. \tag{A74}$$

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

$$\phi_j = \frac{2\hat{\beta}_j \sigma_{y,x_j}}{\sigma^2_y}. \tag{A75}$$

Finally, from Equations (A70), (A74), and (A75), we conclude that:

$$\underbrace{\frac{\sigma_{y,\hat{y}}}{\sigma^2_y}}_{\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0))} = \underbrace{-\frac{\sigma_{y,\hat{y}}}{\sigma^2_y}}_{\phi_0} + \sum_{j=1}^{q} \underbrace{\frac{2\hat{\beta}_j \sigma_{y,x_j}}{\sigma^2_y}}_{\phi_j}. \tag{A76}$$

□

K.5 Proof of the accuracy example in Table A4

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:*

$$\underbrace{2\sigma_{y,\hat{f}(\mathbf{x})} + 2\mathbb{P}(y=1)\hat{P}(\mathbf{x}) + 1 - \mathbb{P}(y=1) - \hat{P}(\mathbf{x})}_{\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0))} = \underbrace{2\mathbb{P}(y=1)\hat{P}(\mathbf{x}) + 1 - \mathbb{P}(y=1) - \hat{P}(\mathbf{x})}_{\phi_0} + \underbrace{2\sigma_{y,\hat{f}(\mathbf{x})}}_{\sum_{j=1}^{q}\phi_j}, \tag{A77}$$

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 prob-

ability of belonging to class 1 (y=1). As shown in Table A2, the individual contribution to the accuracy is defined as:

$$G(y;\mathbf{x};\delta_0)) = y\hat{f}(\mathbf{x}) + (1-y)(1-\hat{f}(\mathbf{x})). \tag{A78}$$

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

$$\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0)) = 2\sigma_{y,\hat{f}(\mathbf{x})} + 2\mathbb{P}(y=1)\hat{P}(\mathbf{x}) + 1 - \mathbb{P}(y=1) - \hat{P}(\mathbf{x}). \tag{A79}$$

Indeed, by taking the expected value of Equation (A78) with respect to the joint distribution of the target variable and the features, we obtain:

$$\begin{aligned}\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0)) &= 2\mathbb{E}\left(y\hat{f}(\mathbf{x})\right) + 1 - \mathbb{E}(y) - \mathbb{E}\left(\hat{f}(\mathbf{x})\right) \\ &= 2\sigma_{y,\hat{f}(\mathbf{x})} + 2\mathbb{E}(y)\mathbb{E}\left(\hat{f}(\mathbf{x})\right) + 1 - \mathbb{E}(y) - \mathbb{E}\left(\hat{f}(\mathbf{x})\right),\end{aligned} \tag{A80}$$

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:

$$\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0)) = 2\sigma_{y,\hat{f}(\mathbf{x})} + 2\mathbb{P}(y=1)\hat{P}(\mathbf{x}) + 1 - \mathbb{P}(y=1) - \hat{P}(\mathbf{x}). \tag{A81}$$

Second, from Equation (A78), we can see that by taking the expected value with respect to the target variable and the expected value with respect to the joint distribution of the features, the benchmark value of accuracy can be written as:

$$\begin{aligned}\phi_0 = \mathbb{E}_y\mathbb{E}_{\mathbf{x}}(G(y;\mathbf{x};\delta_0)) &= 2\mathbb{E}\left(y\right)\mathbb{E}\left(\hat{f}(\mathbf{x})\right) + 1 - \mathbb{E}(y) - \mathbb{E}\left(\hat{f}(\mathbf{x})\right) \\ &= 2\mathbb{P}(y=1)\hat{P}(\mathbf{x}) + 1 - \mathbb{P}(y=1) - \hat{P}(\mathbf{x}).\end{aligned} \tag{A82}$$

Third, according to the axiom 1, we deduce that the sum of the XPER values associated with the

features x_j for the accuracy is equal to:

$$\sum_{j=1}^{q} \phi_j = 2\sigma_{y,\hat{f}(\mathbf{x})}. \tag{A83}$$

Finally, from Equations (A81), (A82), and (A83), we conclude that:

$$\underbrace{2\sigma_{y,\hat{f}(\mathbf{x})} + 2\mathbb{P}(y=1)\hat{P}(\mathbf{x}) + 1 - \mathbb{P}(y=1) - \hat{P}(\mathbf{x})}_{\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0))} = \underbrace{2\mathbb{P}(y=1)\hat{P}(\mathbf{x}) + 1 - \mathbb{P}(y=1) - \hat{P}(\mathbf{x})}_{\phi_0} + \underbrace{2\sigma_{y,\hat{f}(\mathbf{x})}}_{\sum_{j=1}^{q}\phi_j}. \tag{A84}$$

□

K.6 Proof of identical normalized XPER values for a specific DGP

Proposition 6. *Consider a DGP given by a regression model with quadratic effects, such that:*

$$y = \beta_0 + \sum_{j=1}^{q} x_j\beta_j + \sum_{j=1}^{q} x_j^2\omega_j + \varepsilon,$$

where ε is an i.i.d. error term, with $\varepsilon \sim \mathcal{N}(0,1)$, β_0 the intercept, and (β_j, ω_j) the parameters associated to the primary variable x_j. Assuming that each pair of parameters (β_j, ω_j) is identical across all primary variables, setting $\beta_j = a\theta$ and $\omega_j = \theta$, for $j = 1, \ldots, q$, where a is a constant, then we have:

$$y = \beta_0 + \sum_{j=1}^{q}(ax_j + x_j^2)\theta + \varepsilon = \sum_{j=1}^{q} z_j\theta + \varepsilon,$$

with $z_j = ax_j + x_j^2$. We can show that the normalized XPER value associated to the primary variable x_j is equal to:

$$\tilde{\phi}_j = \frac{1}{q}.$$

Proof. Consider a DGP given by a regression model with quadratic effects, such that:

$$y = \beta_0 + \sum_{j=1}^{q} x_j\beta_j + \sum_{j=1}^{q} x_j^2\omega_j + \varepsilon,$$

where ε is an i.i.d. error term, with $\varepsilon \sim \mathcal{N}(0,1)$, β_0 the intercept, and (β_j, ω_j) the parameters

associated to the primary variable x_j. Assuming that each pair of parameters (β_j, ω_j) is identical across all primary variables, setting $\beta_j = a\theta$ and $\omega_j = \theta$, for $j = 1, \ldots, q$, where a is a constant, then we have:

$$y = \beta_0 + \sum_{j=1}^{q} (ax_j + x_j^2)\theta + \varepsilon = \beta_0 + \sum_{j=1}^{q} z_j\theta + \varepsilon,$$

with $z_j = ax_j + x_j^2$. Regardless of the primary variables distribution, the XPER value associated to the primary variable z_j, when considering the model parameters θ as given, is defined as:

$$\phi_j = \frac{2\theta\sigma_{y,z_j}}{\sigma_y^2}.$$

When R^2 is used as the performance metric, the normalized XPER value is defined as:

$$\tilde{\phi}_j = \frac{\phi_j}{R^2 - \phi_0} = \frac{\phi_j}{\sum_{k=1}^{q} \phi_k},$$

where ϕ_0 represents the benchmark value (see Equation (10) in the paper). Therefore, under this framework, the normalized XPER value is equal to:

$$\tilde{\phi}_j = \frac{2\theta\sigma_{y,z_j}}{\sum_{k=1}^{q} 2\theta\sigma_{y,z_k}} = \frac{\sigma_{y,z_j}}{\sum_{k=1}^{q} \sigma_{y,z_k}}. \tag{A85}$$

The covariance between the target variable y and each feature z_j is equal to:

$$\sigma_{y,z_j} = \theta \sum_{k=1}^{q} \left(a^2\sigma_{x_k,x_j} + a\sigma_{x_k^2,x_j} + a\sigma_{x_k,x_j^2} + \sigma_{x_k^2,x_j^2} \right),$$

assuming that each feature z_j is independent from the error term ε. Further assuming that all primary variables are identically distributed and with the same dependence across them, then:

$$\sigma_{y,z_j} = \theta \sum_{k=1}^{q} \tau = \theta \times q \times \tau, \qquad \forall j = 1, \ldots, q, \tag{A86}$$

with

$$\tau = \left(a^2\sigma_{x_k,x_j} + a\sigma_{x_k^2,x_j} + a\sigma_{x_k,x_j^2} + \sigma_{x_k^2,x_j^2} \right), \quad \forall k, j = 1, \ldots, q.$$

Finally, by replacing Equation (A86) in Equation (A85), we find that the normalized XPER value

associated to the primary variable x_j is equal to:

$$\tilde{\phi}_j = \frac{1}{q}.$$

□

K.7 Proof of equivalence between sensitivity and specificity decomposition

Proposition 7. *When decomposing sensitivity and specificity with XPER: (1) the nominal XPER values ϕ_j are identical, up to a normalization factor, and (2) the XPER values in percentage, i.e., $\phi_j/(\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0)) - \phi_0)$, are identical.*

Proof. First, by definition, the population sensitivity and specificity are expressed as follows:

$$\begin{cases} \mathbb{E}_{y,\mathbf{x}}\left(G_{sensi}(y_i;\mathbf{x}_i;\delta_0)\right) & = \frac{1}{\mathbb{P}(Y=1)} \times \mathbb{E}_{y,\mathbf{x}}\left(y_i\hat{f}(\mathbf{x}_i)\right), \\ \mathbb{E}_{y,\mathbf{x}}\left(G_{speci}(y_i;\mathbf{x}_i;\delta_0)\right) & = \frac{1}{\mathbb{P}(Y=0)} \times \mathbb{E}_{y,\mathbf{x}}\left(y_i\hat{f}(\mathbf{x}_i) + 1 - y_i - \hat{f}(\mathbf{x}_i)\right). \end{cases}$$

According to the linearity axiom, we know that:

$$\begin{aligned} \phi_j\left(\mathbb{E}_{y,\mathbf{x}}\left(y_i\hat{f}(\mathbf{x}_i) + 1 - y_i - \hat{f}(\mathbf{x}_i)\right)\right) = {} & \phi_j\left(\mathbb{E}_{y,\mathbf{x}}\left(y_i\hat{f}(\mathbf{x}_i)\right)\right) \\ & + \phi_j\left(\mathbb{E}_{y,\mathbf{x}}\left(1 - y_i\right)\right) \\ & - \phi_j\left(\mathbb{E}_{y,\mathbf{x}}\left(\hat{f}(\mathbf{x}_i)\right)\right). \end{aligned}$$

Thus, we can see that if $\phi_j\left(\mathbb{E}_{y,\mathbf{x}}\left(1 - y_i\right)\right) = 0$ and $\phi_j\left(\mathbb{E}_{y,\mathbf{x}}\left(\hat{f}(\mathbf{x}_i)\right)\right) = 0$, then the contribution of the feature x_j to sensitivity will be equal to its contribution to specificity, up to a normalization factor. As there are no features in the performance metric $PM = \mathbb{E}_{y,\mathbf{x}}\left(1 - y_i\right)$, by applying XPER on this performance metric, it is straightforward to see that $\phi_j\left(\mathbb{E}_{y,\mathbf{x}}\left(1 - y_i\right)\right) = 0$. Regarding the second performance measure $PM = \mathbb{E}_{y,\mathbf{x}}\left(\hat{f}(\mathbf{x}_i)\right)$, by applying XPER, we can see that the benchmark equals the latter. Intuitively, since there is no difference between the performance and its benchmark value, this means that these features do not improve performance, so their contribution must be null. In a simple case with three explanatory variables, we can see that this holds true and understand why

it does. Applying XPER, the feature contribution of feature x_1 to this performance is:

$$\begin{aligned}
\phi_1 = &\frac{1}{3}\left(\mathbb{E}_{x_1}\mathbb{E}_{x_2,x_3}\left(\hat{f}(\mathbf{x}_i)\right) - \mathbb{E}_{x_1,x_2,x_3}\left(\hat{f}(\mathbf{x}_i)\right)\right) \\
&+ \frac{1}{6}\left(\mathbb{E}_{x_1,x_2}\mathbb{E}_{x_3}\left(\hat{f}(\mathbf{x}_i)\right) - \mathbb{E}_{x_2}\mathbb{E}_{x_1,x_3}\left(\hat{f}(\mathbf{x}_i)\right)\right) \\
&+ \frac{1}{6}\left(\mathbb{E}_{x_1,x_3}\mathbb{E}_{x_2}\left(\hat{f}(\mathbf{x}_i)\right) - \mathbb{E}_{x_3}\mathbb{E}_{x_1,x_2}\left(\hat{f}(\mathbf{x}_i)\right)\right) \\
&+ \frac{1}{3}\left(\mathbb{E}_{x_1,x_2,x_3}\left(\hat{f}(\mathbf{x}_i)\right) - \mathbb{E}_{x_2,x_3}\mathbb{E}_{x_1}\left(\hat{f}(\mathbf{x}_i)\right)\right), \\
\phi_1 = &\, 0.
\end{aligned}$$

Similarly for x_2 and x_3, we can easily see that $\phi_2 = \phi_3 = 0$. In a way, in this situation, the marginal effects calculated for the different coalitions cancel out. Thus, the contribution of each feature to the performance is null. Therefore, to decompose either the sensitivity or specificity, we only need to apply XPER on $PM = \mathbb{E}_{y,\mathbf{x}}\left(y_i\hat{f}(\mathbf{x}_i)\right)$, and divide the contributions by $\mathbb{P}(Y=1)$ or $\mathbb{P}(Y=0)$. Let denote by $\tilde{\phi}_j$ the XPER value associated with feature x_j, measuring its contribution to $PM = \mathbb{E}_{y,\mathbf{x}}\left(y_i\hat{f}(\mathbf{x}_i)\right)$. Then, the feature contribution x_j to the sensitivity and specificity are:

$$\begin{cases}
\phi_j^{sensi} = \frac{\tilde{\phi}_j}{\mathbb{P}(Y=1)}, \\
\phi_j^{speci} = \frac{\tilde{\phi}_j}{\mathbb{P}(Y=0)}.
\end{cases}$$

Second, to ease the interpretation of the XPER values, we usually divide their contribution by the difference between the performance metric and its benchmark. For the sensitivity and specificity, the difference between the performance metric and its benchmark is:

$$\begin{cases}
\mathbb{E}_{y,\mathbf{x}}\left(G_{sensi}(y_i;\mathbf{x}_i;\delta_0)\right) - \phi_0^{sensi} = \frac{\mathbb{E}_{y,\mathbf{x}}\left(y_i\hat{f}(\mathbf{x}_i)\right) - \mathbb{E}_y\mathbb{E}_{\mathbf{x}}\left(y_i\hat{f}(\mathbf{x}_i)\right)}{\mathbb{P}(Y=1)}, \\
\mathbb{E}_{y,\mathbf{x}}\left(G_{speci}(y_i;\mathbf{x}_i;\delta_0)\right) - \phi_0^{speci} = \frac{\mathbb{E}_{y,\mathbf{x}}\left(y_i\hat{f}(\mathbf{x}_i)\right) - \mathbb{E}_y\mathbb{E}_{\mathbf{x}}\left(y_i\hat{f}(\mathbf{x}_i)\right)}{\mathbb{P}(Y=0)},
\end{cases}$$

with

$$\begin{cases}
\phi_0^{sensi} = \frac{\mathbb{E}_y\mathbb{E}_{\mathbf{x}}\left(y_i\hat{f}(\mathbf{x}_i)\right)}{\mathbb{P}(Y=1)}, \\
\phi_0^{speci} = \frac{\mathbb{E}_y\mathbb{E}_{\mathbf{x}}\left(y_i\hat{f}(\mathbf{x}_i)\right) + 1 - \mathbb{E}(y_i) - \mathbb{E}(\hat{f}(\mathbf{x}_i))}{\mathbb{P}(Y=0)}.
\end{cases}$$

Therefore, the contribution (expressed in percentage) of the features to the sensitivity and specificity

are identical:

$$\frac{\phi_j^{sensi}}{\mathbb{E}_{y,\mathbf{x}}\left(G_{sensi}(y_i;\mathbf{x}_i;\delta_0)\right)-\phi_0^{sensi}} = \frac{\tilde{\phi}_j}{\mathbb{E}_{y,\mathbf{x}}\left(y_i\hat{f}(\mathbf{x}_i)\right)-\mathbb{E}_y\mathbb{E}_{\mathbf{x}}\left(y_i\hat{f}(\mathbf{x}_i)\right)} = \frac{\phi_j^{speci}}{\mathbb{E}_{y,\mathbf{x}}\left(G_{speci}(y_i;\mathbf{x}_i;\delta_0)\right)-\phi_0^{speci}}.$$

□

K.8 Proof of Equation (A14)

Proposition 8. *The XPER value ϕ_j (see Definition 2) can be expressed as:*

$$\phi_j = \sum_{S\subseteq\mathcal{P}(\{\mathbf{x}_{-j}\})\setminus\{\mathbf{x}_{-j}\}} \omega_S\left[\mathbb{E}_{y,x_j,\mathbf{x}^S}\mathbb{E}_{\mathbf{x}^{\overline{S}}}\left(G\left(y;\mathbf{x};\delta_0\right)\right)-\mathbb{E}_{y,\mathbf{x}^S}\mathbb{E}_{x_j,\mathbf{x}^{\overline{S}}}\left(G\left(y;\mathbf{x};\delta_0\right)\right)\right]+\frac{PI_j}{q}, \tag{A87}$$

where PI_j corresponds to the Permutation Importance (PI) value of feature x_j, as defined in Equation (A13).

Proof. To establish the link between PI_j (as defined in Equation (A13)) and ϕ_j (see Definition 2), let us introduce some additional notations. Let denote $\{\mathbf{x}_{-j}\} = \{\mathbf{x}\}\setminus\{x_j\}$ as the set of explanatory variables excluding the feature x_j. The powerset of the set $\{\mathbf{x}_{-j}\}$ is defined as $\mathcal{P}(\{\mathbf{x}_{-j}\})$. According to Equation (4) in the paper, for $S=\{\mathbf{x}_{-j}\}$ then $w_S = 1/q$, where q denotes the number of model features. Note that for $S=\{\mathbf{x}_{-j}\}$ we have $\mathbf{x}^{\overline{S}}=\{\varnothing\}$. Reminds that the XPER value associated with feature x_j is defined as follows:

$$\phi_j = \sum_{S\subseteq\mathcal{P}(\{\mathbf{x}_{-j}\})} \omega_S\left[\mathbb{E}_{y,x_j,\mathbf{x}^S}\mathbb{E}_{\mathbf{x}^{\overline{S}}}\left(G\left(y;\mathbf{x};\delta_0\right)\right)-\mathbb{E}_{y,\mathbf{x}^S}\mathbb{E}_{x_j,\mathbf{x}^{\overline{S}}}\left(G\left(y;\mathbf{x};\delta_0\right)\right)\right]. \tag{A88}$$

Taking into account that $\mathcal{P}(\{\mathbf{x}_{-j}\}) = (\mathcal{P}(\{\mathbf{x}_{-j}\})\setminus\{\mathbf{x}_{-1}\})\cup\{\mathbf{x}_{-1}\}$, we can rearrange the terms to obtain:

$$\begin{aligned}\phi_j = \sum_{S\subseteq\mathcal{P}(\{\mathbf{x}_{-j}\})\setminus\{\mathbf{x}_{-j}\}} &\omega_S\left[\mathbb{E}_{y,x_j,\mathbf{x}^S}\mathbb{E}_{\mathbf{x}^{\overline{S}}}\left(G\left(y;\mathbf{x};\delta_0\right)\right)-\mathbb{E}_{y,\mathbf{x}^S}\mathbb{E}_{x_j,\mathbf{x}^{\overline{S}}}\left(G\left(y;\mathbf{x};\delta_0\right)\right)\right]\\ &+\frac{1}{q}\left[\mathbb{E}_{y,\mathbf{x}}\left(G\left(y;\mathbf{x};\delta_0\right)\right)-\mathbb{E}_{y,\mathbf{x}_{-j}}\mathbb{E}_{x_j}\left(G(y,\mathbf{x};\delta_0)\right)\right].\end{aligned} \tag{A89}$$

Therefore, we can see that the XPER value ϕ_j can be expressed as the sum of two terms, one of

which is the (normalized) PI_j:

$$\phi_j = \sum_{S \subseteq \mathcal{P}(\{\mathbf{x}_{-j}\}) \setminus \{\mathbf{x}_{-j}\}} \omega_S \left[\mathbb{E}_{y,x_j,\mathbf{x}^S} \mathbb{E}_{\mathbf{x}^{\overline{S}}} \left(G\left(y; \mathbf{x}; \delta_0\right)\right) - \mathbb{E}_{y,\mathbf{x}^S} \mathbb{E}_{x_j,\mathbf{x}^{\overline{S}}} \left(G\left(y; \mathbf{x}; \delta_0\right)\right) \right] + \frac{1}{q} \left[PI_j \right]. \quad \text{(A90)}$$

□

Thus, the XPER value ϕ_j can be expressed as the permutation importance (PI) of feature x_j (normalized by the number of model features) plus an additional term. This additional term vanishes, making ϕ_j exactly equal to the PI, *if and only if* the marginal impact of x_j on performance is identical across all coalitions, i.e., $\mathbb{E}_{y,x_j,\mathbf{x}^S}$ must be consistent across all subsets of features. This condition holds for linear regression models. However, for non-linear models, this condition is generally not valid, leading to different PI and XPER values, as illustrated in Figure 3b in the empirical application.

K.9 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 with the performance metric $G\left(y_i; \mathbf{x}_i; \delta_0\right) = \hat{f}(\mathbf{x}_i)$ is equal to:

$$\phi_{i,j} = \sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\} \setminus \{x_j\})} w_S \left[\mathbb{E}_{\mathbf{x}^{\overline{S}}} \left(\hat{f}\left(x_{i,j}, \mathbf{x}_i^S, \mathbf{x}^{\overline{S}} \right) \right) - \mathbb{E}_{x_j,\mathbf{x}^{\overline{S}}} \left(\hat{f}\left(x_j, \mathbf{x}_i^S, \mathbf{x}^{\overline{S}} \right) \right) \right]. \quad \text{(A91)}$$

Moreover, according Equation (A15), the SHAP value ϕ_j^{SHAP} associated with $\hat{f}(\mathbf{x}_i)$ is equal to:

$$\phi_{i,j}^{SHAP} = \sum_{S \subseteq \mathcal{P}(\{\mathbf{x}\} \setminus \{x_j\})} w_S \left[\mathbb{E}_{\mathbf{x}^{\overline{S}}} \left(\hat{f}(x_{i,j}, \mathbf{x}_i^S, \mathbf{x}^{\overline{S}}) \right) - \mathbb{E}_{x_j,\mathbf{x}^{\overline{S}}} \left(\hat{f}(x_j, \mathbf{x}_i^S, \mathbf{x}^{\overline{S}}) \right) \right]. \quad \text{(A92)}$$

Therefore, in the particular case where $G\left(y_i; \mathbf{x}_i; \delta_0\right) = \hat{f}(\mathbf{x}_i)$, according to Equations (A91) and (A92), the individual XPER value ϕ_j is equal to the SHAP value ϕ_j^{SHAP}. □

K.10 Proof of Equation (A18)

Proposition 9. *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:*

$$\phi_{i,1} = 2\hat{\varepsilon}_i \phi_{i,1}^{SHAP} + \left(\phi_{i,1}^{SHAP}\right)^2 + \mathbb{V}\left(\hat{f}(x_{i,1})\right), \tag{A93}$$

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

$$G(y_i; \hat{f}(\mathbf{x}_i); \delta_0) = \phi_{i,0} + \phi_{i,1}, \tag{A94}$$

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:

$$\phi_{i,1} = -y_i^2 - \hat{f}(x_{i,1})^2 + 2y_i\hat{f}(x_{i,1}) - \phi_{i,0}. \tag{A95}$$

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

$$\phi_{i,0} = \mathbb{E}_{x_1}\left(G(y_i; \hat{f}(\mathbf{x}_i); \delta_0)\right) = -y_i^2 - \mathbb{E}\left(\hat{f}(x_{i,1})^2\right) + 2y_i\mathbb{E}\left(\hat{f}(x_{i,1})\right). \tag{A96}$$

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

$$\phi_{i,0} = -y_i^2 - \mathbb{V}\left(\hat{f}(x_{i,1})\right). \tag{A97}$$

Replacing the expression of $\phi_{i,0}$ in Equation (A95), we obtain:

$$\phi_{i,1} = 2y_i\hat{f}(x_{i,1}) + \mathbb{V}\left(\hat{f}(x_{i,1})\right) - \hat{f}(x_{i,1})^2. \tag{A98}$$

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:

$$\phi_{i,1} = 2\hat{\varepsilon}_i \hat{f}(x_{i,1}) + \hat{f}(x_{i,1})^2 + \mathbb{V}\left(\hat{f}(x_{i,1})\right). \tag{A99}$$

Now, according to Equations (A16) and (A17), as $\mathbb{E}\left(\hat{f}(x_{i,1})\right) = 0$, we can see that:

$$\hat{f}(x_{i,1}) = \phi_{i,1}^{SHAP}. \tag{A100}$$

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

Finally, from Equations (A99) and (A100), we obtain:

$$\phi_{i,1} = 2\hat{\varepsilon}_i \phi_{i,1}^{SHAP} + \left(\phi_{i,1}^{SHAP}\right)^2 + \mathbb{V}\left(\hat{f}(x_{i,1})\right). \tag{A101}$$

□