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

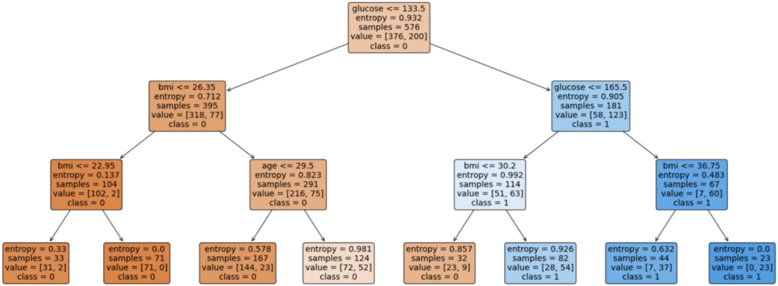
S. Hué [3] C. Hurlin [1] C. Pérignon [2] S. Saurin [1]

[1]University of Orléans [2]HEC Paris [3]AMSE

Econometric Society European Meeting (ESEM)

August 26, 2024

Figure: AUC $= 0.78$



What is the contribution of each feature to the AUC?

We introduce the XPER (eXplainable PERformance) methodology to measure the specific contribution of the input features to the predictive performance of a model.

This method is based on:

- A Shapley Value decomposition (Shapley, 1953).
- A Performance Metric (PM).
- Predictions $\widehat{y}$ of a regression or classification model $f(.)$.

An intuitive primer on XPER

	AUC		ϕ_0		ϕ_1		ϕ_2		ϕ_3
Test sample	0.78	$=$	0.50	$+$	0.14	$+$	0.10	$+$	0.04

with ϕ_0 a benchmark value, and ϕ_j the XPER contribution of feature x_j to the AUC of the model.

Framework and Performance Metrics

We consider a **classification** or a **regression** problem for which:

- A dependent (target) variable denoted y takes values in $\mathcal{Y}$. In case of classification $\mathcal{Y} = \{0, 1\}$, and in case of regression $\mathcal{Y} \subset \mathbb{R}$.
- A q-vector $\mathbf{x} \in \mathcal{X}$ refers to input (explanatory) features, with $\mathcal{X} \subset \mathbb{R}^q$.
- We denote by $f : \mathbf{x} \to \hat{y}$ a model, where $\hat{y} \in \mathcal{Y}$ is either a classification or regression output, such as $\hat{y} = f(\mathbf{x})$.

- The econometric or machine learning model may be parametric or not, linear or not, individual or an ensemble classifier, etc.
- The model is estimated (parametric model) or trained (machine learning algorithm) once for all on an estimation or training sample $\{\mathbf{x}_j, y_j\}_{j=1}^{T}$.
- The statistical performance of the model is evaluated on a test sample $S_n = \{\mathbf{x}_i, y_i, \hat{f}(\mathbf{x}_i)\}_{i=1}^{n}$ for n individuals.

Definition

A sample performance metric $PM_n \in \Theta \subseteq \mathbb{R}$ associated to the model $\hat{f}(.)$ and a test sample S_n is a scalar defined as:

$$PM_n = \tilde{G}_n(y_1, ..., y_n; \hat{f}(\mathbf{x_1}), ..., \hat{f}(\mathbf{x_n})) = G_n(\mathbf{y}; \mathbf{X}),$$

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

Examples:

- Regression model: MSE, MAE, R-squared.
- Classification model: AUC, Accuracy, Sensitivity, Specificity, Brier Score.
- Economic criteria: Profit or cost function.

Assumption 1

The sample performance metric satisfies an additive property such that:

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

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

Assumption 1

The sample performance metric satisfies an additive property such that:

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

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

Example (Mean Squared Error (MSE))

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

Assumption 2

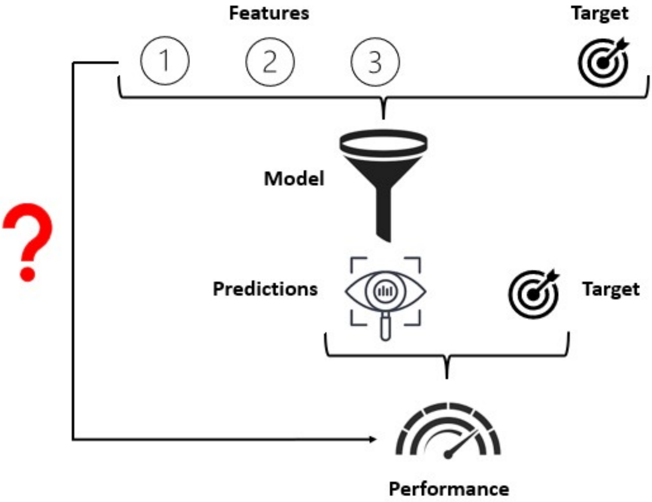
The sample performance metric $G_n(\mathbf{y}; \mathbf{X}; \hat{\delta}_n)$ converges to the population performance metric $\mathbb{E}_{y,\mathbf{x}}(G(y; \mathbf{x}; \delta_0))$, where $\mathbb{E}_{y,\mathbf{x}}(.)$ refers to the expected value with respect to the joint distribution of y and $\mathbf{x}$,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

Example (Mean Squared Error (MSE))

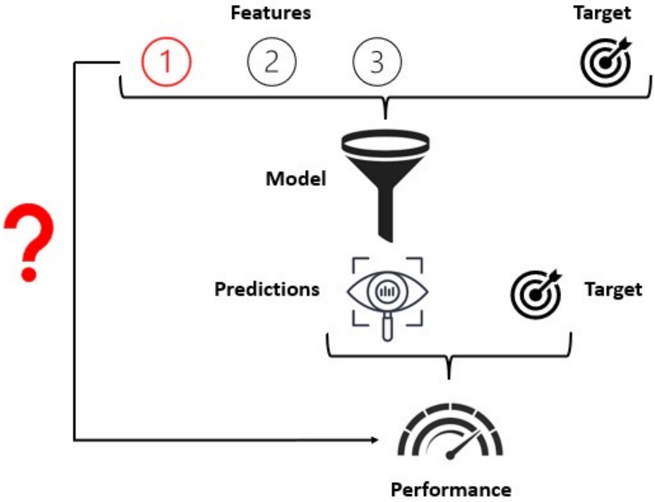
$$\mathbb{E}_{y,\mathbf{x}}(G(y; \mathbf{x}; \delta_0)) = \mathbb{E}_{y,\mathbf{x}}(y - \hat{f}(\mathbf{x}))^2.$$

Theoretical Decomposition

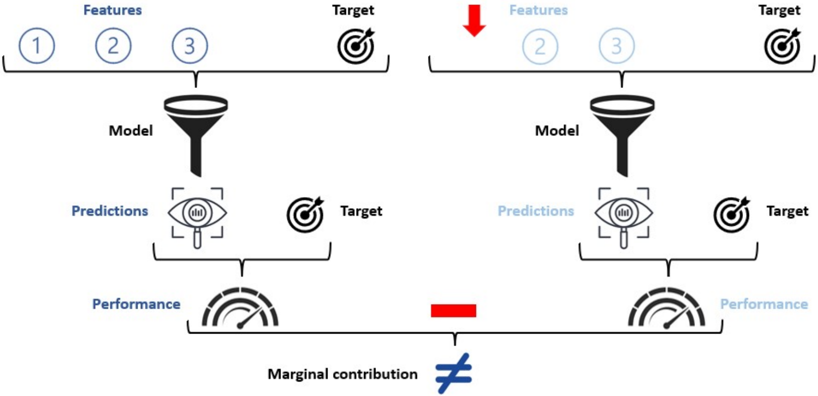
Intuition


Features
1
2
3
Target
Model
Predictions
Target
Performance

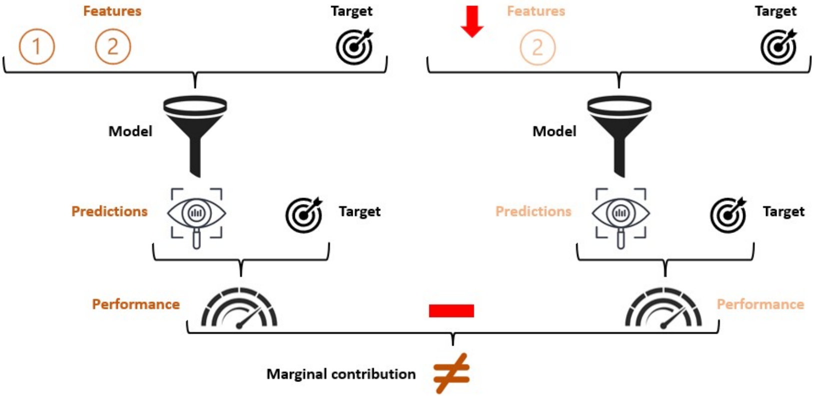
Intuition


Features
Target
1
2
3
Model
Predictions
Target
Performance

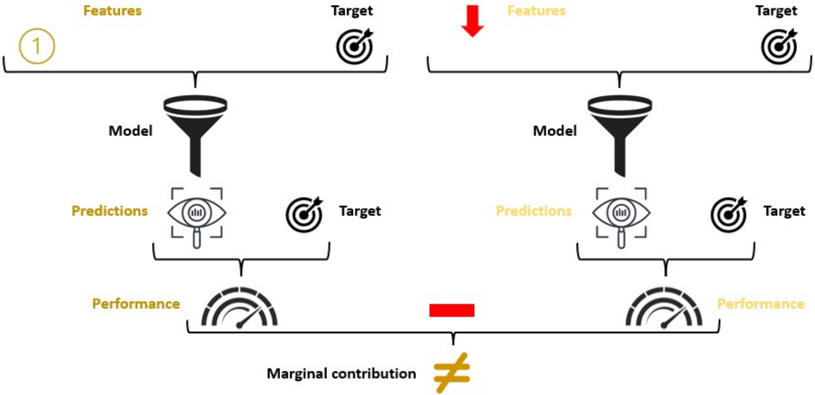
Intuition


Features
1
2
3
Target
Model
Predictions
Target
Performance
Features
2
3
Target
Model
Predictions
Target
Performance
Marginal contribution

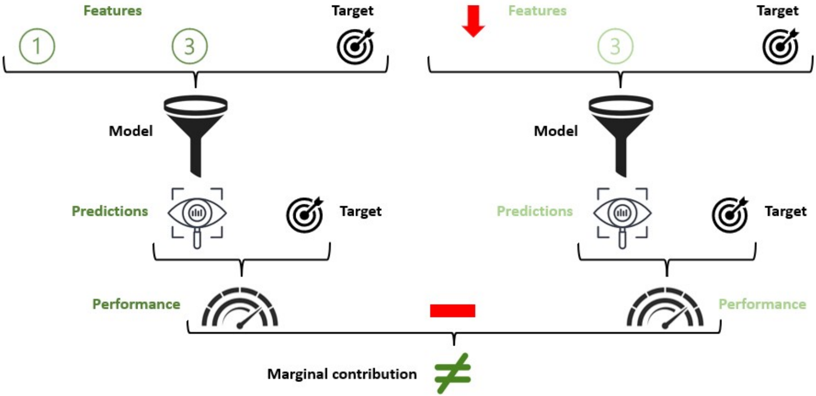
Intuition


Features
1
2
Target
Model
Predictions
Target
Performance
Features
2
Target
Model
Predictions
Target
Performance
Marginal contribution ≠

Intuition


Features
Target
1
Model
Predictions
Target
Performance
Features
Target
Model
Predictions
Target
Performance
Marginal contribution

Intuition


Features
Target
1
3
Model
Predictions
Target
Performance
Features
Target
3
Model
Predictions
Target
Performance
Marginal contribution

Intuition


Coalition 1
Coalition 2
Coalition 3
Coalition 4
1 2 3
1 2
1
1 3
KG
KG
KG
KG
= ϕ_1

Definition (XPER value)

The contribution of feature x_j to the performance metric is:

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

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

The XPER value ϕ_j associated to feature x_j measures its weighted average marginal contribution to the performance metric over all feature coalitions.

Axiom 1. (Efficiency)

The sum of the XPER values ϕ_j, $\forall j = 1, ..., q$ satisfies:

$$\underbrace{\mathbb{E}_{y,\mathbf{x}}(G(y;\mathbf{x};\delta_0))}_{\text{performance metric}} = \underbrace{\phi_0}_{\text{benchmark}} + \sum_{j=1}^{q} \underbrace{\phi_j}_{\text{XPER value}},$$

$$\phi_0 = \mathbb{E}_{\mathbf{x}}\mathbb{E}_{y}\left(G(y;\mathbf{x};\delta_0)\right)$$

with ϕ_0 the performance metric associated to a population where the target variable is independent from all features considered in the model.

Definition (Individual XPER)

The individual XPER value $\phi_{i,j}$ associated to individual i is defined as:

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

For a given realisation $(y_i, \mathbf{x}_i)$, the corresponding individual contribution to the performance metric can be broken down into:

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

where $\phi_{i,j}$ is the realisation of $\phi_{i,j}(y_i;\mathbf{x}_i)$ and $\phi_{i,0}$ is the realisation of $\phi_{i,0}(y_i) = \mathbb{E}_{\mathbf{x}}(G(y_i;\mathbf{x};\delta_0))$.

Example

Empirical Application

Database

Database of auto loans provided by an international bank:

- Target variable y_i:
  - 1: Default
  - 0: No default
- 7,440 consumer loans
- 10 features:
  - 2 categorical features
  - 8 continuous features

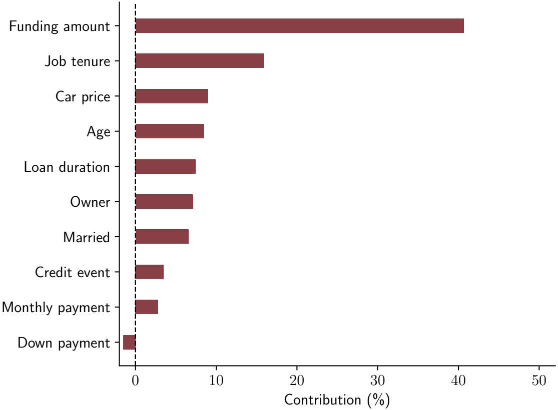
Model

Table: XGBoost Performances

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

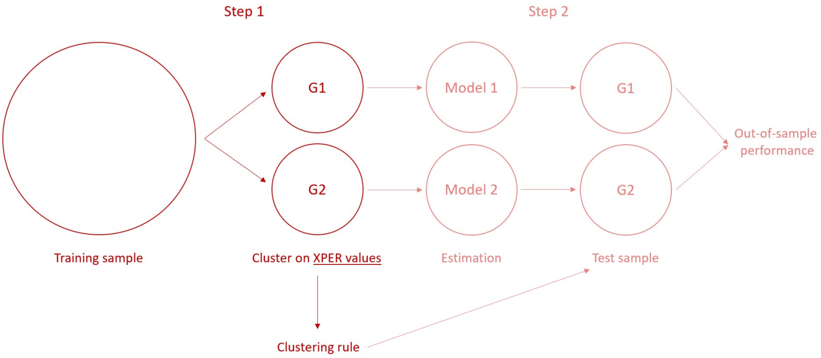
XPER decomposition

Figure: XPER decomposition



Using XPER to boost model performance

Figure: Two-step procedure using XPER



Using XPER to boost model performance

Table: Model performances

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

Conclusion

- We introduce a methodology designed to **measure the feature contributions to the performance** of any regression or classification model.
- Our methodology is **theoretically grounded** on Shapley values and is both model-agnostic and performance metric-agnostic.
- In a loan default forecasting application, XPER appears to be able to significantly **boost out-of-sample performance**.

Python package

Figure: Github link