From Shapley Values to Generalized Additive Models and back

Sebastian Bordt
University of Tübingen
sebastian.bordt@uni-tuebingen.de

Ulrike von Luxburg
University of Tübingen
ulrike.luxburg@uni-tuebingen.de

September 12, 2022

Abstract

In explainable machine learning, local post-hoc explanation algorithms and inherently interpretable models are often seen as competing approaches. In this work, offer a novel perspective on Shapley Values, a prominent post-hoc explanation technique, and show that it is strongly connected with Glassbox-GAMs, a popular class of interpretable models. We introduce n-Shapley Values, a natural extension of Shapley Values that explain individual predictions with interaction terms up to order n. As n increases, the n-Shapley Values converge towards the Shapley-GAM, a uniquely determined decomposition of the original function. From the Shapley-GAM, we can compute Shapley Values of arbitrary order, which gives precise insights into the limitations of these explanations. We then show that Shapley Values recover generalized additive models of order n, assuming that we allow for interaction terms up to order n in the explanations. This implies that the original Shapley Values recover Glassbox-GAMs. At the technical end, we show that there is a one-to-one correspondence between different ways to choose the value function and different functional decompositions of the original function. This provides a novel perspective on the question of how to choose the value function. We also present an empirical analysis of the degree of variable interaction that is present in various standard classifiers, and discuss the implications of our results for algorithmic explanations. A python package to compute n-Shapley Values and replicate the results in this paper is available at `https://github.com/tml-tuebingen/nshap`.

1 Introduction

Local post-hoc explanation algorithms and inherently interpretable models are two of the most prominent approaches in explainable machine learning (Molnar, 2020; Holzinger et al., 2022). Despite a number of arguments about their relative benefits, the differences and similarities between the two remain largely unresolved (Rudin, 2019). In particular, most post-hoc explanations can be computed for any state-of-the art model (Ribeiro et al., 2016; Lundberg and Lee, 2017). In contrast, inherently interpretable models might or might not achieve competitive accuracy on a given task (Rudin et al., 2022). But does the performance of post-hoc explanations depend on whether the model is simple?

In this work, we explore this question in the context of Shapley Values (Lundberg and Lee, 2017), a prominent post-hoc explanation technique, and Glassbox-GAMs (Lou et al., 2012; Agarwal et al., 2021), a popular class of interpretable models. We show that Shapley Values are perfectly faithful to Glassbox-GAMs (defined as functions with additive variable interactions), in these sense that they recover the individual non-linear component functions of the GAM. For any other class of models, Shapley Values necessarily compress information about variable interactions.

More generally, we introduce n-Shapley Values, a natural extension of Shapley Values and Shapley Interaction Values (Lundberg et al., 2020) that explain individual predictions with interaction terms

up to order n. The idea behind this extension is to demonstrate that as explanations become more complex and expressive, they are also able to faithfully explain more complex models. And indeed: n-Shapley recover generalized additive models with variable interactions up to order n in the same way that Shapley Values recover Glassbox-GAMs. This result demonstrates that feature attributions can, at least in the context of Shapley Values, be seamlessly extended to more complex notions of explanations. In high dimensions, n-Shapley Values face computational limitations, but we suggest that they might be useful to debug and assess Shapley Values on low-dimensional problems. For this we release an accompanying python package.

Our results are based on a number of theoretical connections between Shapley Values and additive decompositions of functions. Developing these connections occupies the bulk of the paper. The basis for these connections turns our to be the value function. All Shapley Values are based on value functions, and suitable value functions give rise to a uniquely determined functional decomposition of the original function that we term the Shapley-GAM (Theorem 4). The Shapley-GAM is the limit of n-Shapley Values as n tends towards the number of features, and its usefulness stems from the fact that it highlights the properties of Shapley Values quite clearly. Most importantly, Shapley Values of any order are simple linear combinations of the component functions of the Shapley-GAM (Theorem 5). This provides an alternative motivation for Shapley Values that does not rely on economic game theory, and also links feature attributions with the tools developed in the statistics literature on functional decompositions (Hooker, 2007; Hiabu et al., 2022; Herren and Hahn, 2022). Somewhat surprisingly, *any* functional decomposition of the original function corresponds to Shapley Values as a model explanation technique, if we only choose the value function appropriately (Theorem 6).

Since Shapley Values turn out to be most faithful for models with few interactions, we also study the degree of variable interaction that is present in standard classifiers such as gradient boosted tress and random forests. While our investigation of this question remains preliminary, our results suggest that there is not direct link between accuracy and the average degree of variable interaction in the Shapley-GAM (Section 7).

Taken together, our results offer a fairly precise functionally-grounded analysis of Shapley Values (Doshi-Velez and Kim, 2017). We also believe that they provide an instructive example for how to think about post-hoc explanations (Covert et al., 2021; Krishna et al., 2022). For example, while Shapley Values are guaranteed to recover Glassbox-GAMs, it is impossible to tell from a single local explanation whether it originates from a Glassbox-GAM or any other model. This can only be seen from an aggregation of multiple local explanations (Lundberg et al., 2020), or alternatively, n-Shapley Values of higher order.

2 Related Work

Shapley Values. The seminal paper by Lundberg and Lee (2017) has led to a line of work which investigates the usage of Shapley Values in machine learning (Chen et al., 2020; Heskes et al., 2020; Slack et al., 2020; Albini et al., 2022). There have been extensions of the concept of Shapley Value to incorporate variables interactions Lundberg et al. (2020); Tsai et al. (2022), and debates about the choice of value function (Sundararajan and Najmi, 2020; Janzing et al., 2020). How to efficiently compute Shapley Values is also a question of ongoing interest (Lundberg et al., 2020; Jethani et al., 2021). Shapley Values originate in a literature on economic game theory (Shapley, 1953), and our work is closely related to a particular paper from this literature, namely the seminal work by Grabisch (1997) on additive set functions. Shapley Values have also been explored in various tasks with human decision makers, a topic about which there is much debate (Kumar et al., 2020).

Generalized Additive Models. Generalized additive models originate in statistics (Hastie

and Tibshirani, 1990) and have recently become popular in combination with trees (Lou et al., 2012, 2013) and neural networks (Agarwal et al., 2021). On tabular datasets, interpretable GAMs with few interactions (Caruana et al., 2015) can often achieve competitive accuracy, which has led to an active line of research (Wang et al., 2022; Lengerich et al., 2022). From a statistical perspective, the decomposition of a function as a GAM is underdetermined, which has led to the development of additional uniqueness criteria such as functional ANOVA (Hooker, 2007; Lengerich et al., 2020).

Algorithmic Explanations. Shapley Values are one of many different feature attribution methods (Ribeiro et al., 2016; Sundararajan et al., 2017; Kommiya Mothilal et al., 2021) about which there is a large literature (Lee et al., 2019; Garreau and von Luxburg, 2020; Slack et al., 2021; Covert et al., 2021; Krishna et al., 2022; Han et al., 2022) and much debate (Lipton, 2018; Rudin, 2019; Bordt et al., 2022). Considerable debate also exists around the question whether there is an accuracy-explainability trade-off or a cost of using interpretable models (Rudin, 2019; Moshkovitz et al., 2020). Since our work is exclusively focused on Shapley Values, we do not offer a comprehensive review of the literature on algorithmic explanations. This can be found in many other places (Molnar, 2020; Samek et al., 2021; Holzinger et al., 2022).

3 Background and Notation

Let data points $x \in \mathbb{R}^d$ be distributed according to a probability distribution $\mathcal{D}$. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function that we want to explain. We denote $[n] = \{1, \dots, n\}$ and use subsets of coordinates $S = \{s_1, \dots, s_n\} \subset [d]$, to index both data points $x_S = (x_{s_1}, \dots, x_{s_n})$ and functions $f_S = f_{x_{s_1}, \dots, x_{s_n}}$ where we assume the ordering $s_1 < \dots < s_n$.

3.1 Value Functions and Shapley Values

For a data point $x \in \mathbb{R}^d$ and a subset of coordinates $S \subset [d]$, let $v(x, S)$ be the *value function*. Two value functions of special interest are the *observational SHAP* value function Lundberg and Lee (2017)

$$v(x, S) = \mathbb{E}_{z \sim \mathcal{D}} \left[f(z) \,|\, x_S \right] \tag{1}$$

and the *interventional SHAP* value function (Chen et al., 2020; Janzing et al., 2020)

$$v(x, S) = \mathbb{E}_{z \sim \mathcal{D}} \left[f(z) \,|\, do(x_S) \right]. \tag{2}$$

From the value function, the Shapley Values arise via the well-known Shapley formula (Shapley, 1953). Importantly, different value functions can give rise to different Shapley Values (Sundararajan and Najmi, 2020). We will show that the value function corresponds to a functional decomposition of f if it satisfies $v(x, [d]) = f(x)$ and the following regularity condition.

Definition 1 (Subset-Compliant Value Function)**.** *We say that the value function $v(x, S)$ is subset-compliant if the value $v(x, S)$ depends only on the coordinates of x in S. For a subset-compliant value function, we also write $v(x, S) = v(x_S, S)$.*

3.2 Generalized Additive Models

We employ the following definition of a generalized additive model (GAM) of order n.

Definition 2 (Generalized Additive Model of order n)**.** *We say that $f : \mathbb{R}^d \to R$ is a generalized additive model of order n if f can be written in the form*

$$f(x) = \sum_{S \subset [d],\, |S| \leq n} f_S(x_S) = f_\emptyset + \sum_i f_i(x_i) + \dots + \sum_{i_1 < \dots < i_n} f_{i_1, \dots, i_n}(x_{i_1}, \dots, x_{i_n}). \tag{3}$$

In words, the function f can be described with interaction terms of at most n variables at a time. For $n = 1$, we say that the function f is a Glassbox-GAM (Lou et al., 2012; Caruana et al., 2015). Glassbox-GAMs are often considered interpretable because the feature-wise shape functions can be easily visualized (compare Figure 4). For $n = d$, (3) is called a functional decomposition of f.

4 From Shapley Values to Generalized Additive Models

We now begin our investigation into the relationship between Shapley Values and generalized additive models. We do so by proposing n-Shapley Values, a natural generalization of Shapley Values (Lundberg and Lee, 2017) and Shapley Interaction Values (Lundberg et al., 2020) that is interesting in its own right. We then show that d-Shapley Values represent the original function as a GAM. It turns out that this Shapley-GAM admits a simple analytic representation in terms of the value function.

4.1 n-Shapley Values

For $1 \leq n \leq d$, we define n-Shapley Values recursively as follows. This definition relates to the function f that we want to explain implicitly via the value function.

Definition 3 (n-Shapley Values)**.** *Let $v(x, S)$ be a value function. For any subset of coordinates $S \subset [d]$, consider the measure of contribution*

$$\Delta_S(x) = \sum_{T \subset [d] \setminus S} \frac{(d - |T| - |S|)!|T|!}{(d - |S| + 1)!} \sum_{L \subset S} (-1)^{|S|-|L|} v(x,\, L \cup T). \tag{4}$$

Then, we define n-Shapley values of order $1 \leq n \leq d$ recursively by

$$\Phi_S^n(x) = \begin{cases} \Delta_S(x) & \text{if } |S| = n \\ \Phi_S^{n-1}(x) + B_{n-|S|} \displaystyle\sum_{K \subset [d] \setminus S,\, |K|+|S|=n} \Delta_{S \cup K}(x) & \text{if } |S| < n \end{cases} \tag{5}$$

where the coefficients B_n are the Bernoulli numbers with $B_1 = -1/2$.

The intuition behind this definition is that we successively add higher-order variable interactions to the explanations. This makes the explanations more expressive and also more complex. For $n = 1$, $\Phi_i^1(x)$ are the Shapley Values (Lundberg and Lee, 2017). For $n = 2$, Φ_i^2 and $\Phi_{i,j}^2$ are the Shapley Interaction Values (Lundberg et al., 2020). For $n = d$, $\Phi_i^d, \Phi_{i,j}^d, \ldots, \Phi_{[d]}^d$ are the component functions of the Shapley-GAM (Theorem 4). It turns out that n-Shapley Values enjoy properties similar to the original Shapley Values (efficiency and additivity). However, our motivation for them will be their recovery property (Theorem 7) and straightforward relation to the Shapley-GAM (Theorem 5). More details about n-Shapley Values can be found in Appendix A.

It turns out that n-Shapley Values of lower order can be written as a linear combination of the terms involved in n-Shapley Values of higher order (Appendix Proposition 12). In particular, we can always recover the original Shapley Values by evenly distributing all higher-order interactions back onto the involved features. In Figure 1, we demonstrate that this can be used in order to visualize n-Shapley Values despite their complexity (details in Appendix A.4). Figure 1 depicts multiple n-Shapley Values for a single prediction of a random forest on the Folktables Income classification task (Ding et al., 2021). For this classifier, the number of terms involved in the explanations increases

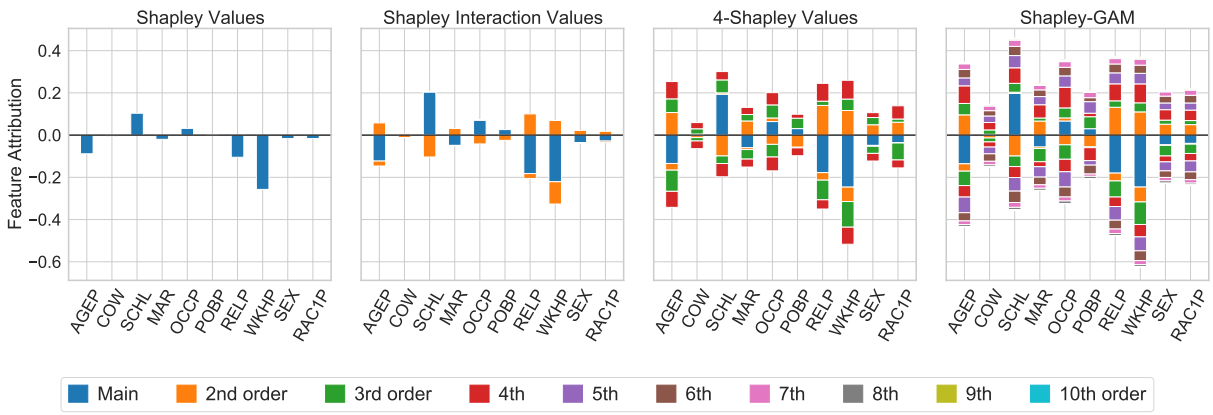


Figure 1: n-Shapley Values generate a sequence of explanations of increasing complexity, ranging from the original Shapley Values to the Shapley-GAM. From left to right: Shapley Values, Shapley Interaction Values, 4-Shapley Values and the Shapley-GAM. The function is a random forest on the Folktables Income classification task, and we used the value function of interventional SHAP. Higher-order interaction effects are distributed evenly onto all involved features, such that the different bars at each feature sum to the Shapley Value (Proposition 12). In each figure, all the different bars sum to the prediction (Section A.4).

significantly with n. Even if the visualizations do not depict all the information that is contained in the n-Shapley Values, they make clear that the original Shapley Values subsume a significant amount of interaction between the different features.

4.2 The Shapley-GAM

If the value function is subset-compliant, then the d-Shapley Values represent the original function as a GAM. This is equivalent to saying that $\Phi^d_S(x)$ is a well-defined function of x_S.

Theorem 4 (The Shapley-GAM)**.** *Let $v(x, S)$ be a subset-compliant value function. Then, the d-Shapley Values represent the original function f as a GAM which we call the Shapley-GAM. In particular, we have*

$$f(x) = \sum_{S \subset [d]} f_S(x_S) \quad \text{with} \quad f_\emptyset = v(\emptyset) \quad \text{and} \quad f_S(x_S) = \Phi^d_S(x) \tag{6}$$

where $\Phi^d_S(x)$ is given by $\Phi^d_S(x) = \sum_{L \subset S} (-1)^{|S|-|L|} v(x_L, L)$. For the special case of the observational SHAP value function (1), the component functions of the Shapley-GAM[1] are given by

$$\begin{aligned} f_\emptyset &= \mathbb{E}[f] & f_{i,j}(x) &= \mathbb{E}[f|x_i, x_j] - \mathbb{E}[f|x_i] - \mathbb{E}[f|x_j] + \mathbb{E}[f] \\ f_i(x_i) &= \mathbb{E}[f|x_i] - \mathbb{E}[f] & f_S(x_S) &= \sum_{L \subset S} (-1)^{|S|-|L|} \mathbb{E}[f|x_L]. \end{aligned} \tag{7}$$

For intuition about Theorem 4, consider Figure 2. It is a well-known fact that the Shapley Value of feature i does not only depend on the value of that feature, but also on the values of

[1]Note that if we fix the point x, then the Shapley-GAM at x is equivalent to the Moebious transform of the measure $v(x, \cdot)$. From this perspective, Theorem 4 can be seen as an application of Theorem 2 in Grabisch (1997).

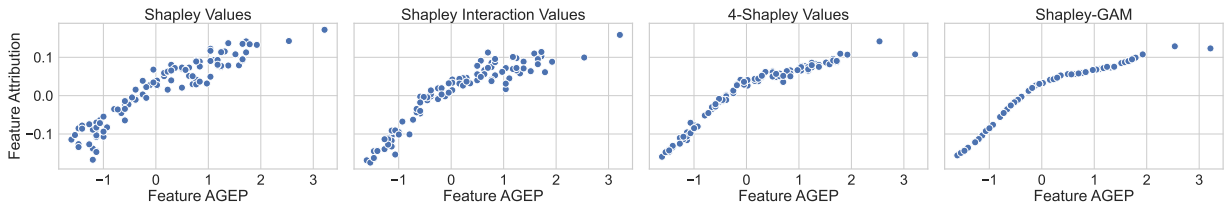


Figure 2: As $n \to d$, the n-Shapley Values converge towards the component functions of the Shapley-GAM. Depcited are the partial dependence plots of $\Phi^n_{\texttt{AGEP}}$ for $n = \{1, 2, 4, 10\}$. The function is a kNN classifier on the Folktables Income classification task. More features in Appendix Figure H.6.

the other features of x (leftmost partial dependence plot in Figure 2). The reason for this is that Shapley Values subsume higher-order variable interactions into the attributions of individual features (according to formula (9), as we will see below). Now, as we successively increase n, more and more variable interactions are appropriately represented in the explanations. This means that they no longer have to be subsumed into lower-order effects, which implies in turn that the lower-order components of the explanations become more distinct (middle parts of Figure 2). For $n = d$, all possible variable interactions can be represented, which implies that the d-Shapley Values become functions of the respective features (rightmost plot in Figure 2).

Note that the Shapley-GAM depends on the value function, which implies that different choices of the value function implicitly correspond to different functional decomposition of f. The significance of this becomes clear when we consider the simple relationship between Shapley Values and the Shapley-GAM in the next section.

5 From Generalized Additive Models to Shapley Values

We now show that n-Shapley Values are nothing but linear combinations of the component functions of the Shapley-GAM. This establishes the significance of the Shapley-GAM, and provides an alternative definition of Shapley Values that does not rely on economic game theory. We also show that for every functional decomposition of f, there is a corresponding subset-compliant value function.

Theorem 5 (n-Shapley Values from the Shapley-GAM)**.** *Let $f(x) = \sum_{S \subset [d]} f_S(x_S)$ be the Shapley-GAM. Then, it holds that*

$$\Phi^n_S(x) = f_S(x_S) + \sum_{K \subset [d] \setminus S,\ n+1 \leq |S|+|K|} C_{n-|S|,|K|}\, f_{S \cup K}(x_{S \cup K}) \tag{8}$$

with coefficients $C_{n,m} = \sum_{k=0}^{n} \binom{n}{k} \frac{B_k}{1+m-k}$. Specifically, the Shapley Value of feature i is given by

$$\Phi^1_i(x) = f_i(x_i) + \frac{1}{2} \sum_{i<j} f_{i,j}(x_i, x_j) + \cdots + \frac{1}{k+1} \sum_{S \subset [d] \setminus \{i\},\, |S|=k} f_{S \cup \{i\}}(x) + \cdots + \frac{1}{d}\, f_{[d]}(x) \tag{9}$$

Theorem 5 specifies the way in which higher-order variable interactions are subsumed into lower-order explanations. In case of the original Shapley Values, this is particularly intuitive: Higher-order effects are evenly distributed onto the involved features.[2] From the perspective of the Shapley-GAM,

[2]For individual value functions, equation (9) is known in the literature on economic game theory (Grabisch, 1997)[Theorem 1]. It was independently re-discovered in Hiabu et al. (2022) and Herren and Hahn (2022).

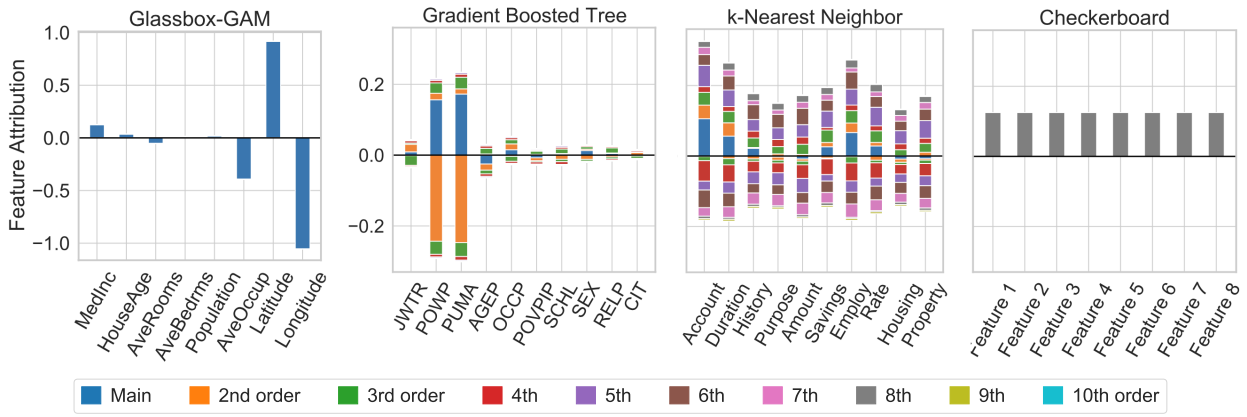


Figure 3: Visualizing the Shapley-GAM of interventional SHAP. Different functions on different datasets require different degrees of variable interaction. (Left) An inherently interpretable GAM on the California Housing dataset. (Middle Left) A gradient boosted tree on the Folktables Travel dataset. (Middle Right) A kNN classifier on the German Credit dataset (Right) The 8-dimensional checkerboard function.

equation (9) describes quite precisely what information that is contained in Shapley Values. In particular, we see that different functions can give rise to the same Shapley Values, and also n-Shapley Values as long as $n < d$ (Grabisch, 2016). We also see that it is impossible to tell from individual Shapley Values if the model consists of main effects or complex variable interactions.

Different functions require different degrees of variable interaction in order to be represented as a GAM. This is illustrated in Figure 3, which depicts the Shapley-GAM of interventional SHAP for four different functions. By definition, a Glassbox-GAM does not require any variable interaction. The other extreme is the n-dimensional checkerboard function, which only consists of interaction terms order n. Learned functions such as random forest, gradient boosted trees and the k-Nearest Neighbor (kNN) classifier lie in between, exhibiting a significant amount of variation between different methods and problems. This last part is also illustrated in additional Figures in Appendix H.

Since we have show that every subset-compliant value function corresponds to a functional decomposition of f, it is natural to ask whether it is also true that every functional decomposition corresponds to a subset-compliant value function. Perhaps somewhat surprisingly, the following Theorem shows that this is indeed the case.

Theorem 6 (From Generalized Additive Models to Value Functions)**.** *Let $f(x) = \sum_{S\subset[d]} g_S(x)$ be any functional decomposition of f. Define the subset-compliant value function*

$$v(x, S) = \sum_{L\subset S} g_L(x). \tag{10}$$

Then the functional decomposition g_S is the Shapley-GAM with respect to the value function (10)*.*

Taken together, Theorem 4 and Theorem 6 establish a bijection between subset-compliant value functions and functional decompositions of f. In a sense, this implies that every functional decomposition implicitly corresponds to a notion of feature importance via its associated value function and the Shapley formula (or, more directly, via equation (9) which is just the same).

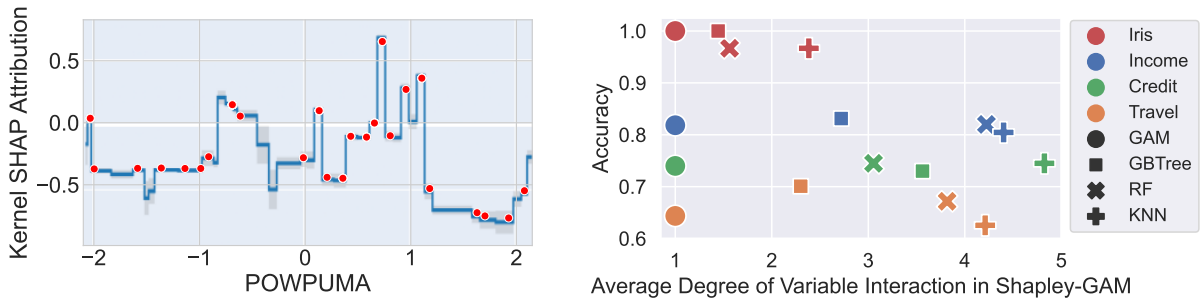


Figure 4: Left: Shapley Values recover Glassbox-GAMs (Theorem 7). To create this figure, we first trained a Glassbox-GAM on the Folktables Travel dataset using the InterpretML package (Nori et al., 2019). Then we computed the Kernel SHAP values for the decision function of the Glassbox-GAM using the `shap` package (Lundberg and Lee, 2017). The Figure depicts the ground-truth variable effect curve of the Glassbox-GAM in blue, and the associated Kernel SHAP values for datapoints from the test set as red dots. Empirically, recovery holds true. Right: An empirical analysis of the average degree of variable interaction in the Shapley-GAM of interventional SHAP for various standard classifiers. The figure depicts accuracy versus the average degree of variable interaction.

6 Shapley Values recover Generalized Additive Models

In this section, we connect Shapley Values with Glassbox-GAMs by showing that n-Shapley Values recover Generalized Additive Models of order n. Together with equation (9), this implies that Glassbox-GAMs are exactly the models for which Shapley Values are most faithful.

Theorem 7 (n-Shapley Values recover GAMs of order n)**.** *Let f be a generalized additive model of order n. Assume that either*

(a) the value function is given by observational SHAP and the input features are independent, or

(b) the value function is given by interventional SHAP.

Then, the n-Shapley Values recover a representation of f as a GAM. Specifically, we have

$$\Phi_S^n(x) = f_S(x_S)$$

where f_S are the component functions of the Shapley-GAM.

Theorem 7 gives a motivation for the usage of Shapley Values does not rely on economic game theory, but on the belief that the model has mostly learned main effects. Unlike our previous results, Theorem 7 depends on the choice of the value function. This is because the recovery property holds if the Shapley-GAM is a GAM or order n - and this depends on the value function.

Figure 4 illustrates Theorem 7 with a Glassbox-GAM that was trained on the Folktables Travel dataset (Caruana et al., 2015). The figure shows the shape curve of the feature POWPUMA in the GAM (blue curve), as well as the associated Kernel SHAP values (red dots). We observe that the Kernel SHAP values lie almost exactly on the shape curve, which implies that recovery property holds fairly precise, even for Kernel SHAP that only approximates the Shapley Values (Lundberg and Lee, 2017).

7 Is there an Accuracy-Complexity Trade-off?

Since Shapley Values turn out to be most faithful for models with few interactions, it becomes interesting to study the degree of variable interaction that is present in various standard classifiers. For low-dimensional problems, this can be done by computing the Shapley-GAM decomposition of different classifiers. Because GAMs of order $n+1$ can represent strictly more functions than GAMs of order n, it is natural to suspect that more accurate classifiers might exhibit higher degrees of variable interaction.

In Figure 4, we depict both the accuracy and the average degree of variable interaction in the Shapley-GAM of interventional SHAP for different classifiers (Glassbox-GAM, gradient boosted tree, random forest, kNN) on four different datasets (Iris, German Credit, Folktables Income, Folktables Travel). Details on the datasets and training procedures are in Appendix G. Note only that we did not attempt to optimize the degree of variable interaction in the random forests or gradient boosted trees. As a first observation, we see that Glassbox-GAMs (dots with variable interaction 1) perform fairly well versus the black-box classifiers, a fact that has been often observed in the literature (Caruana et al., 2015; Agarwal et al., 2021). On the more complex datasets, however, there is usually a model with variable interactions that slightly outperforms the Glassbox-GAM.[3]

Interestingly for us, there is also the example of the kNN classifier. The kNN classifier tends to performs worse on accuracy than the Glassbox-GAM, but exhibits very high degree of variable interaction. This is especially illuminating insofar as the kNN classifier might also be considered inherently interpretable (by explaining the workings of the classifier, and providing the k data points that are responsible for the classification). This implies that a large degree of variable interaction in the Shapley-GAM does not imply that a function is hard to explain per se, but only that the approach taken by Shapley Values - decomposing the function as a GAM - might not lead to a very good representation of the decision boundary.

In our view, it is also questionable whether the relatively high degrees of variable interaction that are exhibited by random forests and gradient boosted tress are actually necessary in order to achieve the respective accuracies. Hence, we conclude that the relation between accuracy and the average degree of variable ineraction in the Shapley-GAM is nuanced: While some degree of interaction seems necessary in order to achieve competitive accuracy, many classifiers seem to exhibit significantly more interaction that that. In some cases, the correlation might even be negative (this is the example of the kNN classifier).

8 Conclusion

This work provides a functionally-grounded characterization of Shapley Values as they are being used in explainable machine learning (Doshi-Velez and Kim, 2017). Some of our result stand in contrast to conventional wisdom around Shapley Values, and offer a novel perspective on local-post hoc explanation algorithms. For example, while Shapley Values depend on the coordinates of x, we have found that they do not contain any information about the local neighborhood of x (Han et al., 2022). In addition, we have seen that Shapley Values are able to faithfully explain non-linear functions, as long as the non-linearity is restricted to the specific form permitted by Glassbox-GAMs. Taken together, this implies that Shapley Values do not perform a local linear approximation of f, which is often believed due to connections with LIME (Ribeiro et al., 2016).

The intimate connections between Shapley Values and Glassbox-GAMs raise the question

[3]The InterpretML package (Nori et al., 2019) allows to include interactions between pairs of variables which reportedly allows to be on pair with black-box models on many data sets.

whether similar results hold for other classes of functions and explanation algorithms. Exploring this question might be a promising direction for future research. Furthermore, the connections between value functions and functional decompositions raise the question whether the different kinds of axiomatizations that have been offered for Shapley Values and their variants in the literature on economic game theory have natural parallels in criteria for functional decompositions and vice-versa. Two concurrent works have already taken important steps in this direction: Hiabu et al. (2022) show that the value function of interventional SHAP can be motivated with a causal assumption on the associated functional decomposition, and Herren and Hahn (2022) outline connections between observational SHAP and functional ANOVA (Hooker, 2007).

We hope that the proposed n-Shapley Values might be a useful tool for further studies of the properties of explanations. In particular, it seems natural to ask whether other feature attribution methods have similar extensions. However, we should also note that n-Shapley Values come with obvious computational limitations. Similar to the original Shapley Values, it might be possible to come up with efficient implementations for certain model classes (Lundberg et al., 2020). In general, however, n-Shapley Values will be exhaustive to compute simply due to the exponentially increasing number of terms involved in the explanations. Also note that while the value function and the Shapley-GAM are uniquely determined, there are probably alternatives to n-Shapley Values. Indeed, we believe that this is evident from the fact that they extend Shapley Interaction Indices (Lundberg et al., 2020), which are not the only possible form of interaction index (Fujimoto et al., 2006; Sundararajan et al., 2020; Tsai et al., 2022; Hamilton et al., 2022).

Finally, we note that our results have a particular practical implication for the computation of Shapley Values. At a high level, our results suggest that Shapley Values should be computed from an output of the function that is as well-approximated by a generalized additive model as possible. In practice, this means that Shapley Values should be computed before applying any final form of non-linearity, and before rounding probabilities to binary predictions.

Acknowledgements

This work was done in part while Sebastian was visiting the Simons Institute for the Theory of Computing. Sebastian would like to thank Rich Caruana, Gyorgy Turan, Michal Moshkovitz and Tosca Lechner for many fruitful discussions about variable interactions. The authors would also like to thank Markus Scheuer and René Gy for linking Lemma 9 to the literature on Bernoulli numbers. This work has been supported by the German Research Foundation through the Cluster of Excellence “Machine Learning – New Perspectives for Science" (EXC 2064/1 number 390727645), the BMBF Tübingen AI Center (FKZ: 01IS18039A), and the International Max Planck Research School for Intelligent Systems (IMPRS-IS).

References

R. Agarwal, L. Melnick, N. Frosst, X. Zhang, B. Lengerich, R. Caruana, and G. E. Hinton. Neural additive models: Interpretable machine learning with neural nets. *NeurIPS*, 2021.

E. Albini, J. Long, D. Dervovic, and D. Magazzeni. Counterfactual shapley additive explanations. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022.

S. Bordt, M. Finck, E. Raidl, and U. von Luxburg. Post-hoc explanations fail to achieve their purpose in adversarial contexts. *arXiv preprint arXiv:2201.10295*, 2022.

R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015.

H. Chen, J. D. Janizek, S. Lundberg, and S.-I. Lee. True to the model or true to the data? *arXiv preprint arXiv:2006.16234*, 2020.

I. Covert, S. Lundberg, and S. Lee. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research (JMLR)*, 22(209):1–90, 2021.

F. Ding, M. Hardt, J. Miller, and L. Schmidt. Retiring adult: New datasets for fair machine learning. In *Neural Information Processing Systems (NeurIPS)*, 2021.

F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

K. Fujimoto, I. Kojadinovic, and J.-L. Marichal. Axiomatic characterizations of probabilistic and cardinal-probabilistic interaction indices. *Games and Economic Behavior*, 55(1):72–99, 2006.

D. Garreau and U. von Luxburg. Explaining the explainer: A first theoretical analysis of lime. In *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

H. Gould and J. Quaintance. Bernoulli numbers and a new binomial transform identity. *J. Integer Seq.*, 2014.

M. Grabisch. K-order additive discrete fuzzy measures and their representation. *Fuzzy sets and systems*, 92(2):167–189, 1997.

M. Grabisch. Bases and transforms of set functions. In *On Logical, Algebraic, and Probabilistic Aspects of Fuzzy Set Theory*. Springer, 2016.

R. Gy. Combinatorial identity involving bernoulli numbers. Mathematics Stack Exchange, 2022. URL `https://math.stackexchange.com/q/4520567`.

M. Hamilton, S. Lundberg, L. Zhang, S. Fu, and W. T. Freeman. Axiomatic explanations for visual search, retrieval, and similarity learning. In *ICLR*, 2022.

T. Han, S. Srinivas, and H. Lakkaraju. Which explanation should i choose? a function approximation perspective to characterizing post hoc explanations. *arXiv preprint arXiv:2206.01254*, 2022.

T. Hastie and R. Tibshirani. *Generalized Additive Models. Chapman Hall & CRC*. 1990.

A. Herren and P. R. Hahn. Statistical aspects of SHAP: Functional ANOVA for model interpretation. *arXiv preprint arXiv:2208.09970*, 2022.

T. Heskes, E. Sijben, I. G. Bucur, and T. Claassen. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *NeurIPS*, 2020.

M. Hiabu, J. T. Meyer, and M. N. Wright. Unifying local and global model explanations by functional decomposition of low dimensional structures, 2022.

A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, and W. Samek. xxai-beyond explainable artificial intelligence. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, pages 3–10. Springer, 2022.

G. Hooker. Generalized functional anova diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 2007.

D. Janzing, L. Minorics, and P. Blöbaum. Feature relevance quantification in explainable ai: A causal problem. In *International Conference on artificial intelligence and statistics*, pages 2907–2916. PMLR, 2020.

N. Jethani, M. Sudarshan, I. C. Covert, S.-I. Lee, and R. Ranganath. Fastshap: Real-time shapley value estimation. In *ICLR*, 2021.

R. Kommiya Mothilal, D. Mahajan, C. Tan, and A. Sharma. Towards unifying feature attribution and counterfactual explanations: Different means to the same end. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2021.

S. Krishna, T. Han, A. Gu, J. Pombra, S. Jabbari, S. Wu, and H. Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*, 2022.

I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler. Problems with shapley-value-based explanations as feature importance measures. In *ICML*, 2020.

E. Lee, D. Braines, M. Stiffler, A. Hudler, and D. Harborne. Developing the sensitivity of lime for better machine learning explanation. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, 2019.

B. Lengerich, S. Tan, C.-H. Chang, G. Hooker, and R. Caruana. Purifying interaction effects with the functional anova: An efficient algorithm for recovering identifiable additive models. In *AISTATS*, 2020.

B. J. Lengerich, R. Caruana, M. E. Nunnally, and M. Kellis. Death by round numbers and sharp thresholds: How to avoid dangerous ai ehr recommendations. *medRxiv*, 2022.

Z. C. Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 2018.

Y. Lou, R. Caruana, and J. Gehrke. Intelligible models for classification and regression. In *18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012.

Y. Lou, R. Caruana, J. Gehrke, and G. Hooker. Accurate intelligible models with pairwise interactions. In *ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *NeurIPS*, 2017.

S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2020.

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

M. Moshkovitz, S. Dasgupta, C. Rashtchian, and N. Frost. Explainable k-means and k-medians clustering. In *ICML*, 2020.

H. Nori, S. Jenkins, P. Koch, and R. Caruana. Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*, 2019.

M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you? explaining the predictions of any classifier. In *22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016.

C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 2022.

W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3): 247–278, 2021.

L. Shapley. A value for n-person games., 1953.

D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2020.

D. Slack, A. Hilgard, S. Singh, and H. Lakkaraju. Reliable post hoc explanations: Modeling uncertainty in explainability. In *Neural Information Processing Systems (NeurIPS)*, 2021.

M. Sundararajan and A. Najmi. The many shapley values for model explanation. In *ICML*, 2020.

M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *ICML*. PMLR, 2017.

M. Sundararajan, K. Dhamdhere, and A. Agarwal. The shapley taylor interaction index. In *ICML*, 2020.

C.-P. Tsai, C.-K. Yeh, and P. Ravikumar. Faith-shap: The faithful shapley interaction index. *arXiv preprint arXiv:2203.00870*, 2022.

Z. J. Wang, A. Kale, H. Nori, P. Stella, M. E. Nunnally, D. H. Chau, M. Vorvoreanu, J. Wortman Vaughan, and R. Caruana. Interpretability, then what? editing machine learning models to reflect human knowledge and values. In *28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022.

A n-Shapley Values

This section details the properties of n-Shapley Values.

A.1 Bernoulli numbers

The Bernoulli numbers[1] B_n are defined by $B_0 = 1$ and

$$\sum_{k=0}^{n} \binom{n+1}{k} B_k = 0 \quad \forall n \geq 1. \tag{11}$$

In this paper, the Bernoulli numbers arise as the coefficients that make n-Shapley Values sum to the prediction (Proposition 11). In fact, equation (11) arises directly from the proof of Proposition 11. The Bernoulli numbers can be computed recursively by re-writing into (11)

$$B_n = \frac{-1}{n+1} \sum_{k=0}^{n-1} B_k \binom{n+1}{k} \quad \forall n \geq 1. \tag{12}$$

In a certain sense, the entire combinatorics around n-Shapley Values relies on the properties of the Bernoulli numbers. In particular, the proofs of Theorem 4 and Theorem 5 rely on the following two Lemmas.

Lemma 8. *For all $n \geq 1$, it holds that*

$$\sum_{k=1}^{n} \frac{B_k}{n-k+1} \binom{n}{k} = \frac{-1}{n+1}. \tag{13}$$

Proof. We re-arrange the sum to get

$$\sum_{k=1}^{n} \frac{B_k}{n-k+1} \binom{n}{k} = \frac{1}{n+1} \sum_{k=0}^{n} \binom{n+1}{k} B_k - \frac{B_0}{n+1} = \frac{-1}{n+1} \tag{14}$$

where the second equality follows from (11). □

Lemma 9. *For all $n, m \geq 0$, it holds that*

$$\sum_{k=0}^{n} \sum_{l=0}^{m} \binom{n}{k} \binom{m}{l} \frac{(n-k)!(m-l)!}{(n+m-k-l+1)!} (-1)^l B_{k+l} = \begin{cases} 1 & \text{if } n = 0 \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

Lemma 9 follows from standard results for the Bernoulli numbers (Gould and Quaintance, 2014)[Theorem 2]. A proof is contained in Appendix F.

A.2 Additivity and Efficiency

From the recursive definition of the n-Shapley Values in Definition 3, a straightforward calculation shows that

$$\Phi_S^n(x) = \sum_{k=0}^{n-|S|} \sum_{K \subset [d] \setminus S, |K|=k} B_k \, \Delta_{S \cup K}(x) \tag{16}$$

which is an alternative non-recursive definition of n-Shapley Values.

[1] An introduction and discussion about Bernoulli numbers can be found, for example, in the corresponding Wikipedia article at `https://en.wikipedia.org/wiki/Bernoulli_number`.

n	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
B_n	1	$\frac{-1}{2}$	$\frac{1}{6}$	0	$\frac{-1}{30}$	0	$\frac{1}{42}$	0	$\frac{-1}{30}$	0	$\frac{5}{66}$	0	$\frac{-691}{2730}$	0	$\frac{7}{6}$	0	$\frac{-3617}{510}$	0	$\frac{43867}{798}$	0

Table A.1: The first 20 Bernoulli numbers.

Proposition 10 (Additivity)**.** *For all $1 \leq n \leq d$ and all $f, g : \mathbb{R}^n \to \mathbb{R}$, we have*

$$\Phi_S^n(x; f+g) = \Phi_S^n(x; f) + \Phi_S^n(x; g). \tag{17}$$

Proof. By definition, Φ_S^n is linear in Δ_S, and Δ_S is linear in the value function v. Therefore, the linearity of Φ_S^n in f follows from the linearity of v in f, i.e. from the fact that $v_{f+g}(x, S) = v_f(x, S) + v_g(x, S)$. □

Proposition 11 (Efficiency)**.** *For all $1 \leq n \leq d$, it holds that*

$$\sum_{\substack{S \subset [d] \\ 1 \leq |S| \leq n}} \Phi_S^n(x) = v([d]) - v(\emptyset). \tag{18}$$

Proof. For $n = 1$, the statement follows from the efficiency of the original Shapley Values. We assume that the statement holds for $n-1$ and re-arrange the sum

$$\begin{aligned}
\sum_{\substack{S \subset [d] \\ 1 \leq |S| \leq n}} \Phi_S^n(x) &= \sum_{\substack{S \subset [d] \\ 1 \leq |S| < n}} \Phi_S^n(x) + \sum_{\substack{S \subset [d] \\ |S| = n}} \Phi_S^n(x) \\
&= \sum_{\substack{S \subset [d] \\ 1 \leq |S| < n}} \left(\Phi_S^{n-1}(x) + B_{n-|S|} \sum_{\substack{K \subset [d] \setminus S \\ |K|+|S|=n}} \Delta_{S \cup K}(x) \right) + \sum_{\substack{S \subset [d] \\ |S| = n}} \Delta_S(x) \\
&= \sum_{\substack{S \subset [d] \\ 1 \leq |S| \leq n-1}} \Phi_S^{n-1}(x) + \sum_{\substack{S \subset [d] \\ 1 \leq |S| < n}} \sum_{\substack{K \subset [d] \setminus S \\ |K|+|S|=n}} B_{n-|S|}\, \Delta_{S \cup K}(x) + \sum_{\substack{S \subset [d] \\ |S| = n}} \Delta_S(x).
\end{aligned} \tag{19}$$

Notice that the first term is equivalent to $v([d]) - v(\emptyset)$ by the induction hypothesis. It remains to show that

$$\sum_{\substack{S \subset [d] \\ 1 \leq |S| < n}} \sum_{\substack{K \subset [d] \setminus S \\ |K|+|S|=n}} B_{n-|S|}\, \Delta_{S \cup K}(x) + \sum_{\substack{S \subset [d] \\ |S| = n}} \Delta_S(x) = 0. \tag{20}$$

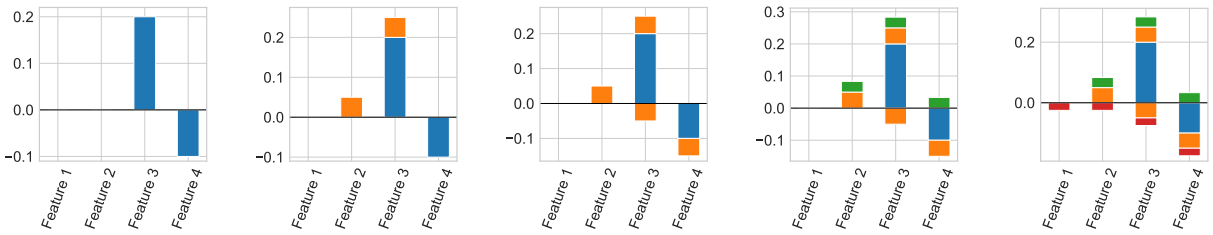
Notice that both sums are over sets of length n. In the first sum, each sets occurs multiple times. In the second sum, each set occurs exactly once. By counting the occurrences of each set in the first sum we see that (20) holds if

$$\sum_{s=1}^{n-1} B_{n-s} \binom{n}{s} + 1 = 0. \tag{21}$$

If we set $B_0 = 1$, this holds if and only if

$$\sum_{k=0}^{n-1} B_k \binom{n}{k} = 0, \tag{22}$$

which is the defining property of the Bernoulli numbers (11). In summary, we see that the Bernoulli numbers are the coefficients that balance the terms in the first sum in equation (20). □



(a) Example 1 (b) Example 2 (c) Example 3 (d) Example 4 (e) Example 5

Figure A.1: Examples to illustrate the proposed visualization technique for n-Shapley Values.

A.3 Relationship Between n-Shapley Values of Different Order

The following proposition is a straightforward extension of Theorem 5.

Proposition 12 (Relationship Between n-Shapley Values of Different Order)**.** *For $m \leq n$, let Φ_S^m and Φ_S^n be the m- and n-Shapley Values, respectively. Then, the m-Shapley Values can be computed from the n-Shapley Values by*

$$\Phi_S^m(x) = \Phi_S^n + \sum_{\substack{K \subset [d] \setminus S, \\ m-|S| < |K| \leq n-|S|}} \beta_{m-|S|,|K|} \, \Phi_{S \cup K}^n(x). \tag{23}$$

Specifically, it holds that

$$\Phi_i^1 = \Phi_i^n + \frac{1}{2} \sum_{j \neq i} \Phi_{i,j}^n + \cdots + \frac{1}{n} \sum_{\substack{K \subset [d] \setminus \{i\} \\ |K| = n-1}} \Phi_{K \cup i}^n \tag{24}$$

which is the basis for the visualizations in the paper.

Proof. The proposition follows from the counting argument used in the proof of Theorem 5. □

A.4 Visualizing n-Shapley Values

Due to the large number of terms involved in n-Shapley Values of higher order, visualizing these explanations is difficult.[2] However, Proposition 12 (which is really a variant of Theorem 5) states that higher-order variable interactions in n-Shapley Values are related to the original Shapley Values via a simple lump-sum formula. This gives rise to the idea of simply visualizing, for each feature, the respective components of the sum.

To illustrate this idea, let us consider a simple example. Let us begin with four different features and the usual Shapley Values. Say the first two features have attribution zero, the third feature has attribution 0.2, and the fourth feature has attribution -0.1. These Shapley Values can be visualized as usual, depicted in Figure A.1a. Now, let us add a second-order interaction effect, say $\Phi_{2,3}^2 = 0.1$. Because this interaction effect would ultimately be added to the attributions of feature 2 and feature 3 with a factor of $\frac{1}{2}$, let us simply add two corresponding bars to the attributions of these features, with the color indicating that it is a second-order effect. From the resulting Figure A.1b, it can then be seen that we have two main effects and a single positive interaction effect between features 2

[2] Empirically, the Shapley-GAMs investigated in this paper turn out not to be sparse.

and 3. If there were another interaction effect, say $\Phi^2_{3,4} = -0.1$, we would proceed in the same way, taking care of the sign. From the resulting Figure A.1c, it can be seen that there are two main effects and a number of second-order interactions. With higher-order interactions we proceed accordingly, as illustrated for $\Phi^3_{2,3,4} = 0.1$ (Figure A.1d) and $\Phi^4_{1,2,3,4} = -0.1$ (Figure A.1d).

Note that while this form of visualization faithfully depicts the relative magnitude of the different variable interactions, it is in general not possible to tell from the figures which variables interact with each other, for example when there are a number of different second-order effects.

B Proof of Theorem 4

Proof of Theorem 4. We are going to show that

$$\Phi^d_S(x) = \sum_{L \subset S} (-1)^{|S|-|L|} v(x_L, L). \tag{25}$$

Note that the RHS evaluates the value function v only for sets $L \subset S$. From the assumption that the value function is subset-compliant, it follows that the RHS is a well-defined function of x_S. According to Proposition 11 (efficiency), the d-Shapley Values sum to $v(x) - v(\emptyset)$ which implies the Theorem.

To show (25), we consider the non-recursive definition of n-Shapley Values 16 and then substitute the definition of $\Delta_S(x)$ from Definition 3.

$$\begin{aligned}
\Phi^d_S(x) &= \sum_{k=0}^{d-|S|} \sum_{K \subset [d]\setminus S,\, |K|=k} B_k \, \Delta_{S\cup K}(x) \\
&= \sum_{k=0}^{d-|S|} \sum_{K \subset [d]\setminus S,\, |K|=k} B_k \sum_{T \subset [d]\setminus (S\cup K)} \frac{(d-|T|-|S|-|K|)!|T|!}{(d-|S|-|K|+1)!} \sum_{L \subset S\cup K} (-1)^{|S|+|K|-|L|} v(x,\, L\cup T). \\
&= \sum_{K \subset [d]\setminus S} \sum_{T \subset [d]\setminus (S\cup K)} B_{|K|} \, \frac{(d-|T|-|S|-|K|)!|T|!}{(d-|S|-|K|+1)!} \sum_{L \subset S\cup K} (-1)^{|S|+|K|-|L|} v(x,\, L\cup T).
\end{aligned} \tag{26}$$

Where the last equation follows from the realization that we are summing over all possible subsets of $[d] \setminus S$.

In equation (26), we are summing over the value of the same sets multiple times. Let us fix a set $M = L \cup T$ and count how often it occurs in the sum. First note that $v(x, M)$ occurs exactly once for every set K, namely by choosing $T = M \setminus (S \cup K)$ and $L = M \cap (S \cup K)$. Since the coefficients do not only depend on the size of K, but also on $|T|$ and $|L|$, let us partition the set $K = K_1 \cup K_2 = \{K \cap M\} \cup \{K \setminus M\}$. Let $n_1 = |M \setminus S|$ and $n_2 = |[d] \setminus (S \cup M)|$ denote the maximum sizes of both partitions. With this counting argument, we arrive at

$$(-1)^{|S|-|M|} \sum_{K_1 \subset M\setminus S} \sum_{K_2 \subset [d]\setminus (S\cup M)} B_{|K_1|+|K_2|} \, \frac{(n_2-|K_2|)!(n_1-|K_1|)!}{(n_1+n_2-|K_1|-|K_2|+1)!} (-1)^{|K_2|} \tag{27}$$

occurrences of the term $v(x, M)$. Notice that equation (27) is equal to

$$(-1)^{|S|-|M|} \sum_{k_1=0}^{n_1} \sum_{k_2=0}^{n_2} \binom{n_1}{k_1} \binom{n_2}{k_2} \frac{(n_2-k_2)!(n_1-k1)!}{(n_1+n_2-k_1-k_2+1)!} (-1)^{k_2} B_{k_1+k_2} \tag{28}$$

The desired result now follows from the properties of the Bernoulli numbers. In particular, since $M \subset S \iff n_1 = 0$, we see from Lemma 9 that (28) equals $(-1)^{|S|-|M|}$ if $M \subset S$ and 0 otherwise. Comparing the terms for all possible sets $M \subset [d]$, we see that (26) equals (25). □

C Proof of Theorem 5

Proof of Theorem 5. According to Theorem 4, the d-Shapley Values can be written as

$$\Phi_S^d(x) = f_S(x) \tag{29}$$

where $f_S(x)$ are the component functions of the Shapley-GAM. Hence, the d-Shapley Values are a linear combination of the component functions of the Shapley-GAM. From the recursive definition of the n-Shapley Values, we see that

$$\Phi_S^n(x) = \Phi_S^{n+1}(x) - B_{1+n-|S|} \sum_{K \subset [d] \setminus S, |K|+|S|=n+1} \Phi_{S \cup K}^{n+1}(x) \tag{30}$$

that is the n-Shapley Values are a linear combination of the terms involved in the $n+1$-Shapley Values. By induction, we see that the n-Shapley Values are linear combinations of the component functions of the Shapley-GAM.

It remains to determine the coefficients $C_{n,m}$. We present a counting argument that is based on the recurrence relation (30). In this counting argument, we first determine the coefficients $D_{n,m}$ where the first index corresponds to the distance between $|S|$ and the order of the Shapley Values, and the second index corresponds to the different between the size of the interaction effect and the order of the Shapley Values. Suppose that we are computing n-Shapley Values. If we use equation (30) to proceed recursively from d-Shapley Values to n-Shapley Values, then the first time that the component function $f_{S \cup K}$ is being added to Φ_S^m is during the computation of the $(|S| + |K| - 1)$-Shapley Values. According to equation (30), the linear coefficient will simply be $D_{|K|-1,1} = -B_{|K|}$. The second time that the component function $f_{S \cup K}$ is being added to Φ_S^m is during the computation of the $(|S| + |K| - 2)$-Shapley Values. This is because we have previously added $-B_1 f_{S \cup K}$ to all the terms of order $|S| + |K| - 1$ that are a subset of $S \cup K$. There are $\binom{|K|}{1}$ such terms, and we are now adding all of them to f_S, using the coefficient $-B_{|K|-1}$. This means that we arrive at a total coefficient of

$$D_{|K|-2,2} = -B_{|K|} + B_{|K|-1}\binom{|K|}{1}B_1. \tag{31}$$

By a similar argument we arrive at a coefficient of

$$D_{|K|-3,3} = -B_{|K|} + B_{|K|-1}\binom{|K|}{1}B_1 - B_{|K|-2}\binom{|K|}{2}B_2 - B_{|K|-2}\binom{|K|}{2}B_1\binom{2}{1}B_1. \tag{32}$$

for the $(|S| + |K| - 3)$-Shapley Values. In general, that is when we compute n-Shapley Values, the component function $f_{S \cup K}$ is being added to Φ_S^n once for every possible pathway that goes from a set of order $n+1$ to the set $S \cup K$ by successively adding different numbers of elements. For $k \geq 1$, let

$$P_k = \left\{ (p_1, \ldots, p_k) \in \mathbb{N}_{\geq 0}^k \;\middle|\; \sum_{i=1}^{k} p_i = k \quad \text{and} \quad p_i = 0 \implies (p_j = 0 \, \forall j > i) \right\} \tag{33}$$

be the set of pathways of length k. This means that we have $P_1 = \big\{(1)\big\}$,

$$\begin{aligned} P_2 &= \big\{(2,0),(1,1)\big\}, \\ P_3 &= \big\{(3,0,0),(2,1,0),(1,2,0),(1,1,1)\big\}, \\ P_4 &= \big\{(4,0,0,0),(3,1,0,0),(2,2,0,0),(2,1,1,0), \\ &\qquad (1,3,0,0),(1,2,1,0),(1,1,2,0),(1,1,1,1)\big\} \end{aligned} \tag{34}$$

and so on. By accounting for the coefficients B_k and the signs along each path, the coefficients can be written as

$$D_{n,m} = \sum_{(p_1,\ldots,p_m)\in P_m} (-1)^{\sum_{i=1}^m \operatorname{sign}(p_i)} \binom{n+m}{n+p_1} B_{n+p_1} \prod_{i=2}^{m} B_{p_i} \binom{m-\sum_{j=1}^{i-1} p_j}{p_i} \tag{35}$$

From this, we derive the special case

$$\begin{aligned}
D_{0,m} &= \sum_{(p_1,\ldots,p_m)\in P_m} (-1)^{\sum_{i=1}^m \operatorname{sign}(p_i)} \binom{m}{i_1} B_{p_1} \prod_{i=2}^{m} B_{p_i} \binom{m-\sum_{j=1}^{i-1} p_j}{p_i} \\
&= \sum_{(p_1,\ldots,p_m)\in P_m} (-1)^{\sum_{i=1}^m \operatorname{sign}(p_i)} \prod_{i=1}^{m} B_{p_i} \binom{m-\sum_{j=1}^{i-1} p_j}{p_i} \\
&= -B_m - \sum_{p_1=1}^{m-1} a_{p_1} \binom{m}{p_1} \sum_{(\hat{p}_1,\ldots,\hat{p}_{m-p_1})\in P_{m-p_1}} (-1)^{\sum_{i=1}^{m-p_1} \operatorname{sign}(p_i)} \prod_{j=1}^{m-p_1} B_{\hat{p}_j} \binom{m-i_1-\sum_{s=1}^{j-1} \hat{p}_s}{\hat{p}_j} \\
&= -B_m - \sum_{p_1=1}^{m-1} a_{p_1} \binom{m}{p_1} \beta_{0,m-p_1} \\
&= -B_m - \sum_{p_1=1}^{m-1} a_{p_1} \binom{m}{p_1} \frac{1}{m-p_1+1} \\
&= -\sum_{k=1}^{m} \frac{B_k}{m-k+1} \binom{m}{k} \\
&= \frac{1}{m+1}
\end{aligned} \tag{36}$$

where the last equality is due to Lemma 8. Now, this implies that

$$\Delta_S(x) = \Phi_S^{|S|}(x) = f_S(x) + \sum_{K\subset[d]\setminus S,\ |K|\geq 1} D_{0,|K|}\, f_{S\cup K}(x) = \sum_{K\subset[d]\setminus S} \frac{1}{1+|K|} f_{S\cup K}(x) \tag{37}$$

which is a version of Theorem 1 in Grabisch (1997). Using (37) and the explicit formula for n-Shapley Values (16), we get

$$\begin{aligned}
\Phi_S^n(x) &= \sum_{k=0}^{n-|S|} \sum_{K\subset[d]\setminus S,\,|K|=k} B_k\, \Delta_{S\cup K}(x) \\
&= \sum_{k=0}^{n-|S|} \sum_{K\subset[d]\setminus S,\,|K|=k} B_k \sum_{T\subset[d]\setminus(S\cup K)} \frac{1}{1+|T|} f_{S\cup K\cup T}(x)
\end{aligned} \tag{38}$$

From which we see that the component function $f_{S\cup\tilde{K}}$ is being added to $\Phi_S^n(x)$ exactly

$$C_{n-|S|,|\tilde{K}|} = \sum_{k=0}^{n-|S|} \binom{n-|S|}{k} \frac{B_k}{1+|\tilde{K}|-k} \tag{39}$$

times which concludes the proof. □

D Proof of Theorem 6

Proof of Theorem 6. According to Theorem 4, the Shapley-GAM decomposition is given by

$$f_S(x) = \sum_{L \subset S} (-1)^{|S|-|L|} v(x_L, L). \tag{40}$$

By substituting the definition of the value function (10)

$$\begin{aligned} f_S(x) &= \sum_{L \subset S} (-1)^{|S|-|L|} v(x_L, L) \\ &= \sum_{L \subset S} (-1)^{|S|-|L|} \sum_{T \subset L} g_T(x) \\ &= \sum_{L \subset S} \sum_{T \subset L} g_T(x) (-1)^{|S|-|L|} \\ &= \sum_{T \subset S} g_T(x) \sum_{L \subset S \setminus T} (-1)^{|S|-|L|-|T|} \\ &= g_S(x) \end{aligned} \tag{41}$$

Where we have re-arranged the sum to count the number of occurrences of the set T, and then used the fact that inner sum averages to zero except for $T = S$. □

E Proof of Theorem 7

Proof of Theorem 7. We assume that the function f can be written as a GAM of order n, that is

$$f(x) = \sum_{S \subset [d],\, |S| \leq n} g_S(x_S). \tag{42}$$

Notice that this GAM is not necessarily the Shapley-GAM, but just some way to write the function f as a GAM. Now, let f_S be the component functions of the Shapley-GAM. According to Theorem 5, n-Shapley Values can be written as a linear combination of the component functions of the Shapley-GAM

$$\Phi_S^n(x) = f_S(x_S) + \sum_{K \subset [d] \setminus S,\, |S|+|K|>n} C_{n-|S|,|K|}\, f_{S \cup K}(x_{S \cup K}) \tag{43}$$

with linear coefficients $C_{n,m}$. According to equation (43), $\Phi_S^n(x)$ is equal to $f_S(x_S)$ plus some weighted components of the Shapley-GAM of order greater than n. As a consequence, should the Shapley-GAM also be a GAM of order n, then the second sum vanishes and we arrive at $\Phi_S^n(x) = f_S(x_S)$ which is what we want to show.

It remains to show that the Shapley-GAM is a GAM of order n. According to Theorem 4, the component functions of the Shapley-GAM are given by

$$f_S(x) = \sum_{L \subset S} (-1)^{|S|-|L|} v(x_L, L). \tag{44}$$

We want to show that the component functions of degree greater than n vanish. Let us first consider

observational SHAP. Here we have

$$\begin{aligned}\sum_{L\subset S}(-1)^{|S|-|L|}v(x_L,L) &= \sum_{L\subset S}(-1)^{|S|-|L|}\mathbb{E}[f(x)|x_L]\\ &= \sum_{L\subset S}(-1)^{|S|-|L|}\mathbb{E}\left[\sum_{T\subset[d],|T|\leq n} g_T(x_T)\Big|x_L\right]\\ &= \sum_{L\subset S}(-1)^{|S|-|L|}\sum_{T\subset[d],|T|\leq n}\mathbb{E}\left[g_T(x_T)|x_L\right]\\ &= \sum_{T\subset[d],|T|\leq n}\sum_{L\subset S}(-1)^{|S|-|L|}\mathbb{E}\left[g_T(x_T)|x_L\right]\end{aligned} \tag{45}$$

Consider the inner sum. If $|S|>n$, we can always pick an element $i\in S\setminus T$ and write

$$\sum_{L\subset S\setminus\{i\}}(-1)^{|S|-|L|}\Big(\mathbb{E}\left[g_T(x_T)|x_L\right]-\mathbb{E}\left[g_T(x_T)|x_{L\cup\{i\}}\right]\Big) \tag{46}$$

If the input features are independent, then $g_T(x_T)$ and x_i are independent, from which we get by the properties of the conditional expectation that

$$\mathbb{E}\left[g_T(x_T)|x_{L\cup\{i\}}\right]=\mathbb{E}\left[g_T(x_T)|x_L\right] \tag{47}$$

It follows that the inner sum is zero for all sets T, and that the component functions of the Shapley-GAM of degree greater than n are equal to zero, too.

Let us now consider interventional SHAP. Just as for observational SHAP, we arrive at equation (46) using the linearity of the expectation operator. Hence, we require that

$$\mathbb{E}\left[g_T(x_T)|do(x_{L\cup\{i\}})\right]=\mathbb{E}\left[g_T(x_T)|do(x_L)\right] \tag{48}$$

which follows from the properties of the causal do-operator. Intuitively, since g_T does not depend on the value of feature i, intervening on that feature has no effect. □

F Proof of Lemma 9

Proof according to MSE 4520057. Let us first consider the case $n=0$. For $n=0$ and $m=0$, we have

$$\binom{0}{0}\binom{0}{0}\frac{(0-0)!(0-0)!}{(0+0-0-0+1)!}(-1)^0B_0=1. \tag{49}$$

For $n=0$ and $m\geq 1$, we have

$$\begin{aligned}\sum_{l=0}^{m}\binom{m}{l}\frac{1}{(m-l+1)}(-1)^lB_l &= \frac{1}{m+1}\sum_{l=0}^{m}\binom{m+1}{l}(-1)^lB_l\\ &= \frac{-2}{m+1}\binom{m+1}{1}B_1+\sum_{l=0}^{m}\binom{m+1}{l}\\ &= -2B_1+0=1.\end{aligned} \tag{50}$$

where we used (11) and the fact that the odd Bernoulli numbers vanish except for $n=1$. For $m=0$ and $n\geq 1$, we also have from (11)

$$\sum_{k=0}^{n}\binom{n}{k}\frac{1}{(n-k+1)}(-1)^0B_k=\frac{1}{n+1}\sum_{k=0}^{n}\binom{n+1}{k}B_k=0. \tag{51}$$

It remains to show the general case $n, m \geq 1$. According to a derivation by Gy (2022), the problem in this case is equivalent to

$$(-1)^n \sum_{l=0}^{m} \frac{B_{n+l+1}}{n+l+1} \binom{m}{l} + (-1)^m \sum_{k=0}^{n} \frac{B_{m+k+1}}{m+k+1} \binom{n}{k} = -\frac{1}{(n+m+1)\binom{n+m}{m}} \tag{52}$$

Now, Theorem 2 in Gould and Quaintance (2014) with $s = 1$ states that for any sequence of numbers $(a_n)_{n\geq 0}$, it holds that

$$\sum_{k=0}^{m} \binom{m}{k} \frac{a_{n+k+1}}{n+k+1} = \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} \frac{b_{m+k+1}}{m+k+1} + \frac{(-1)^{n+1} a_0}{(m+n+1)\binom{m+n}{n}} \tag{53}$$

where the sequence $(b_n)_{n\geq 0}$ is the binomial transform of the sequence $(a_n)_{n\geq 0}$, given by

$$b_n = \sum_{k=0}^{n} \binom{n}{k} a_k. \tag{54}$$

Setting $a_n = B_n$, we have from (11) that the binomial transform of the Bernoulli numbers is simply

$$b_n = \sum_{k=0}^{n} \binom{n}{k} B_k = (-1)^n B_n \tag{55}$$

where the factor $(-1)^n$ takes care of the special case $n = 1$. Using (53) with $a_n = B_n$ and $b_n = (-1)^n B_n$, we get

$$(-1)^n \sum_{k=0}^{m} \binom{m}{k} \frac{B_{n+k+1}}{n+k+1} = -\sum_{k=0}^{n} (-1)^m \binom{n}{k} \frac{B_{m+k+1}}{m+k+1} - \frac{1}{(m+n+1)\binom{m+n}{n}} \tag{56}$$

where we multiplied both sides with $(-1)^n$. This is the same as (52) which concludes the proof. □

G Datasets and Models

In our experiments, we use the following datasets and models.

G.1 Datasets

Folktables Income. Folktables is a Python package that provides access to datasets derived from recent US Censuses `https://github.com/zykls/folktables`. We used this package to obtain the data from the 2016 Census in California. The machine learning problem is the ACSIncome prediction task, that is to predict whether an individual's income is above $50,000, based on 10 personal characteristics (Ding et al., 2021). The dataset contains of 152 149 observations.

Folktables Travel Time. Folktables is a Python package that provides access to datasets derived from recent US Censuses `https://github.com/zykls/folktables`. We used this package to obtain the data from the 2016 Census in California. The machine learning problem is the ACSTravelTime prediction task, that is to predict whether an individual has to commute to work longer than 20 minutes, based on 10 personal characteristics (Ding et al., 2021). The dataset contains 133 549 observations.

German Credit. The German Credit Dataset is a dataset with 20 different features on individual's credit history and personal characteristic. The machine learning problem is to predict credit risk in binary form. We obtained the dataset from the UCI machine learning repository and reduced the number of features to 10 without any observed drop in accuracy. The dataset contains 1000 observations.

California Housing. The California Housing dataset was derived from the 1990 U.S. census. The regression problem is to predict the median house value, based on 8 characteristics. We obtained the dataset form the `scikit-learn` library. The dataset contains 20 640 observations.

Iris. The Iris dataset is a simple flower dataset. The machine learning problem is to classify whether the flower is of a particular kind or not, based on 4 different features. We obtained the dataset form the `scikit-learn` library. The dataset contains 150 observations.

G.2 Models

Glassbox-GAM. We train the Glassbox-GAMs with the `interpretML` library (Nori et al., 2019) and default parameters (no interactions).

Gradient Boosted Tree. We use the `xgboost` library and train with 100 trees per model. This setting allows to achieve competitive accuracy for gradient boosted trees.

Random Forest. We use the `scikit-learn` library and train with 100 trees per forest. This setting allows to achieve competitive accuracy for random forests.

k-Nearest Neighbor. We use the `scikit-learn` library. The hyperparameter k was chosen with cross-validation to be 30, 80, 25, 10, 1 for the datasets as listed above.

H Additional Plots and Figures

H.1 Folktables Income

H.1.1 Glassbox-GAM

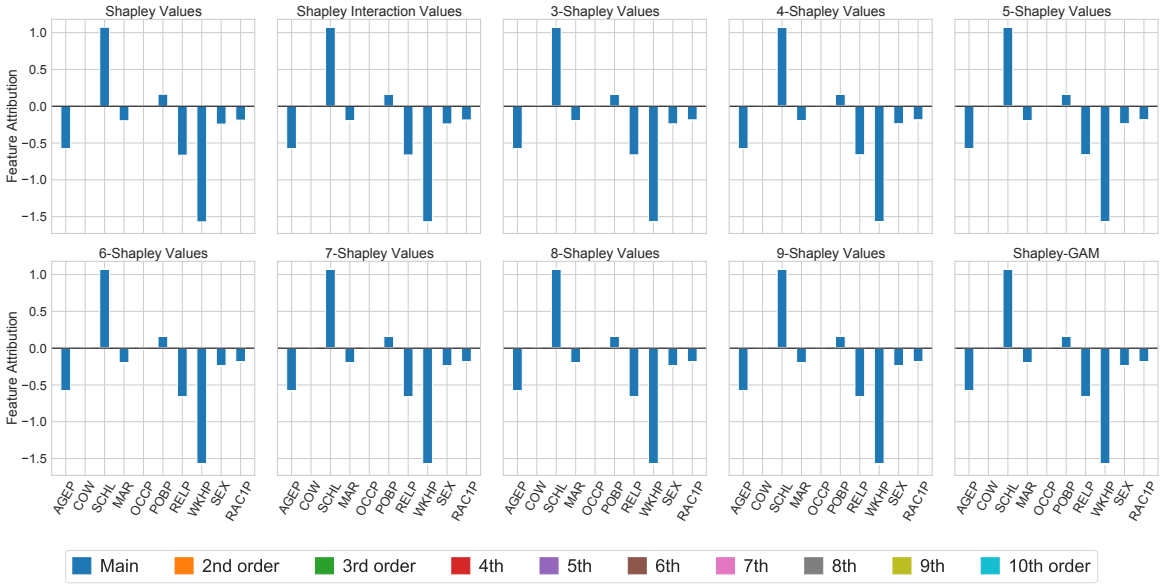


Figure H.2: n-Shapley Values for a Glassbox-GAM and the first observation in our test set of the Folktables income dataset.

H.1.2 Gradient Boosted Tree

Figure H.3: n-Shapley Values for a Gradient Boosted Tree and the first observation in our test set of the Folktables income dataset.

H.1.3 Random Forest

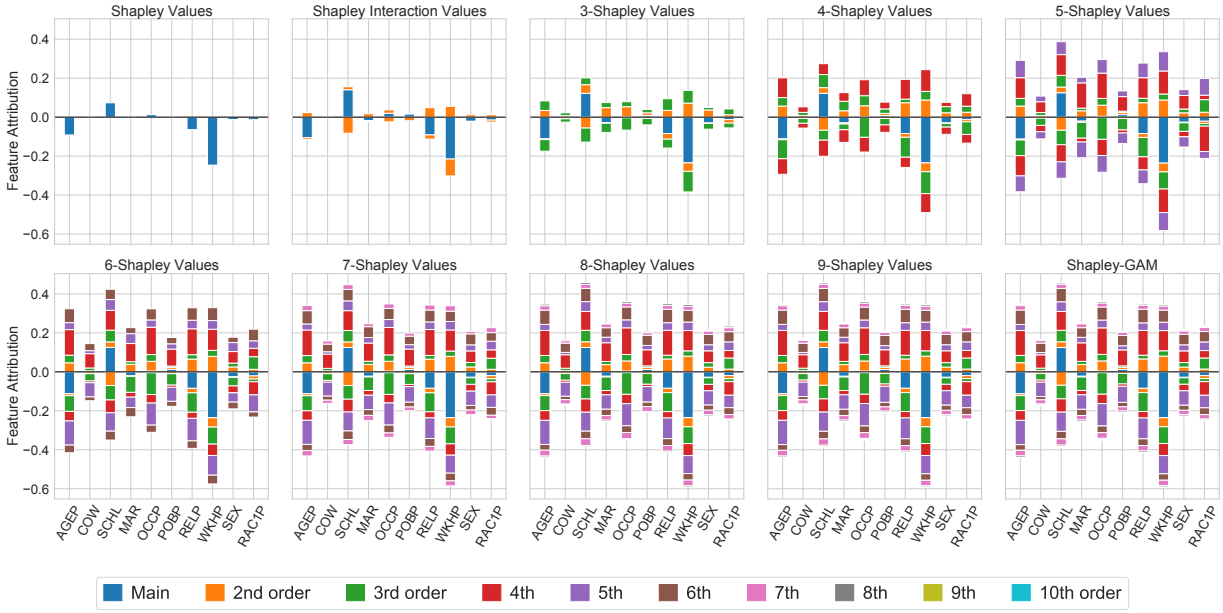


Figure H.4: n-Shapley Values for a Random Forest and the first observation in our test set of the Folktables income dataset.

H.1.4 k-Nearest Neighbor

Figure H.5: n-Shapley Values for a kNN classifier and the first observation in our test set of the Folktables income dataset.

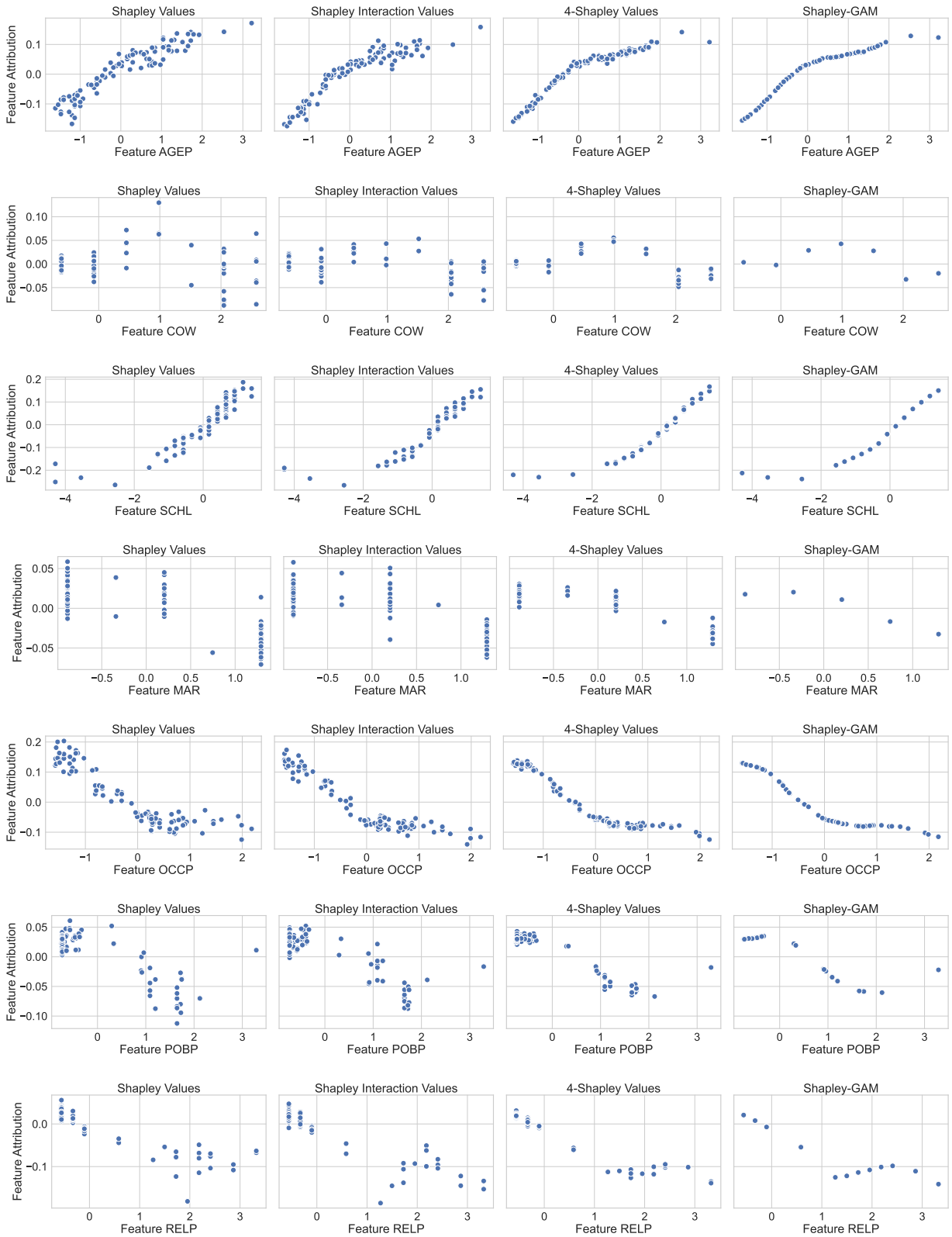


Figure H.6: Partial dependence plots for the kNN classifier on the Folktables Income dataset (compare Figure 2 in the main paper). Depicted are the partial dependence plots of Φ_i^n for $n = \{1, 2, 4, 10\}$ and 7 different features.

H.2 Folktables Travel

H.2.1 Glassbox-GAM

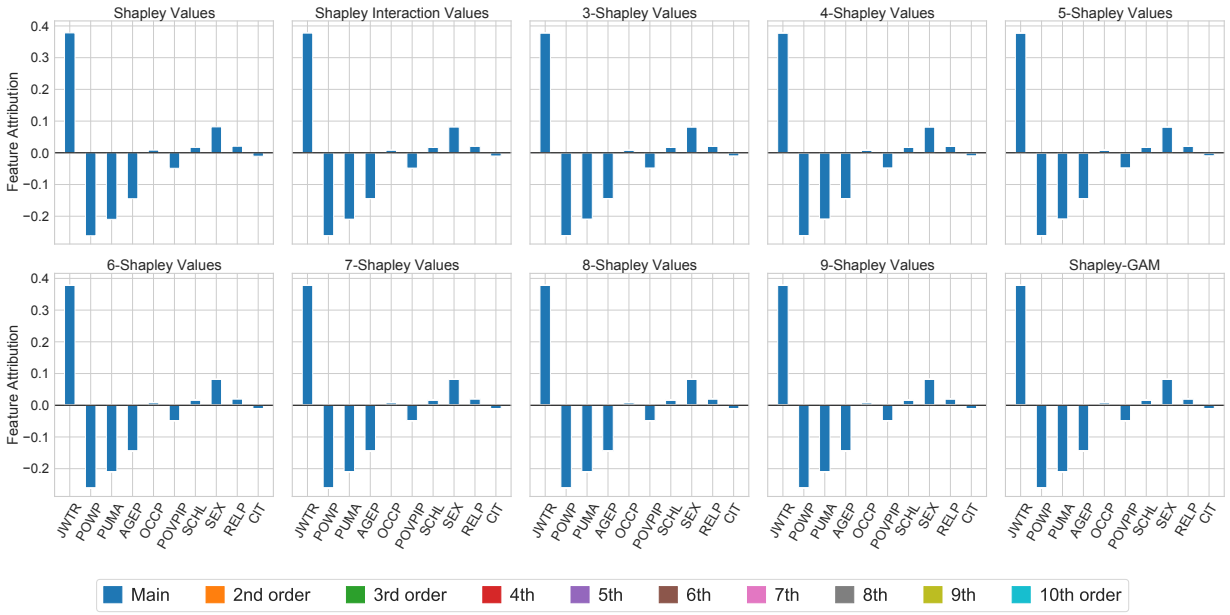


Figure H.7: n-Shapley Values for a Glassbox-GAM and the first observation in our test set of the Folktables Travel dataset.

H.2.2 Gradient Boosted Tree

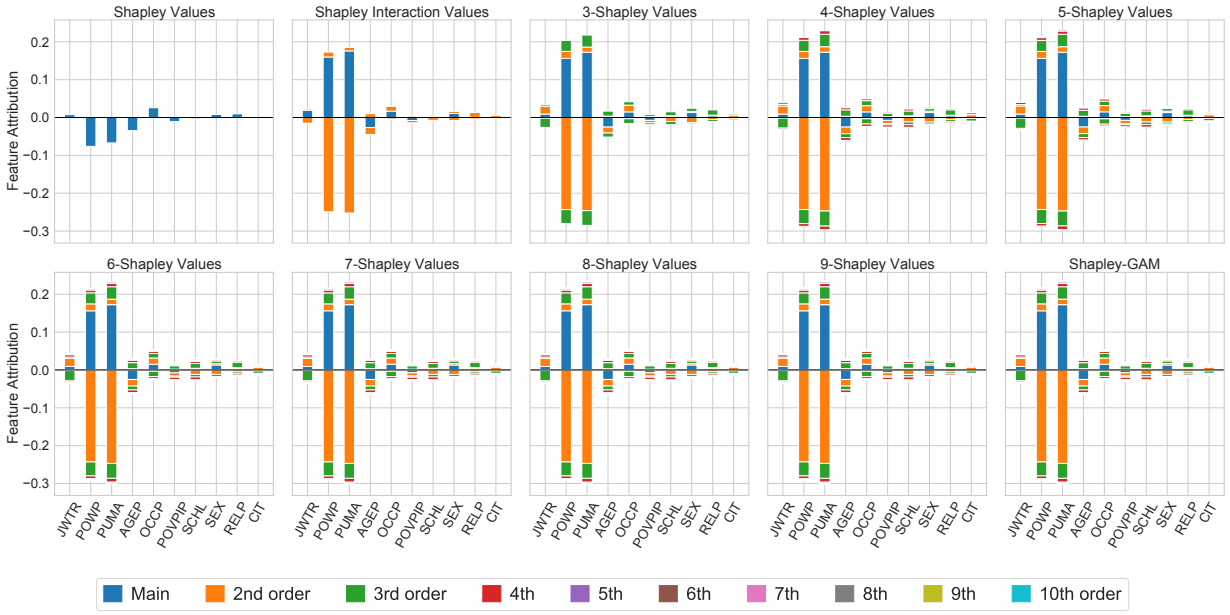


Figure H.8: n-Shapley Values for a Gradient Boosted Tree and the first observation in our test set of the Folktables Travel dataset.

H.2.3 Random Forest

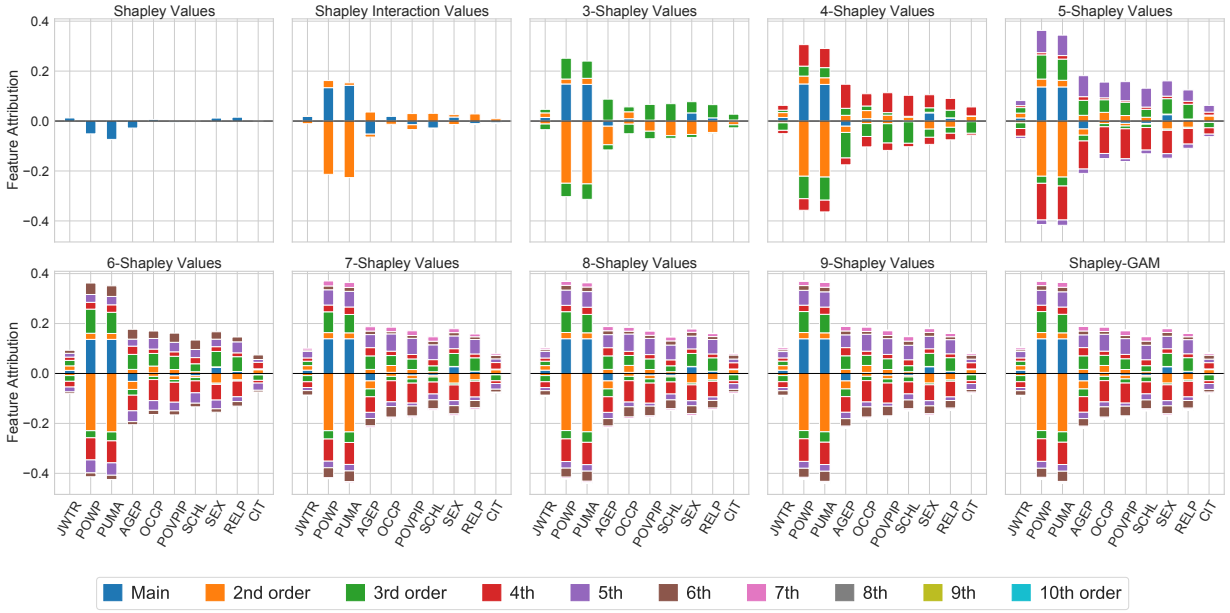


Figure H.9: n-Shapley Values for a Random Forest and the first observation in our test set of the Folktables Travel dataset.

H.2.4 k-Nearest Neighbor

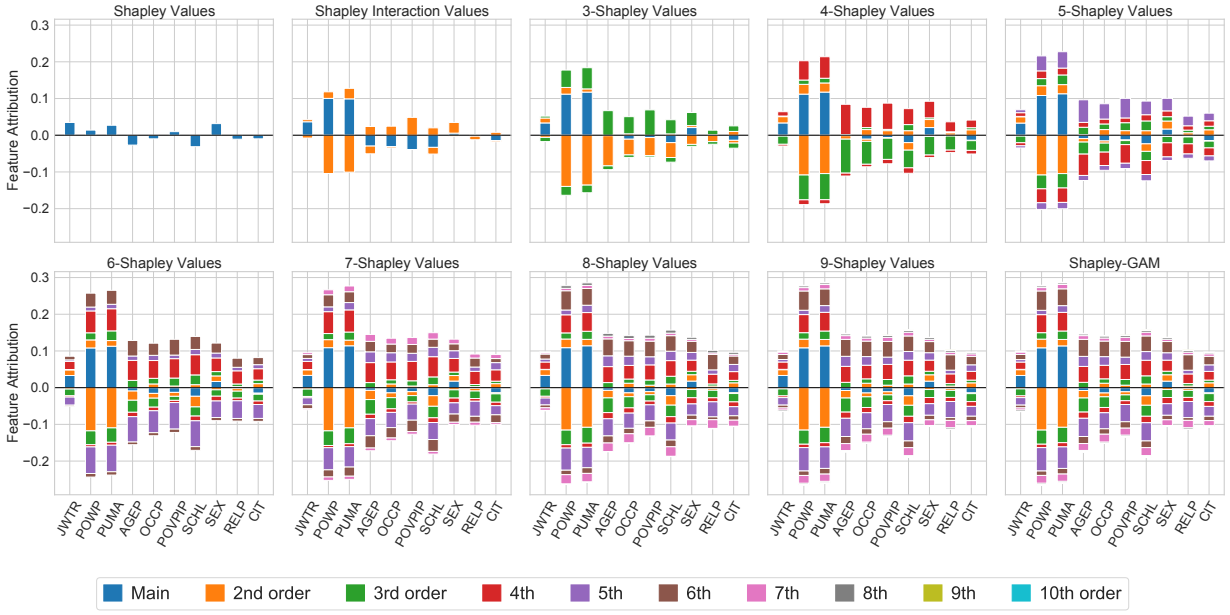


Figure H.10: n-Shapley Values for a kNN classifier and the first observation in our test set of the Folktables Travel dataset.

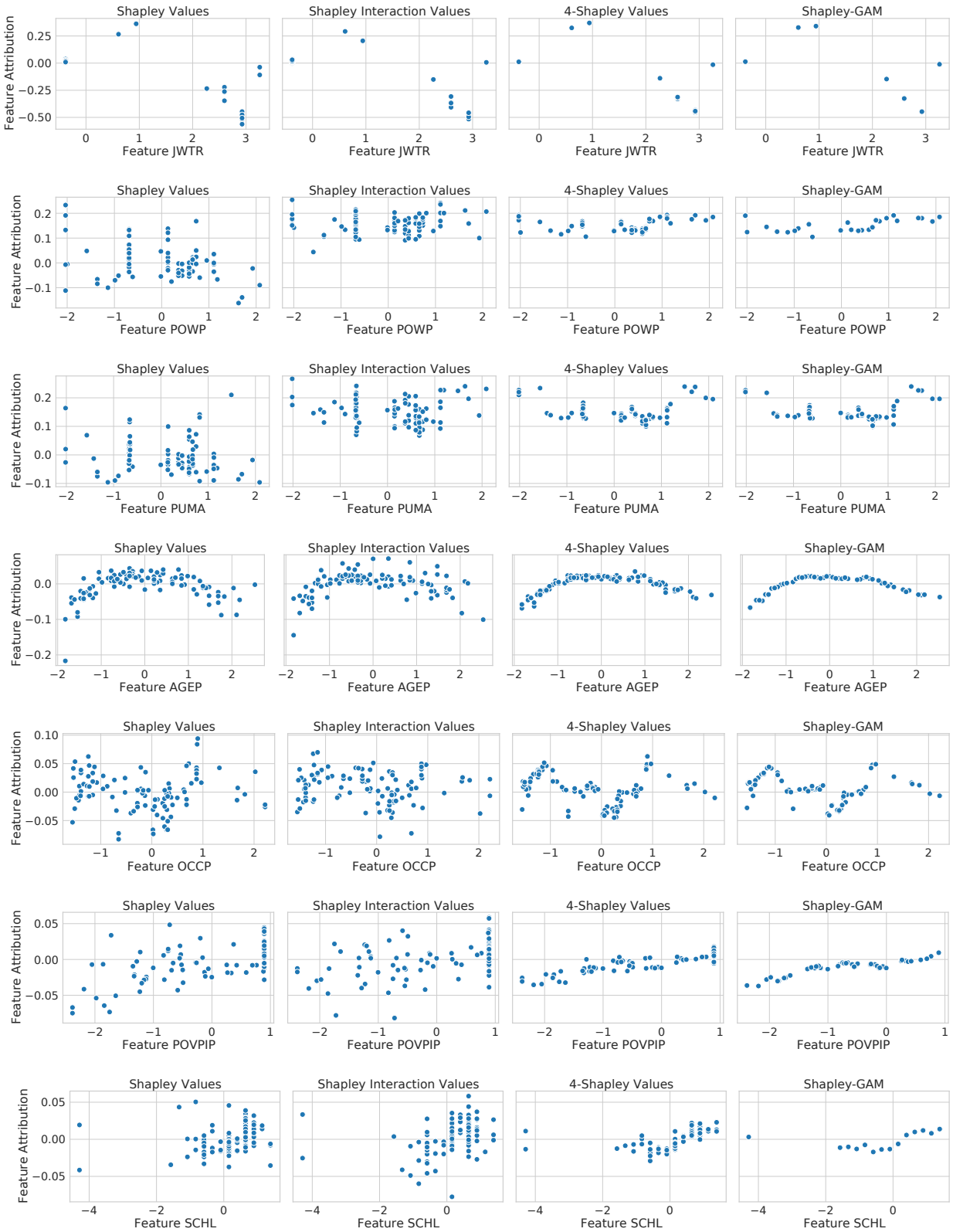


Figure H.11: Partial dependence plots for the random forest on the Folktables Travel dataset. Depicted are the partial dependence plots of Φ_i^n for $n = \{1, 2, 4, 10\}$ and 7 different features.

H.3 German Credit

H.3.1 Glassbox-GAM

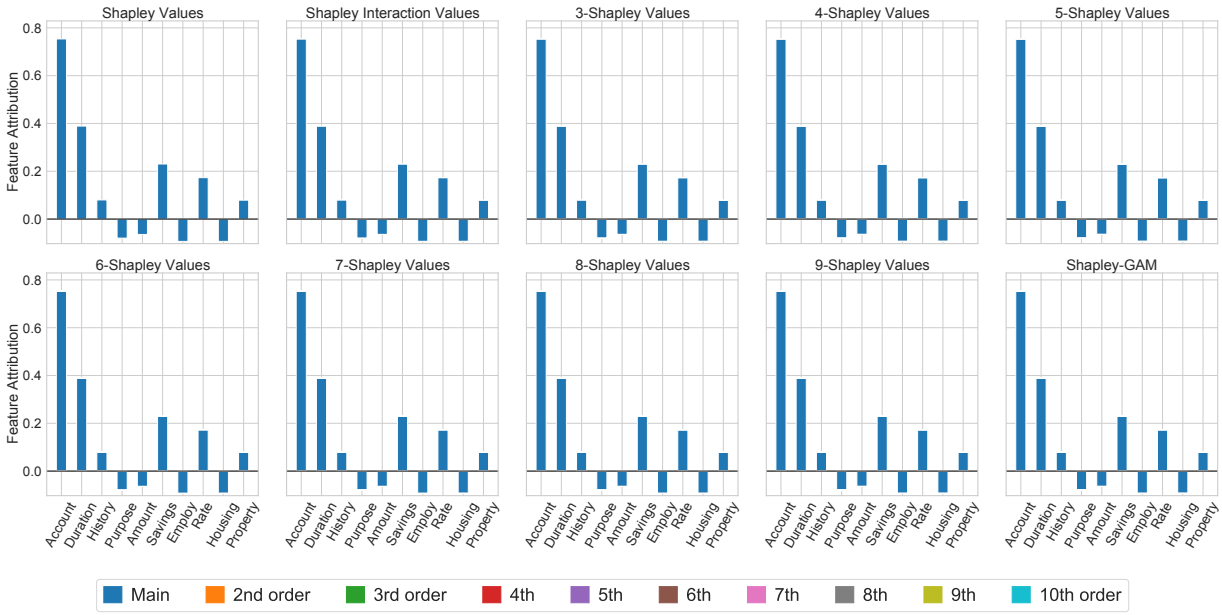


Figure H.12: n-Shapley Values for a Glassbox-GAM and the first observation in our test set of the German Credit dataset.

H.3.2 Gradient Boosted Tree

Figure H.13: n-Shapley Values for a Gradient Boosted Tree and the first observation in our test set of the German Credit dataset.

H.3.3 Random Forest

Figure H.14: n-Shapley Values for a Random Forest and the first observation in our test set of the German Credit dataset.

H.3.4 k-Nearest Neighbor

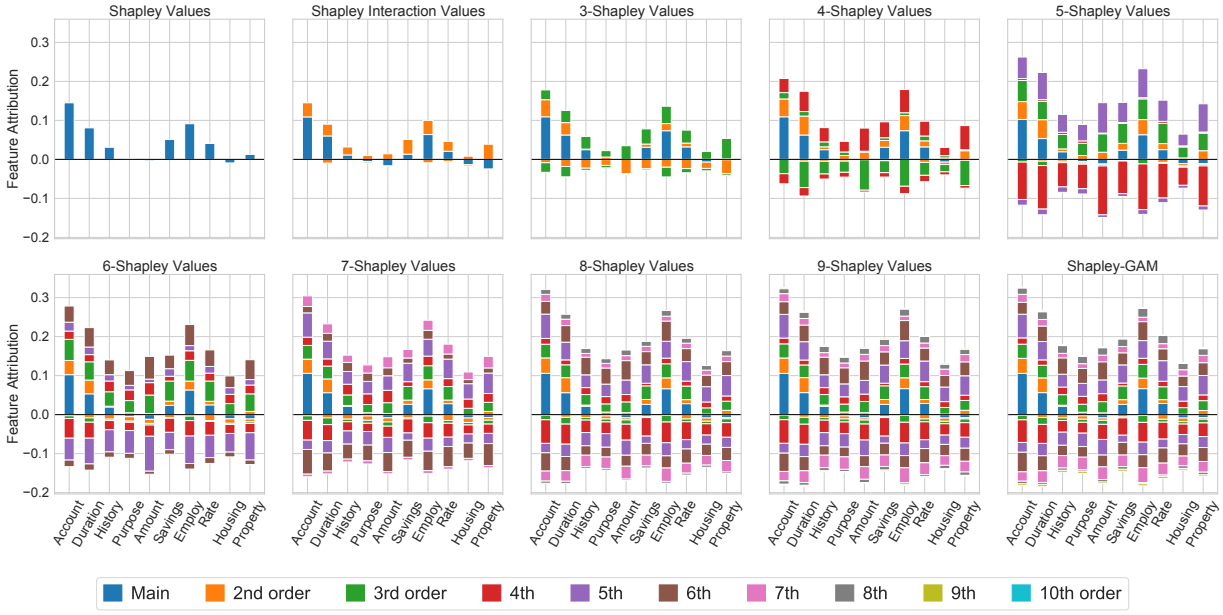


Figure H.15: n-Shapley Values for a kNN classifier and the first observation in our test set of the German Credit dataset.

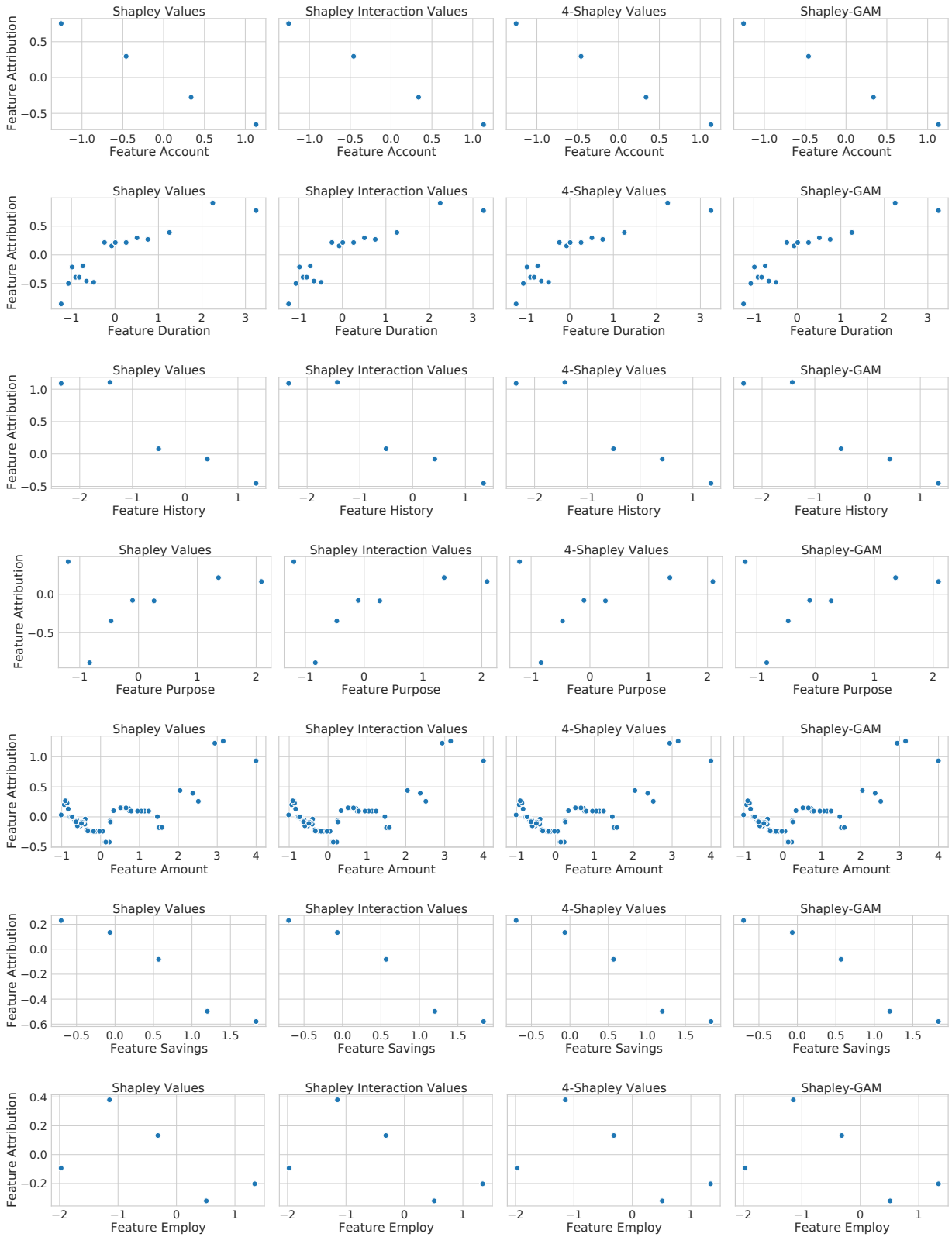


Figure H.16: Partial dependence plots for the Glassbox-GAM on the German Credit dataset. Depicted are the partial dependence plots of Φ_i^n for $n = \{1, 2, 4, 10\}$ and 7 different features.

H.4 California Housing

H.4.1 Glassbox-GAM

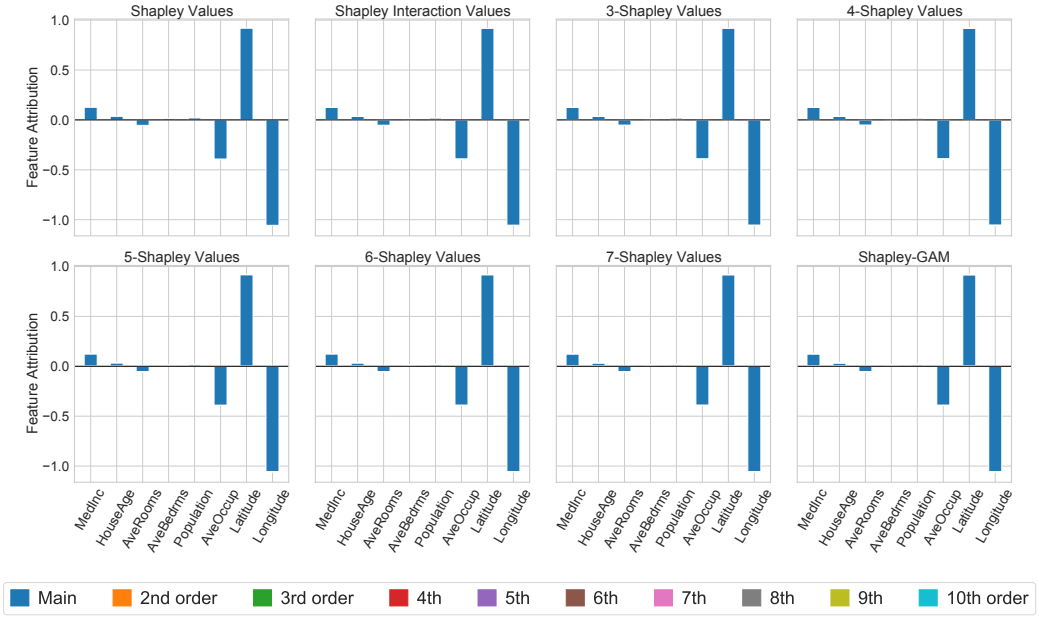


Figure H.17: n-Shapley Values for a Glassbox-GAM and the first observation in our test set of the California Housing dataset.

H.4.2 Gradient Boosted Tree

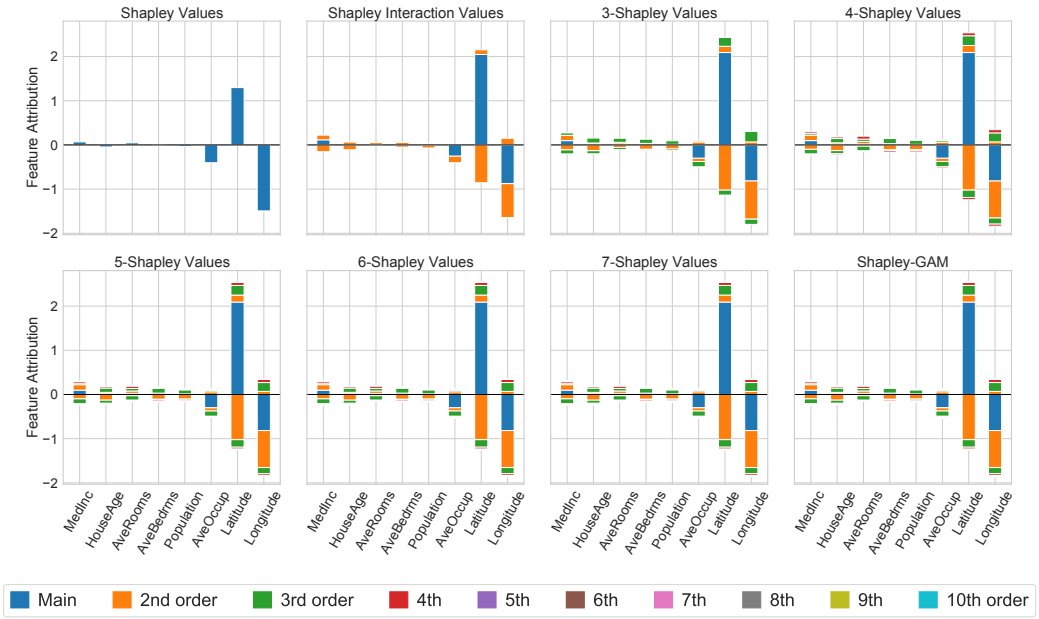


Figure H.18: n-Shapley Values for a Gradient Boosted Tree and the first observation in our test set of the California Housing dataset.

H.4.3 Random Forest

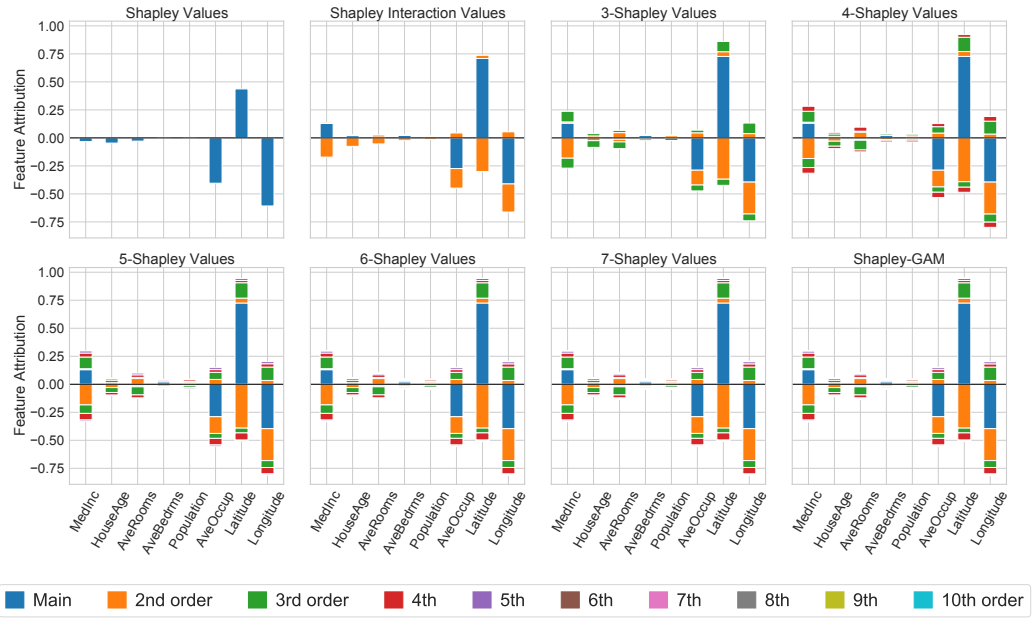


Figure H.19: n-Shapley Values for a Random Forest and the first observation in our test set of the California Housing dataset.

H.4.4 k-Nearest Neighbor

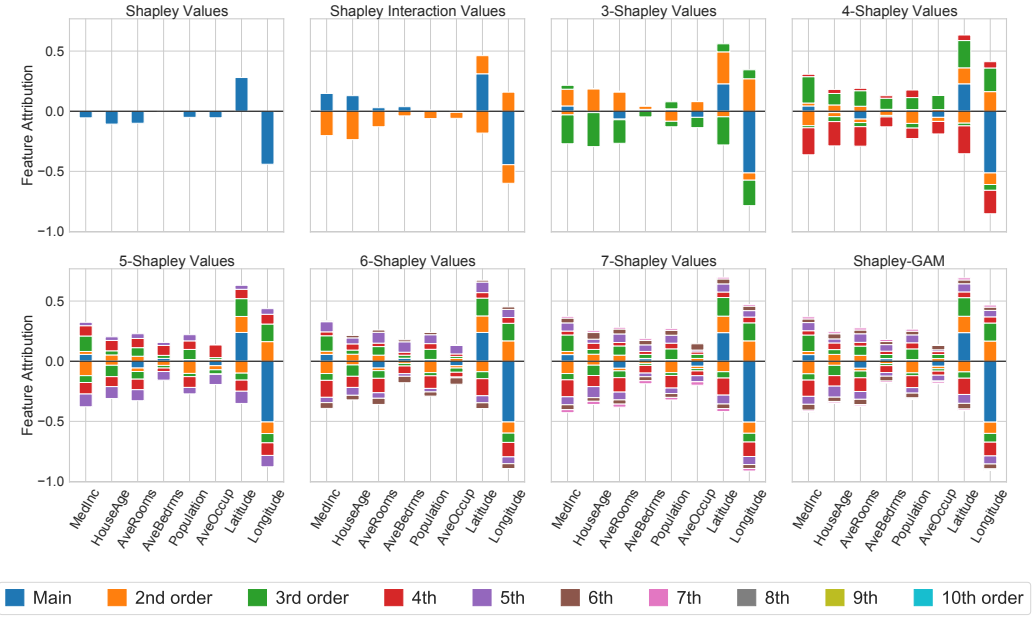


Figure H.20: n-Shapley Values for a kNN classifier and the first observation in our test set of the California Housing dataset.

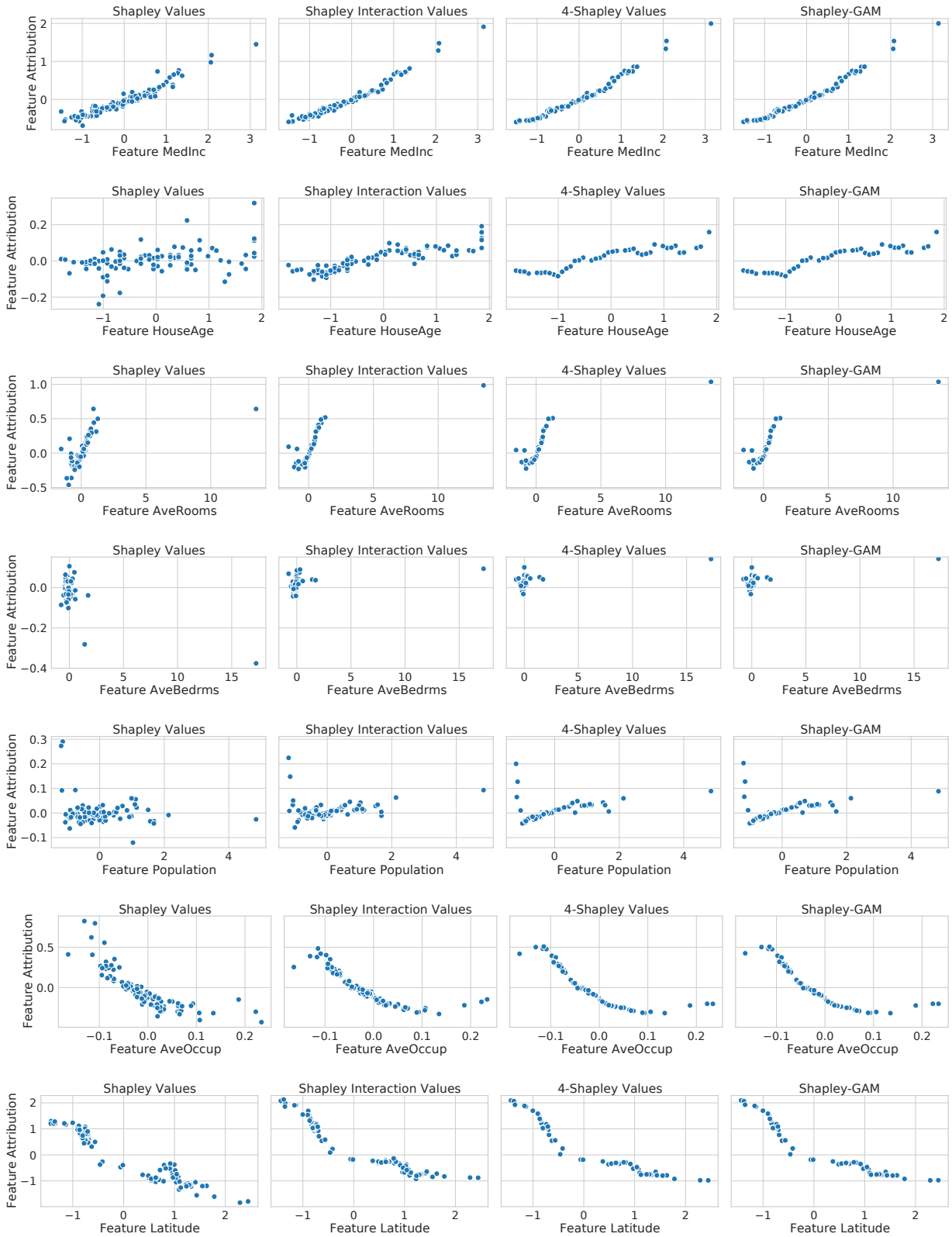


Figure H.21: Partial dependence plots for the gradient boosted tree on the California Housing dataset. Depicted are the partial dependence plots of Φ_i^n for $n = \{1, 2, 4, 10\}$ and 7 different features.