X-SHAP: towards multiplicative explainability of Machine Learning

Luisa Bouneder
Emerton Data
luisa.nouneder@emerton-data.com

Yannick Léo
Emerton Data
yannick.leo@emerton-data.com

Aimé Lachapelle[*]
Emerton Data
aime.lachapelle@emerton-data.com

Abstract

This paper introduces X-SHAP, a model-agnostic method that assesses multiplicative contributions of variables for both local and global predictions. This method theoretically and operationally extends the so-called additive SHAP approach. It proves useful underlying multiplicative interactions of factors, typically arising in sectors where Generalized Linear Models are traditionally used, such as in insurance or biology. We test the method on various datasets and propose a set of techniques based on individual X-SHAP contributions to build aggregated multiplicative contributions and to capture multiplicative feature importance, that we compare to traditional techniques.

1 Introduction

Interpretation of prediction model outputs can be as important as the prediction of machine learning models, e.g. insurance pricing, credit rejection or acceptance, recommendation to decision markers, medical diagnostic. The users need to understand the factors underlying the prediction. Model interpretability offers the possibility to better audit the robustness and fairness of predictive models. Simple models such as linear regressions or GLMs are quite accurate and easily interpretable. On the contrary, the development of more complex models, such as machine learning ensemble models or deep learning models leads, to highly accurate but more complex models that are difficult to interpret. The trade-off between building a more accurate model vs. keeping a simple and interpretable model is not an easy choice. In many cases, the simple interpretable model is still preferred. In order to solve the accuracy-interpretability trade-off, a large number of interpretable methods have been proposed [17, 10, 20, 24, 11, 4, 6]. It is noteworthy that all these methods focus on additive contributions computation, none of them being able to tackle multiplicative contributions assessment.

In this paper, we introduce, X-SHAP, a model-agnostic interpretability method that provides multiplicative contributions for individual predictions. Our main contributions are summarized as follows:

1. We extend the additive analytical solution to the model-agnostic multiplicative interpretability problem,
2. We introduce X-SHAP, an algorithm that provides approximate multiplicative contributions at individual levels,
3. We propose the X-SHAP toolbox, a new set of techniques used to understand global and segmented model structure by aggregating multiple local contributions,

[*] 16 Avenue Hoche, 75008 Paris, www.emerton-data.com

Preprint. Under review.

4. We empirically verify desirable properties and compare the X-SHAP approach to both the additive algorithm Kernel SHAP, and to well-known metrics on various supervised problems.

2 Related work

The simplest way to interpret any prediction model's outputs is to analyze the model itself when it is not too complex. This is the case for simple models like Generalized Linear Models [14, 3, 12] or decision trees [18], yet, more complex models are not directly interpretable.

To raise adoption of complex models, specific interpretable methods have been developed. Although neural networks have a black box nature, some interpretable approaches exist [4, 20]. For instance, DeepLIFT [20] (Deep Learning Important FeaTures) decomposes the output prediction of a neural network on a specific input by backpropagating the contributions of all neurons in the network to each feature of the input. In order to interpret tree based machine learning ensemble models such as random forests or gradient boosting, Lundberg et al. [11] proposes a polynomial time explainer based on game theory that measures local feature interaction effects.

There are two types of model-agnostic interpretability methods. The first type consists of finding the training points that are most responsible for the prediction [8, 6]. The second type of general explainer performs a local linear regression around the prediction and extracts contributions from local linear models [17]. In this case, when features are not independent, contributions are produced via Shapley values, a concept in cooperative game theory, introduced in [19], that assigns a unique distribution (among the players) of a total surplus generated by the coalition of all players. These are the SHAP methods [24, 10].

In many fields such as actuarial [3, 7], epidemiology [9], economy [22] and medicine [13] phenomenon are multiplicative by nature, very often with a traditional use of models (e.g. log-GLM), the available interpretability methods provide additive interpretations. Little attention has been paid to multiplicative contributions assessment despite the existence of theoretical extension of additive Shapley values [23] to multiplicative provided by Ortmann [15] to positive cooperative games. In this paper, we propose to fill this gap by extending the Kernel SHAP local interpretation method to multiplicative problems.

3 Problem and notations

3.1 Model-agnostic interpretability problem

Let X be an input dataset composed of n observations x_i and m features where $X = \{{x_i}^j\}$ with $\forall i \in [1, n], \forall j \in [1, m], {x_i}^j \in \mathbb{R}$. x_i refers to a single observation of the dataset X. The set of features $\{j\}_{j \in [1,m]}$ is noted F. Let us introduce a strictly positive target vector $Y = \{y_i\}_{i \in [1,n]}$ such that $\forall i \in [1, n], y_i > 0$. Let f denotes the associated predictive model $f : \mathbb{R}^k \to \mathbb{R}^{+*}, \forall i \in [1, n], \hat{y}_i = f(x_i)$. Let us assume that the predictive model f is already trained on the dataset (X_{train}, Y_{train}) with same properties as (X, Y).

The usual method used to explain machine learning models is the additive contributions of features.

Definition 1. Additive feature contributions. Let f be a predictive model associated with (X, Y) and x_i a single observation of X with $\hat{y}_i = f(x_i)$. The prediction of x_i can be decomposed by the sum of the additive feature contributions:

$$\phi^0 + \sum_{j=1}^{m} \phi_i^j(x_i) = f(x_i) = \hat{y}_i \tag{1}$$

where ϕ^0 is a baseline value for predictions, independent of the observations explained, m is the number of features, ϕ_i^j is the additive contribution of feature j to the model prediction $\hat{y}_i$ for the observation x_i. ϕ or ϕ_f denotes the set of additive contributions related to f.

In this paper, we focus on use the multiplicative contributions of features.

Definition 2. Multiplicative feature contributions. Let f be a predictive model associated with (X, Y) and x_i a single observation of X with $\hat{y}_i = f(x_i)$. The prediction of a single observation x_i, also refers to x to simplify, can be decomposed by the product multiplicative feature contributions:

$$\psi^0 \times \prod_{j=1}^{m} \psi_i^j(x_i) = f(x_i) = \hat{y}_i \tag{2}$$

where ψ^0 is a baseline value for predictions, independent of the observations explained, m is the number of features, ψ_i^j is the multiplicative contribution of feature j to the model prediction $\hat{y}_i$ for the observation x_i. We note ψ or ψ_f, the set of multiplicative contributions related to f.

Model-agnostic interpretability problem feeds as follows: given any predictive model f associated with the dataset (X, Y), the multiplicative (resp. additive) model-agnostic interpretability problem consists of finding, for any prediction $(x_i, \hat{y}_i)$, a multiplicative (resp. additive) feature contributions ψ (resp. ϕ).

3.2 Notation and definitions

Notation 1. Arithmetic and geometric means. Considering n real values $\forall i \in [1, n], x_i \in \mathbb{R}$, the arithmetic mean is noted $< x >_+ = \frac{1}{n} \sum_{i=1}^{n} x_i$ and the geometric mean is noted $< x >_\times = (\prod_{i=1}^{n} x_i)^{\frac{1}{n}}$

Definition 3. Coalition vector. We define the coalition vector c of dimension m as a simple binary vector $c \in [0, 1]^m$ representing a set of activated features with c of F. The complementary coalition vector, noted $\bar{c}$, is defined as follows: $\forall j \in [1, m], \bar{c}^j = 1 - c^j$. c can also be noted c_k when multiple coalitions have to be enumerated.

Definition 4. Sub-observation and sub-dataset. Considering an observation x_i of a dataset X and a coalition vector $c \subset F$, the induced sub-observation is given by ${x_i}^c = x_i \times c$. One can extend to the sub dataset $X^c = X * c$.

Definition 5. Augmented observation. Considering an observation x_i of size m of a dataset X, the augmented dataset is defined as the duplicate (n times) of x_i: $X_i = x_i \times \mathbb{1}_n$. Thus, the size of the matrix X_i is $n \times m$

Definition 6. Perturbated coalition dataset. Considering an augmented observation X_i of an observation $x_i \in X$ and a coalition vector c of F, we define the perturbated coalition as $\mathrm{M}^c(X, x_i) = X_i^c + X^{\bar{c}}$

4 Short review of the Kernel SHAP method used for additive contributions

Before introducing the X-SHAP method, end for the sake of comparison and clarity, we remind the Kernel SHAP method from which it is derived.

A theoretical solution to the additive version of the model-agnostic interpretation problem is introduced in [23, 24, 10]. It shows that additive Shapley values defined in eq. (3) are the unique solution to the additive model-agnostic interpretability problem defined in section 3.1 that respects local accuracy, missingness and consistency properties defined in [10]. The solution is given by:

$$\phi^j(x) = \sum_{c \subset F \setminus \{j\}} \frac{|c|!(|F| - |c| - 1)!}{|F|!} (f_{c \cup \{j\}}(x_{c \cup \{j\}}) - f_c(x_c)) \tag{3}$$

where x is the considered observation, f the predictive model and $f_c(x_c)$ is the prediction of the model restrained to the space of features c applied to sub-observation x_c, F is the set of features. For all coalitions, combinatory fractions are noted as the weights W.

In practice, for a given dataset X, the additive contribution of a feature j is averaged among multiple observations. It can be proven that the problem of computing the Shapley value is an NP-complete problem. Therefore, Lundberg and Lee [10] propose the Kernel SHAP method to approximate the additive feature contributions $\tilde{\phi} \approx \phi$. A python library of the Kernel SHAP algorihtm is implemented and available[2]. To do so, [10] makes two main simplifications:

1. First, in order to obtain linear computation of the Shapley values, as proposed in Castro et al. [5], not all the coalitions are enumerated. The selection of coalitions is done in order of importance in the Shapley values formula (eq.(3)) measured by the weights W. First come coalitions of size 1 (all singletons) and their respective complementary (of size $m-1$), then all coalitions of size 2 paired with their complementary (of size $m-2$), and so on
2. Second, a representative sample X^{ref} of the whole dataset X containing $n^{ref} << n$ observations is considered to compute contributions. Thus, the average reference target value is $\hat{y}^{ref} =< f(X^{ref} >_+$

Then, in order to compute the additive contributions of an observation x_i, the perturbated coalition dataset $\mathrm{M}^c(X, x_i)$ is built for each coalition $c \in C$ as follows: $\forall c \in C, \mathrm{M}^c(X^{ref}, x_i) = {X_i}^c + {X^{ref}}^{\bar{c}}$. The average coalition target value is obtained by applying f on the perturbated coalation dataset and averaging: $\hat{y}^c(x_i) =< f(\mathrm{M}(X, x_i)^c) >_+$. For each coalition, the gap between the coalition target value and the reference target value $\Delta^c(x_i) = \hat{y}^c(x_i) - \hat{y}^{ref}$ intuitively captures the impact of the coalition c. Therefore, the last step of the Kernel SHAP method consists of applying a weighted linear regression on $\Delta(x_i) = \{\Delta^c(x_i)\}_{c \in C}$ to compute the approximated additive feature contributions. The closed form for the weighted regression is:

$$\tilde{\phi}(x_i) = (W \cdot C^T C)^{-1} W \cdot C^T \Delta(x_i) \tag{4}$$

where $\tilde{\phi}$ is the estimated additive contributions of f for the observation x_i from Kernel SHAP method. As the coalitions C are selected by order of weights W in the Shapley values formula, the approximation $\tilde{\phi} \approx \phi$ is verified in practice if a sufficient number of coalitions is selected.

5 Generalization to multiplicative contributions, X-SHAP

5.1 Theoretical extension: analytical solution to multiplicative contributions problem

The X-SHAP algorithm adapts the Kernel SHAP method to multiplicative feature contributions. Thanks to the theoretical extension of the Shapley values, developed in Ortmann [15] in game theory, we easily extend the solution and desirable properties to the model-agnostic interpretabiltity problem.

In this section, we show that there is a unique solution of the multiplicative model-agnostic interpretability problem that verifies the geometrical efficiency (also refered to as local accuracy by Lundberg and Lee [10]) and preserving-ratios properties.

Property 1. ***(Local accuracy)*** *Taking a predictive model f associated with a dataset (X, Y), the associated contributions function ψ is geometrically efficient if it verifies the relation:*

$$\forall i \in [1, n],\ \psi^0 \times \prod_{j=1}^{m} \psi_i^j(x_i) = f(x_i) = \hat{y}_i \tag{5}$$

Property 2. ***(Preserving-ratios)*** *For all f and (X, Y), the associated contributions ψ is said to preserve ratios when one has:*

$$\forall x \in X, \forall j_1 \neq j_2, \frac{\psi^{j_1}(x)}{\psi^{j_1}(c \setminus \{j_2\}, x)} = \frac{\psi^{j_2}(x)}{\psi^{j_2}(c \setminus \{j_1\}, x)} \tag{6}$$

Theorem 1. *For any predictive model f associated with a dataset (X, Y), there is a unique multiplicative feature contributions ψ that is geometrically efficient and preserves ratios for the predictive model f and for any observations x X. The solution is given by:*

$$\psi^j(x) = exp(\sum_{c \subset F \setminus \{j\}} \frac{|c|!(|F| - |c| - 1)!}{|F|!} (\ln(f_{c \cup \{j\}}(x_{c \cup \{j\}})) - \ln(f_c(x_c)))) \tag{7}$$

[2] `https://github.com/slundberg/shap`

Definition 7. Given a predictive model f and a dataset (X, Y) and an observation x, a feature $j \in [1, m]$ is called inessential, if for every coalition $c \in F, j \notin c$, one has $f_{c\cup\{j\}}(x_{c\cup\{j\}}) = f_c(x_c)$

Corollary 1. *Given a predictive model f associated with a dataset (X, Y) and j an inessential feature. Then, the contribution of the feature j, $\psi^j(x) = 1$.*

5.2 Practical extension: the X-SHAP algorithm

Following the theoretical generalization of additive contributions to multiplicative contributions, X-SHAP extends the computation of the approximate multiplicative contributions $\tilde{\psi}(x_i)$ of each prediction $x_i \in X$: $\tilde{\psi}^0 \times \prod_{j=1}^{m} \tilde{\psi}^j(x_i) = \hat{y}_i$. While facing the same computational challenges, Thus, the algorithm X-SHAP (Algorithm 1) follows similar initial steps as the SHAP, such as building a representative reference dataset X^{ref} and selecting the coalitions C with greatest weights. Then, as the predictive model is multiplicative, the whole algorithm of the Kernel SHAP has to be consequently adjusted. Thus, the arithmetic mean is transformed into geometric mean and the linear regression to a logarithm-generalized linear regression. The details of the algorithm are developed in Algorithm 1.

Algorithm 1: X-SHAP for computing multiplicative feature contributions for a single observation x_i following additive Kernel SHAP implementation [10]

```
function **x_shap_explainer** $(f, x_i, X^{ref}, C, W)$:
    **Input** : $f$ the predictive function of the model, $x_i$ the observation to interpret, $X^{ref}$ the reference
              dataset, $C$ the $K$ selected coalitions, $W$ the associated weights of the coalitions
    **Output** : Vector $\tilde{\psi}_f(x_i)$ of X-SHAP contributions
    $\hat{y}^{ref}_{\times} \leftarrow < f(X^{ref}) >_{\times}$ // Average reference target value
    $X_i \leftarrow x_i \times \mathbb{1}_n$ // Augmented observation
    $\mathrm{M}^{c_k}(X, x_i) \leftarrow X_i^{c_k} + X^{\bar{c_k}}, \forall\, c_k \; in \; C$ // Pertubated coalition datasets
    $\hat{y}^{c}_{\times}(x_i) \leftarrow < f(\mathrm{M}^{c}(X, x_i) >_{\times}, \forall\, c_k \; in \; C$ // Coalition average target values
    $\Delta_{\times}(x_i) \leftarrow (\hat{y}^{c_1}_{\times}(x_i)/\hat{y}^{ref}_{\times}, ..., \hat{y}^{c_K}_{\times}(x_i)/\hat{y}^{ref}_{\times})$ // Coalitions-reference gaps
    $C_s \leftarrow feature\_selection(C)$ // Feature selection using Lasso (optional)
    $\tilde{\psi}_f(x_i) \leftarrow \exp((W \cdot C_s^T C_s)^{-1} W \cdot C_s^T \ln(\Delta_{\times}(x_i)))$ // GLM to obtain contributions
```

Given a fixed number of selected coalitions, the complexity in time and space is polynomial.

5.3 Interpretation

Impact interpretation. X-SHAP measures the multiplicative factor associated with a feature j of the observation x_i. If the X-SHAP contribution $\psi^j(x_i) > 1$, the value of feature j in observation x_i increases the model prediction compared to the baseline. On the contrary, when $\psi^j(x_i) < 1$, the feature value decreases the model prediction from baseline. Finally, if $\psi^j(x_i) = 1$, the feature is inessential and thus impactless.

Link with log-GLMs. In the specific case where the predictive model f is a logarithmic Generalized Linear Model such as $\hat{y}_i = \exp(\alpha) \times \prod_{j=1}^{k} \exp(\beta^j \times x_i^j)$ where $\hat{y}_i$ is the prediction for observation x_i, β^j is the coefficient for feature j and α is a constant, the link between the multiplicative feature contributions ψ^j and the coefficients β^j of the GLM regression can be expressed as follows.

Proposition 1. *Let us assume features independence, then one has the following relation between terms of GLM's parameters β^j and contributions $\psi^j(x_i)$:*

$$\forall j \in [1, m], \psi^j(x_i) = \exp(\beta^j \times (x_i^j - < X^j >_+)) \tag{8}$$

As expected, the multiplicative feature contribution measures the impact on the model output of the deviation of x_i^j from expected value in $< X^j >_+$. Therefore X-SHAP allows a reconciliation with log-GLMs.

6 X-SHAP metrics

In addition to the computation of multiplicative contributions, a set of tools is developed including metrics and visualizations. In this section, we present the main metrics used in section results.

Definition 8. X-SHAP multiplicative contributions of a group of observations. The multiplicative contributions $\psi^j(G)$ of a group of distinct observations $G = \{x_i\}_{i \subset [1,n]^{|G|}}$ is defined as the geometric mean of the multiplicative contributions of the observations $\psi^j(x_i)$ expressed as:

$$\psi^j(G) =< \psi_j(x_i) >_{\times, x_i \in G} \tag{9}$$

Definition 9. X-SHAP local feature importance. Let $I^j(x_i)$ denotes the local importance of feature j for observation x_i. It measures the absolute multiplicative impact of the multiplicative contribution on the model's prediction. It is defined as:

$$I^j(x_i) = \max(\frac{1}{\psi^j}(x_i), \psi^j(x_i)) \tag{10}$$

Definition 10. X-SHAP global feature importance. The global feature importance of the feature j, noted I^j, is defined as the geometric mean of local feature importances:

$$I^j =< I^j(x_i) >_{\times, i \in [1,n]} \tag{11}$$

Definition 11. X-SHAP partial dependence. Given a feature j and a range of values $[x_1^j, x_2^j]$ of x^j, the partial dependence of the feature j on $[x_1^j, x_2^j]$ is:

$$PD^j([x_1^j, x_2^j]) = \frac{< \psi^j([x_1^j, x_2^j]) >_\times}{< \psi^j >_\times} \times < \hat{Y} >_\times \tag{12}$$

where $\psi^j([x_1^j, x_2^j])$ is the contribution vector of feature j restricted to values $x_i^j \in [x_1^j, x_2^j], \forall i \in [1, n]$.

7 Data

Three real-world datasets with continuous targets are used to present our results:

- Boston dataset[3]: this dataset contains 13 numerical attributes and 506 observations. The regression task is to predict the median value of owner occupied houses.
- Diabetes dataset[4]: this dataset contains 10 numerical attributes and 442 observations. The regression task is to predict the progression of the disease one year after the baseline.
- Auto Insurance dataset[5]: this dataset contains 23 numerical and categorical attributes and 8161 observations. The task is to predict the severity of motor accidents as an expected material claim amount.

Boston and Diabetes datasets are both sets for which the regression problem is easily solved. Moreover they both have a small number of features. These two characteristics make them good candidates to check the coherence and performance of the X-SHAP algorithm.

The Auto Insurance dataset has more features. It is used to test the X-SHAP method on a real-world example when modeling experts (e.g. actuaries) would typically use GLMs in order to explore the multiplicative effects.

Each dataset is randomly split into a train set (70% of original size) and a test set. Both a random forest regressor (RF) and a gradient boosting (GB) are fit on the training sets.

The reference data is taken from the training set and the X-SHAP values are computed on the test set.

8 Results

We analyze the results from different perspectives (local, global, and segmented) in order to verify the consistency between X-SHAP explanations, classical explanations tools and intuition.

[3] https://archive.ics.uci.edu/ml/machine-learning-databases/housing/

[4] https://www4.stat.ncsu.edu/~boos/var.select/diabetes.tab.txt

[5] https://www.kaggle.com/c/auto-insurance-fall-2017/data

Precision of approximations. First, we implement sanity checks to observe empirically properties satisfied by X-SHAP contributions:

1. Local accuracy (property 1) is verified for predictions of the three datasets. The products of all the contributions are equal to the prediction with a mean percentage error $< 10^{-16}$
2. The estimation of the analytical multiplicative contributions (eq. 1) performed by the X-SHAP algorithm is accurate as soon as a sufficient number of coalitions is selected. We observe a quick convergence to analytical contributions. With the three datasets, the stability of the computations is reached when $|C| > 500$.

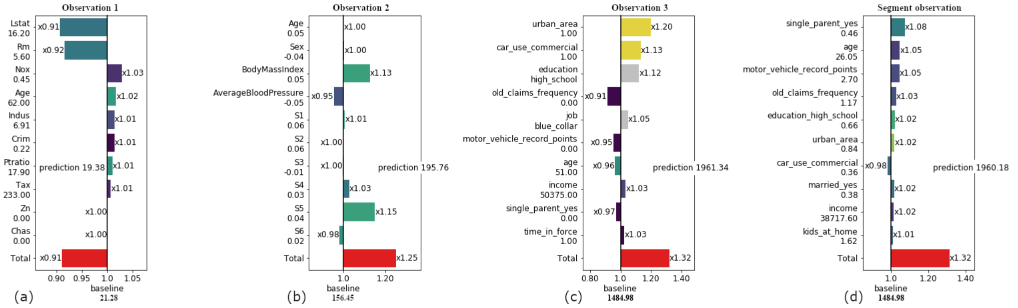

Figure 1: **X-SHAP multiplicative contributions.** X-SHAP multiplicative feature contributions $\tilde{\psi}^j(x_i)$ of top 10 features from the (a) Boston dataset, (b) Diabetes dataset and (c) Auto Insurance dataset. (d) Multiplicative contributions of young persons $\tilde{\psi}^j(G_{<30yo})$ from the Auto Insurance dataset. Read as follows: in (a), the prediction is 0.91 times the baseline. $Lstat$ which the value is 16.20 decreases the baseline by a factor of 0.91 while the age feature which the value is 62 contributes by $\times 1.02$.

Local explanations. Since X-SHAP provides a multiplicative breakdown of a model predictions, X-SHAP gives the possibility to locally depict, for each prediction $(x_i, \hat{y}_i)$, how the values of the features contribute. In Figure 1), starting from the reference value, the contributions are multiplied and have positive or negative impact on the final result (in red). These impacts depend on each observation value x_i^j.

Summary plots of contributions. We extend SHAP summary plots ([10]) to analyze the impact of feature values to the model's prediction.

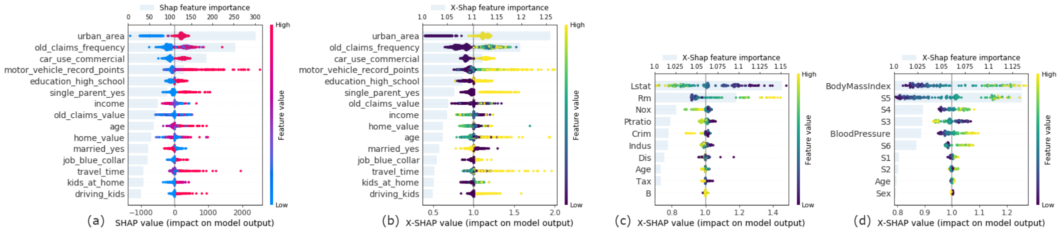

Figure 2: **Summary plots of contributions.** (a) and (b) : comparison of the Kernel SHAP and X-SHAP summary plots for top 15 features for all observations in X_{test} of the Auto Insurance dataset and RF model. (a) Kernel SHAP additive values. (b) X-SHAP multiplicative values. (c) X-SHAP multiplicative values for the Boston data set and RF model. (d) X-SHAP multiplicative values for Diabetes dataset and RF model. Dots represent pairs (contribution, feature). A heatmap associates the underlying feature value. Outliers are not displayed. The underlying bar chart represents the value of the global feature importance of each feature.

Summary plots, depicted in Figure 2, help to visualize how features interact with the model. Figure 2(a) presents the Kernel SHAP value [10] while 2(b) presents X-SHAP values. From these plots we can check consistency between the two algorithms. For most of the features presented there is a clear link between their value and their associated contribution, for example the feature $urban\_area$ identifies whether the person lives a in urban area (high density area). From the X-SHAP summary plot people living in dense areas have a higher average material claim cost than those living in rural areas. Similarly, people with a history of material claim cost ($old\_claims\_frequency$ feature) are more at risk to have material accidents.

Partial dependence of features. Estimating the overall marginal effect of a feature helps to better understand the relation between features and model output. Figure 3 shows the comparison of the X-SHAP partial dependence $PD^j([x_i^j, x_{i+1}^j])$ with the partial dependence, defined in Trevor Hastie [21], for four different features from the Auto Insurance datasets. Both methodologies agree on the behavior of the dependency between the model and the features. Differences in values is mainly due to the way averages are computed: X-SHAP uses a geometric mean which is smaller than the arithmetic mean and less sensitive to outliers.

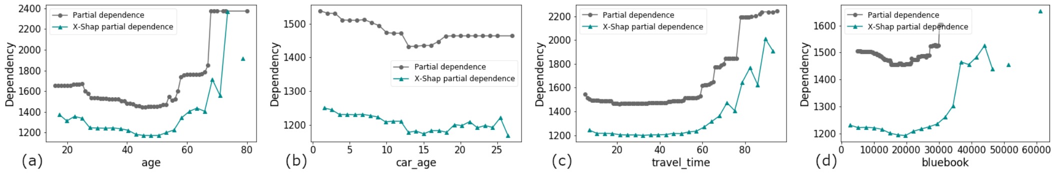

Figure 3: **Partial dependance plots.** Comparison between X-SHAP partial dependence $PD^j([x_1^j, x_2^j])$ (eq. 12) and traditional additive partial dependence [21] over four different features of the Auto Insurance dataset: (a) age, (b) car age, (c) travel time and (d) bluebook. For the X-SHAP dependence plots data was discretized in 25 bins.

Feature importance To understand a model from a global perspective, a used approach is the feature importance. Standard libraries implement such feature importance computation methods. X-SHAP feature importance is computed using the definition 10. The larger the metric, the greater the effect of the feature on the model prediction. Figure 4 compares feature importance of RF model for Diabetes dataset: (a) inner implementation from RF model, (b) Kernel SHAP feature importance (defined as the mean of contribution absolute value), and (c) X-SHAP feature importance. Once again Kernel SHAP and X-SHAP assigns almost the same order of importance (only two order inversions). Moreover X-SHAP results are consistent with intuition since it is commonly acknowledged by experts that Body Mass Index is a major determinant of the evolution of the disease.

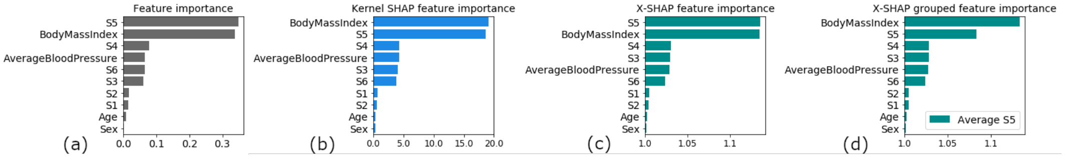

Figure 4: **Feature importance.** Comparison of the results of (a) feature importances given by the RF model, (b) Kernel SHAP feature importances, defined as the mean for each feature $j \in F$ of the absolute value of contribution for all observations i and (c) X-SHAP feature importances I^j. (d) X-SHAP feature importances I^j are depicted for the group of patients having a $S5$ feature value close to the average observed in the cohort.

Interpretation of a group of predictions. X-SHAP contributions can be aggregated to represent a certain group of observations sharing one or more characteristics, thus enabling another explanation level. This level can be adapted for all defined metrics: contributions, partial dependence and feature importances. For instance, Figure 1(d) exhibits the interpretation of the young segment whereas Figure 4(d) presents the X-SHAP feature importance for the patients for which the $S5$ (lamotrigine blood measurement) feature value was close to the average observed in the cohort. While for the whole test set the features Body Mass Index and S5 have a similar effect magnitude, for this specific group there is a clear gap between the importance of these two features.

9 Conclusion

The increased need to providing highly accurate and interpretable multiplicative models has driven the development of X-SHAP, a model-agnostic interpreter that provides local approximations of the multiplicative contributions accompanied with theoretical proofs and empirical checks. In addition, we introduce the X-SHAP toolbox, a new set of tools to analyze local, global and segmented model structure by aggregating multiple local contributions of each or part of individual predictions.

Although the X-SHAP algorithm has a polynomial complexity, interesting opportunities regarding the decrease of complexity in time can arise while exploring the advantage of developing model-specific approximations of the multiplicative contributions for tree based ensemble models or neural networks.

Broader Impact

X-SHAP offers a robust and model-agnostic methodology to assess multiplicative contributions. This unique method strengthens the set of techniques and tools contributing to making machine learning more transparent, auditable and accessible. This method is expected to prove useful for multiplicative underlying structures of modeled phenomena, such as areas where modelers are used to apply log-GLMs (e.g. actuaries modeling claims, epidemiology spreading modeling, disease risk factors estimation, energy consumption forecasting). It is provided as a tool that can help these experts adopt machine learning models with appropriate interpretability framework that stick to their habits.

References

[1] David Alvarez-Melis and Tommi S. Jaakkola. On the robustness of interpretability methods. *CoRR*, abs/1806.08049, 2018. URL `http://arxiv.org/abs/1806.08049`.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica, May*, 23:2016, 2016.

[3] Katrien Antonio and Jan Beirlant. Actuarial statistics with generalized linear mixed models. *Insurance: Mathematics and Economics*, 40(1):58–76, 2007.

[4] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7), 2015.

[5] Javier Castro, Daniel Gómez, and Juan Tejada. Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730, 2009.

[6] Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)*, pages 598–617. IEEE, 2016.

[7] Mark Goldburd, Anand Khare, and Dan Tevet. *Generalized Linear Models For Insurance Rating*. Casualty Actuarial Society, 4350 North Fairfax Drive, Suite 250 Arlington, Virginia 22203, USA, 2016. ISBN 978-0-9968897-3-5.

[8] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1885–1894. JMLR.org, 2017.

[9] Matthias Land and Olaf Gefeller. A multiplicative variant of the shapley value for factorizing the risk of disease. In *Game practice: contributions from applied game theory*, pages 143–158. Springer, 2000.

[10] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017. URL `http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf`.

[11] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):2522–5839, 2020.

[12] Alexander J McNeil and Jonathan P Wendin. Bayesian inference for generalized linear mixed models of portfolio credit risk. *Journal of Empirical Finance*, 14(2):131–149, 2007.

[13] Neil Mehta and Samuel Preston. Are major behavioral and sociodemographic risk factors for mortality additive or multiplicative in their effects? *Social Science & Medicine*, 154:93–99, 2016.

[14] John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

[15] Karl Ortmann. A cooperative value in a multiplicative model. *Central European Journal of Operations Research*, 21(3):561–583, September 2013. doi: 10.1007/s10100-012-0247-6. URL `https://ideas.repec.org/a/spr/cejnor/v21y2013i3p561-583.html`.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[17] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016. URL `http://arxiv.org/abs/1602.04938`.

[18] S Rasoul Safavian and David Landgrebe. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3):660–674, 1991.

[19] Lloyd S Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28): 307–317, 1953.

[20] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3145–3153. JMLR. org, 2017.

[21] Jerome Friedman Trevor Hastie, Robert Tibshirani. *The Elements of Statistical Learning*, chapter 10.13.2. Springer-Verlag New York, 2009. ISBN 978-0-387-84857-0.

[22] H Wang, BW Ang, and Bin Su. Multiplicative structural decomposition analysis of energy and emission intensities: Some methodological issues. *Energy*, 123:47–63, 2017.

[23] H Peyton Young. Monotonic solutions of cooperative games. *International Journal of Game Theory*, 14(2):65–72, 1985.

[24] Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41:647–665, 12 2013. doi: 10.1007/s10115-013-0679-x.

A Proofs

Proof of Theorem 1.

Proof. The proof of Theorem 1 can be directly deducted from the results of Ortmann (2012) [15] while traducing game theory problem to the interpretability problem. It originally derives direclty from the unicity of the additive shapley values. On one hand, the proof of the geometrical efficiency can be derived from additive version as follows:

$$
\begin{aligned}
&\forall i \in [1,n], \phi^0 + \sum_{j=1}^{k} \phi_i^{(j)} = \ln(\hat{y}_i) \qquad (13)\\
\Leftrightarrow\ &\forall i \in [1,n], \ln(\psi^0) + \sum_{j=1}^{k} \ln(\psi_i^{(j)}) = \ln(\hat{y}_i)\\
\Leftrightarrow\ &\forall i,\ \exp(\ln(\psi^{(0)}) + \sum_{j=1}^{k} \ln(\psi_i^{(j)})) = \hat{y}_i\\
\Leftrightarrow\ &\forall i,\ \psi^{(0)} \times \prod_{j=1}^{k} \psi_i^{(j)} = \hat{y}_i
\end{aligned}
$$

On the other hand, the proof of preserving-ratios is directly extended from the additive shapley values that preserves differences.

$$
\begin{aligned}
&\forall j_1 \neq j_2, \phi^{j_1}(x) - \phi^{j_1}(c \setminus \{j_2\}, x) = \phi^{j_2}(x) - \phi^{j_2}(c \setminus \{j_1\}, x) \qquad (14)\\
\Leftrightarrow\ &\forall j_1 \neq j_2, \ln(\psi^{j_1}(x)) - \ln(\psi^{j_1}(c \setminus \{j_2\}, x)) = \ln(\psi^{j_2}(x)) - \ln(\psi^{j_2}(c \setminus \{j_1\}, x))\\
\Leftrightarrow\ &\forall j_1 \neq j_2, \exp(\ln(\psi^{j_1}(x)) - \ln(\psi^{j_1}(c \setminus \{j_2\}, x))) = \exp(\ln(\psi^{j_2}(x)) - \ln(\psi^{j_2}(c \setminus \{j_1\}, x)))\\
\Leftrightarrow\ &\forall j_1 \neq j_2, \frac{\exp(\ln(\psi^{j_1}(x)))}{\exp(\ln(\psi^{j_1}(c \setminus \{j_2\}, x)))} = \frac{\exp(\ln(\psi^{j_2}(x)))}{\exp(\ln(\psi^{j_2}(c \setminus \{j_1\}, x)))}\\
\Leftrightarrow\ &\forall x \in X, \forall j_1 \neq j_2, \frac{\psi^{j_1}(x)}{\psi^{j_1}(c \setminus \{j_2\}, x)} = \frac{\psi^{j_2}(x)}{\psi^{j_2}(c \setminus \{j_1\}, x)}
\end{aligned}
$$

All in all, as the additive Shapley values are the unique solution of the model-agnostic additive interpretable problem that respect local accuracy and preserve differences, the multiplicative Shapley values are the unique solution of the model-agnostic multiplicative interpretable problem that both respect local accuracy and preserve ratios. □

Proof of Proposition 1.

Lemma 1. *Given a predictive log-GLM model f associated to a dataset (X, Y) and ψ the multiplicative shapley values, the relation between ψ^0 and f is:*

$$
\psi^0 = \exp(\alpha) \prod_{j=1}^{k} \exp(\beta^j \times X^j) \qquad (15)
$$

Proof. Starting from the known relation in additive version between ϕ^0 and f:

$$
\begin{aligned}
&\phi^0 =< \hat{y}_i >_+\\
\Leftrightarrow\ &\psi^0 =< \hat{y}_i >_\times\\
\Leftrightarrow\ &\psi^0 = (\prod_{i=1}^{n} \exp(\alpha) \times \prod_{j=1}^{k} \exp(\beta^j \times x_i^j))^{\frac{1}{n}}\\
\Leftrightarrow\ &\psi^0 = \exp(\alpha) \prod_{j=1}^{k} \exp(\beta^j \times \sum_{i=1}^{n} \frac{x_i^j}{n})\\
\Leftrightarrow\ &\psi^0 = \exp(\alpha) \prod_{j=1}^{k} \exp(\beta^j \times < X^j >_+)
\end{aligned}
$$

□

Data set	Model	MSE	R2	mean_APE	median_APE	std_APE	max_APE
Boston	RF	$4.96e-29$	1.0	$2.33e-16$	$2.00e-16$	$1.69e-16$	$7.28e-16$
	GB	$3.44e-29$	1.0	$1.84e-16$	$1.70e-16$	$1.64e-16$	$5.96e-16$
Diabetes	RF	$2.95e-27$	1.0	$2.79e-16$	$2.40e-16$	$2.15e-16$	$9.54e-16$
	GB	$2.81e-27$	1.0	$2.61e-16$	$2.14e-16$	$1.90e-16$	$8.80e-16$
Auto ED	RF	$6.04e-25$	1.0	$3.10e-16$	$2.52e-16$	$2.41e-16$	$1.5e-15$
	GB	$6.83e-25$	1.0	$3.39e-16$	$2.94e-16$	$2.69e-16$	$2.22e-15$

Table 1: **Local accuracy.** Scores of X-SHAP estimated contributions output against model predictions $\hat{y}$ for all three data sets and two models tested

Lemma 2. *Given a predictive log-GLM model f associated to a dataset (X, Y) and ψ the multiplicative shapley values, the relation between $\psi^j, \forall j \in [1, m]$ and f is:*

$$\prod_{j=1}^{m} \psi^j(x_i) = \prod_{j=1}^{m} \exp(\beta^j \times (x_i^j - \bar{X}^j)) \tag{16}$$

Proof. Introducing the expression of ψ^0 using the GLM's parameters found in eq. (15) into the two definitions of $\hat{y}_i$ (using log-GLM definition and feature contribution in eq. (2)) gives the following proof:

$$\hat{y}_i = \psi^0 \times \prod_{j=1}^{k} \psi_i^j = \exp(\alpha) \times \prod_{j=1}^{k} \exp(\beta^j \times x_i^j)$$

$$\Leftrightarrow \exp(\alpha) \prod_{j=1}^{m} \exp(\beta^j \times \bar{X}^j) \times \prod_{j=1}^{m} \psi^j(x_i) = \exp(\alpha) \times \prod_{j=1}^{m} \exp(\beta^j \times x_i^j)$$

$$\Leftrightarrow \prod_{j=1}^{m} \psi^j(x_i) = \prod_{j=1}^{m} \exp(\beta^j \times (x_i^j - < X^j >_+))$$

□

The proof of the **Proposition 1** is then straight forward.

Proof. From Lemma 2 and assuming the feature independence in eq. (16), the proof is straight forward:

$$\prod_{j=1}^{k} \psi^j(x_i) = \prod_{j=1}^{k} \exp(\beta^j \times (x_i^j - < X^j >_+))$$

$$\Rightarrow \forall j \in [1, m] \psi^j(x_i) = \exp(\beta^j \times (x_i^j - < X^j >_+))$$

□

B Precision of approximations.

We implement sanity checks to observe empirically properties satisfied by X-SHAP contributions.

Local accuracy. Property 1 is verified for predictions of the three datasets. The products of all the contributions are equal to the prediction with a mean percentage error $< 10^{-16}$ (see Table 1)

Precision. The X-SHAP algorithm performs an approximation of the analytical multiplicative contributions (eq. 1). According is accurate as soon as a sufficient number of coalitions is selected. We observe a quick convergence to analytical contributions. With the three datasets, the stability of the computations is reached when $|C| > 500$.