Master's Thesis

Levels Structures for Feature Attribution in Explainable AI

Application of Winter values an extensions of Shapley values to nested coalitions in the SHAP library

Benedek Fulop

Student number: 12290335

Date of final version: August 31, 2024

Master's programme: Data Science and Business Analytics

Supervisor: Prof. dr. Ilker Birbil

Second reader:

Faculty of Economics and Business

Contents

Chapter 1

Introduction

More and more complex machine learning models get deployed with increasing frequency. Given the momentous impact the widespread adoption of such highly accessible statistical models will have on society calls for explainable and safe models are increasing. We have seen the recent adoption of the Artificial Intelligence Act in the EU by the European Parliament (2024) and in the US the Blueprint for an AI Bill of Rights The White House (2022). These legislative efforts emphasise or require automatic explanations for deployments in high-risk contexts. For practitioners working on models in healthcare, finance or justice further standards and scrutiny frequently apply. Most crucially, users should have a clear, concise, prerequisite-free, and digestible explanation of how their data is used and how the models achieve their results. If we are to build the trust and discussion spaces needed for the long-term adoption of large statistical implements.

Such developments have led to an explosion of work in Explainable Artificial Intelligence (XAI). One of the more popular methods in the field is SHapley Additive exPlanation (SHAP) Lundberg and Lee (2017). The method's popularity can be largely attributed to its approach of unifying several methods using the mathematics from cooperative game theory. The framing is to view the explanation problem as a transferable utility game, with the features of the players and the model under evaluation as the payoff function for the cooperation. This work follows this philosophy adopting the concept of Winter values to provide explanations consistent with finite nested partitions, levels structures, of the feature space. These Winter values are an extension of Owen values to levels structures, which themselves are an extension of Shapley values to games with coalition structures. Such feature coalitions and nested coalitions represent restrictions to the interaction effects in the explanations. These are useful to keep perturbations on manifold and therefore, useful for explaining multi-modal models. We develop a new computation and implementation strategy for the Winter values. We conclude that Winter values closely approximate the Shapley values using either clustering or custom levels structures. We further showcase the robustness of the custom levels structures based Winter values to one form of adversarial attacks aimed at concealing model biases.

The remainder of this thesis is organized as follows. Chapter 2 contains a literature review, reviews the context in XAI, other works on coalitional explanations, works using the PartitionExplainer, and Winter values. Chapter 3 details the mathematical approach, and provides examples of the computation, and the code used by the PartitionExplainer. Chapter 4 showcases several useful properties of the methods. Chapter 5 concludes and outlines steps and directions for future research.

The developed exact method for Winter values is currently being merged into the open-source SHAP library. All the code for the work presented can be found in this repository https://github.com/CousinThrockmorton/shap.

Chapter 2

Literature Review

2.1 Model Agnostic Explanation Methods

To situate this research within the field of model explainability we can turn to several taxonomies such as Arrieta et al. (2020), Speith (2022), Schwalbe and Finzel (2023). Such categorisation all differentiate between transparent modelling methods and post-hoc explainable models. Methods such as linear regression or decision trees are considered inherently explainable, and interpretable, due to being decomposable with simple calculations used for imputation. More complex models like neural networks or gradient-boosting models construct functions that can be very large and therefore require explanation methods to describe them. Such post-hoc methods may rely on the form of the model they aim to explain these are model-specific methods or they attempt to describe the functional space based on the in and outputs, called model agnostic methods. Model agnostic methods became very popular since they can be applied to any model, do not require retraining, do not affect performance, and can be used by a third party to understand the model. The core idea behind these model-agnostic methods is to perturb the model's inputs to see how the model functions. This is usually done for a single data point, hence they are dubbed local methods. These methods draw on the cooperative game theory literature treating the model as a transferable utility cooperative game, features as players in the game, and the output is the payoff.

One of the foundational works in this field is the Shapley value, a solution concept for such cooperative games to fairly allocate the cooperation payoff Shapley et al. (1953). The value assigns to each player their average marginal contribution over all possible permutations of how the coalition can be formed. Following this, a wide variety of cooperative game theory solution concepts have been developed. In this work, we will focus on the literature that examines games where not all coalitions can be formed stemming from Aumann and Dreze (1974).

Being defined on all the possible ways the coalitions can form, that is the input of the solution concepts is in 2^P space therefore Deng and Papadimitriou (1994) argue:

"There is a catch, however: If the game is defined by the coalition values, there may be little to be said about the computational complexity of the various solution concepts, because the input is already exponential in n, and thus, in most cases, the computational problems above can be solved very 'efficiently".

Therefore, the computational burden associated with such explanations is considerable, especially for models with long imputation times. Nonetheless, there are several algorithms for hierarchical coalition values with polynomial run time in the number of features Besner (2022).

2.2 The PartitionExplainer

Lundberg and Lee (2017) used this game theoretic understanding to unify several existing approaches namely LIME Ribeiro et al. (2016), DeepLIFT Shrikumar et al. (2017), Layer-Wise Relevance propagation Bach et al. (2015) to one framework and library under the name SHapley Additive exPlanation (SHAP). This approach established the library as one of the most popular methods in XAI and was soon expanded with other game theory-based explanation methods amongst them the PartitionExplainer. This method uses coalitions of features to reduce the computational complexity of explanations and, for this is the default method for text and image data in the current version of the SHAP library. As far as we know, there have been no studies examining the logic and characteristics of the PartitionExplainer.

The PartitionExplainer is described in the SHAP documentation as a method that

"computes Shapley values recursively through a hierarchy of features, this hierarchy defines feature coalitions and results in the Owen values from game theory.".

Therefore, most studies use the PartitionExplainer due to its ability to group correlated features. For example, for image classification Podgorelec et al. (2020), spurious correlation detection on x-rays Sun et al. (2023), cell-penetrating peptides biochemical data Maroni et al. (2024), predictive process monitoring Warmuth and Leopold (2022), text classification Gücükbel (2023), mutual fund categorization with text data Vamvourellis et al. (2022) and many more. There is an issue, however, with the SHAP documentation, as we will show later in Section 3, the PartitionExplainer does not calculate the Owen values as described in Vamvourellis et al. (2022), Gücükbel (2023) but one special case of Winter values. This confusion is likely due to the Winter values frequently being described as the extension of Owen values to hierarchies or recursive Owen values in the game theory literature.

Winter and Owen values provide coalitional values by considering a subset of all possible permutations of features to those consistent with the coalition structure. In explainability, this

is analogous to including knowledge of the feature interactions in the model in the explanation method. For instance, we may know of the dependence structure of features and include this knowledge to get more consistent explanations Heskes et al. (2020), Frye et al. (2019). However, the causal dependence structure of the model for complex models is hard to come by in most cases.

There are many situations where practitioners may know something less comprehensive about the restrictions of feature interactions in the model. They may use inherently coalitional data like biological-knowledge graphs Martínez Mora et al. (2024). They develop coalitional-based explanations for graph data analogous to the Owen values and call for future work on Winter values-based explanations. Ferrettini et al. (2022) also develop explanations using Owen values using various clustering methods such as PCA, VIF and Spearman correlation. Maybe most interestingly we may know the model uses multi-modal data or an ensemble formulation. Burton and Al Moubayed (2023) use this information to extend the SHAP library. They find the clustering method of the PartitionExplainer lacking for text-tabular data as the correlations may or may not respect the type difference. Their masker essentially calls the PartitionExplainer separately for the tabular and text data providing more consistent explanations.

Other works have also explicitly called for the adoption of Winter values for model explainability. Rozemberczki et al. (2022) discusses a wide variety of Shapley value-based explanations and calls for coalition explanations. Patil and Främling (2023) discuss the use and interpretation of intermediate concepts for explanations and develop coalitional explanations using an alternate feature attribution strategy based on contextual importance and utility.

The SHAP method also received a fair amount of criticism in the form of a wide range of adversarial methods designed to disguise the actual biased or unfair model behaviour from an auditor using SHAP. This can be done via manipulating the model. Slack et al. (2020) use a "scaffolding" to detect the perturbations that often fall out of the base distribution of the data and conceal the models' bias. Dimanov et al. (2020) devise a new loss function to retrain models with little change in accuracy and conceal dependence on unfair features. Or by poisoning the data distribution Baniecki et al. (2022),Baniecki and Biecek (2022) or by biasing the data sampling to cherry-pick samples for the explanations Laberge et al. (2022). However, all of these methods assess their methods against some of the TreeExplainer, KernelExplainer and ExactExplainer methods. As far as we are aware there have been no inquiries into the robustness of the PartitionExplainer method of the SHAP library.

Contributions:

- Develop and adopt Winter values to explain machine learning models.
- Document the SHAP PartitionExplainer, a method for explanations under clustered bi-

nary partition trees.

- Demonstrate use cases and properties of coalitional explanations.
- Set future research directions for model agnostic and coalition explanations.

Chapter 3

Methods

We will first introduce the mathematical formulation of the Shapley value, before discussing coalitions and Owen values. Then nested coalition structures, level structures, and their interpretation at last the resulting Winter values. We will then develop several simple illustrative examples of the calculation and characteristics of the value.

In this section we will follow the notation and formulation of Winter (2002) with some slight modifications for the language of machine learning explainability. Let the set of players, from now on dubbed the features of the model, be $P = \{1,2,3,..,n\}$. The payoff function of the game, the model we aim to explain, f, is a function from the set of all possible coalitions $\Pi = 2^P$ to the set of real numbers $\mathbb{R}$. The additive feature attribution of interest ϕ assigns to each model f a vector of payoffs $\phi(f) = \{\phi_1, \phi_2, ..., \phi_n\}$ in $\mathbb{R}^n$. Then, feature i's measure of influence on the model f we denote by $\phi_i(f)$.

3.1 The Shapley value

The Shapley value assigns the marginal contribution to each feature in the model with respect to a uniform distribution over all possible permutations of features Π. Writing a permutation of the features $\pi \epsilon \Pi$ as a function $P \to P$. For each feature i we will denote $p^i_\pi = \{j : \pi(i) > \pi(j)\}$ as the set of all features preceding i in the order π. Using this we define the marginal contribution of feature i to the permutation π as $f(p^i_\pi \cup i) - f(p^i_\pi)$. The definition of Shapley value is the following:

$$\phi_i(f) = \frac{1}{|P|!} \sum_{\pi \epsilon \Pi} [f(p^i_\pi \cup i) - f(p^i_\pi)]. \tag{3.1}$$

The formula is quite intuitive summing over every possible permutation of the features averaging the marginal contribution of features.

We can formulate an equivalent equation as in the original Shapley paper Shapley et al. (1953) using the subset notation $S \subseteq P$ for coalitions instead of permutations.

$$\phi_i(f) = \sum_{S \subseteq P, S \ni i} \frac{1}{|P|} \frac{1}{\binom{|P|-1}{|S|}} [f(S \cup \{i\}) - f(S)]. \tag{3.2}$$

This formulation is more useful for the practical computation of explanations. However, the most useful is the axiomatic formulation that guarantees several useful properties for the results. Four axioms of (1) efficiency, (2) symmetry, (3) dummy, and (4) additivity uniquely characterise the above Shapley values.

Efficiency requires that the explanation values for the features precisely distribute the entire prediction value of the model.

EFFICIENCY. $\sum_{i \epsilon N} \phi_i(f) = f(N)$.

The notion of symmetry we will be using is the following: The features $i, j \in P$ are considered symmetric with respect to the model f if their marginal contribution is the same to any coalition. That is $\forall S \subset P$ with $i, j \notin S$, $f(S \cup i) = f(S \cup j)$.

SYMMETRY. If features i and j are symmetric with respect to model f, then $\phi_i(f) = \phi_j(f)$.

The Dummy axiom requires zero value attributed to features whose marginal contribution is zero for every coalition.

DUMMY. If i is a dummy feature, $f(S \cup i) - f(S) = 0$, $\forall S \subset P$, then $\phi_i(f) = 0$.

The value also is required to be additive on the space of all models.

ADDITIVITY. $\phi(f + g) = \phi(f) + \phi(g)$, where the model $f + g$ is defined by $(f + g)(S) = f(S) + g(S)$ $\forall S$.

These axioms uniquely characterise the Shapley value in equation 3.1, 3.2. For proof of uniqueness see Winter (2002). These axioms are intuitive and general and therefore, can be applied to any model on any feature space resulting in comparable model-agnostic explanations.

3.2 Level Structures, The Owen value Owen (1977)

The knowledge of feature interactions can be described via coalitions of features, or nested coalitions of features, called levels structure. To start off, consider the set of features P partitioned into groups according to coalition structure $B = (S_1, ..., S_m)$, that is $\bigcup S_j = P$ and $S_i \cap S_j$ for $i \neq j$. We can then think of the set of orders $\Pi(B)$ by ordering the groups and then the components of the groups. Formally, $\Pi(B) = \{\pi \in \Pi;$ if $i, j \in S_k$ and $\pi(i) < \pi(r) < \pi(j)$, then $r \in S_k\}$ is the set of all permutations consistent with B.

The Owen value of feature i in the model f with coalition structure B is given by:

$$\phi_i(f, B) = \frac{1}{|\Pi(B)|} \sum_{\pi \epsilon \Pi(B)} [f(p^i_\pi \cup i) - f(p^i_\pi)]. \tag{3.3}$$

The benefit of this formulation is reducing the number of elements in the sum from $\Omega(2^{|P|-1})$ for a single feature for Shapley values to $\Omega(2^{|S_i|+m-1})$ where $|S_i|$ is the number of features in the coalition containing i and m the number of coalitions in B. For Owen values too we can write the formula using the coalitions.

$$\phi_i(f, B) = \sum_{S \subseteq P \setminus \{j\}} \sum_{T \subseteq \mathcal{S}_j \setminus \{i\}} \frac{1}{|P|} \frac{1}{\binom{|P|-1}{|S|}} \times \frac{1}{|\mathcal{S}_j|} \frac{1}{\binom{|\mathcal{S}_j|-1}{|T|}} \left[v(S \cup T \cup \{i\}) - v(S \cup T) \right], \tag{3.4}$$

In this formulation we sum over all other coalitions the feature is not part of at some level of the partition tree $S \subseteq P \setminus \{j\}$ and over all other coalitions in the features "own" group $T \subseteq S_j \setminus i$. If we examine the "weights" applied to the marginals we can see that these are simply the binomial coefficients for the features/coalitions.

3.3 The Winter value Winter (1989)

We can further stack these partitions $B = (B_1, B_2, ..., B_m)$ such that B_i is a refinement of B_{i+1}, specifically if $S \in B_i$, then $S \subset T$ for some $T \in B_{i+1}$.

We can interpret such level structures as representing the strength of relationships between features. This means that B_m the coarsest partition has the weakest relations and B_1 the strongest relationship between features. There are numerous examples we can give for such relationships. In the game theory literature Winter (1989) used trade relationships to describe levels structures where B_m describes free trade agreements, then lower levels B_{m-1} describes the forming countries, then states, regions, municipalities and so on. We can also draw a parallel between levels structures and partonomies or meronomies from linguistics describing part-whole relationships. We can also think of various graph, subgraph structures as financial, biological or social networks. Therefore, the levels structures can encode the correlation, and computational relations known about the structure of the model f, broader than a strict dependence structure.

We can then make payoffs dependent on the cooperation structure described by thinking of the permutations Π as the order features collect their payoffs. We will only consider orders in which no feature i follows player j if there is another feature k who is "closer" to feature i and hasn't appeared yet.

For a given level structure $B = (B_1, B_2, ..., B_m)$, define.

$\Pi_m = \{\pi \in \Pi$; for each $l, j \in S \in B_m$ and $i \in N, \pi(l) < \pi(i) < \pi(j)$ implies $i \in S\}$

and,

$\Pi_r = \{\pi \in \Pi_{r+1}$; for each $l, j \in S \in B_r$ and $i \in N, \pi(l) < \pi(i) < \pi(j)$ implies $i \in S\}$.

Meaning we permute the top level B_m coalitions first then lower levels successively with Π_1 containing all permutations consistent with B. This construction also results in the total payoffs for a coalition at a level that is independent of any coalition structures at the lower levels, see Theorem 3. Winter (1989).

The Winter value then can be written as:

$$\phi_i(f, B) = \frac{1}{|\Pi_1|} \sum_{\pi \epsilon \Pi_1} [f(p_\pi^i \cup i) - f(p_\pi^i)] \tag{3.5}$$

In the coalitional formulation, the only difference from the Owen value is that it is defined over multiple layers of nested coalitions.

$$\phi_i(f, B) = \sum_{S \subseteq B_{j+1} \setminus \{j\}} \sum_{T \subseteq \mathcal{B}_j \setminus \{i\}} \frac{1}{|B_{j+1}|} \frac{1}{\binom{|B_{j+1}|-1}{|S|}} \times \frac{1}{|B_j|} \frac{1}{\binom{|B_j|-1}{|T|}} \left[v(S \cup T \cup \{i\}) - v(S \cup T)\right]. \tag{3.6}$$

The Winter value relaxes the symmetry axiom of the Shapley value and in their place posits two axioms for individual features and coalitions.

INDIVIDUAL SYMMETRY. If k and j are two symmetric features with respect to the model f, where every level $1 \leq i \leq m$, and any non-singleton coalition $S \in B_1$ then $k \in S$ if $j \in S$, and $\phi_i(B, f) = \phi_j(B, f)$. Note the symmetry individual symmetry axiom can be removed if the last layer is of the individual features $B_1 = (\{1\}, \{2\}, ..., \{n\})$.

COALITIONAL SYMMETRY. Let $B = (B_1, B_2, ..., B_m)$ be a levels structure. For each level $1 \leq i \leq m$ if $[S], [T] \in [B_i]$ are symmetric features in the model $([B_i], F^i)$ and S, T are subsets of the same coalition in B_j for $j > i$ then $\sum_{r \in S} \phi_r(B, f) = \sum_{r \in T} \phi_r(B, f)$

3.4 Illustrative Examples - Winter values

As an illustrative example let's consider a machine learning model for a dating app. The model consumes multiple types of data that we will consider to be singular features, for simplicity, of the user's pictures, location data, text prompts, and other characteristics and preferences. In this case explainability is important for developers wanting to improve model behaviour, regulators aiming to ensure safety, fairness, and privacy of users, or the company wanting to provide customers with advice on how to use their algorithm.

First, we will illustrate how the coalitional explanations offer more consistent intuitive explanations in the case of related features. Consider the game behind the explanations for the images for a second in Figure 3.1. The images are comprised of a complex set of features extracted via convolution in most machine-learning applications, shown as shutters here. Other potential features are represented by the data symbol the chip symbol at the root represents the overall model. For simplicity let's consider the model where features *1*, and *2* are substitutable for each other as is often the case for most image models.

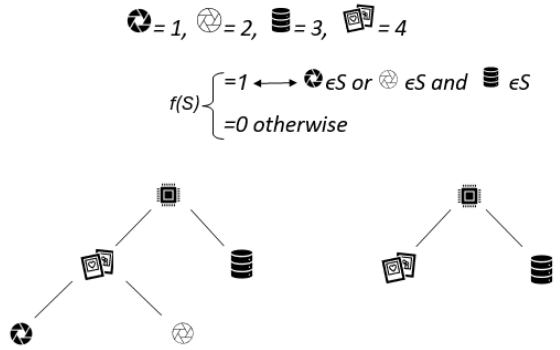
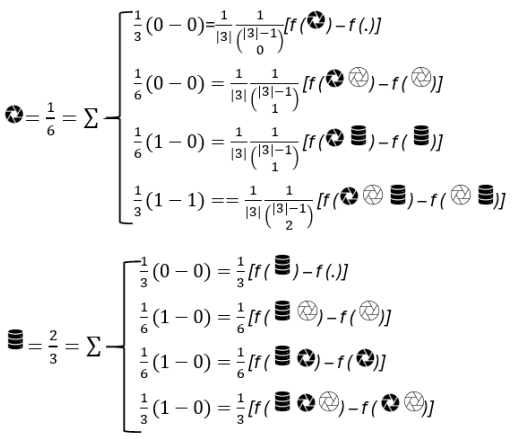
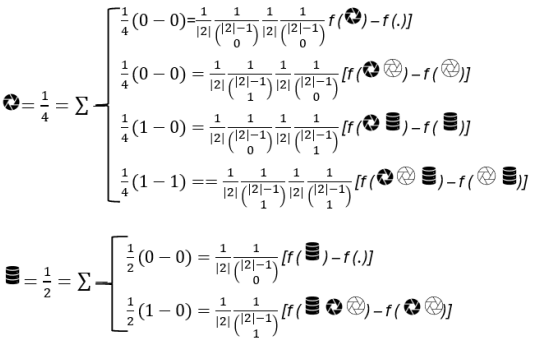


Figure 3.1: The computation of coalitional values



(a) Shapley values calculation



(b) Owen values calculation

Figure 3.2: Calculations for Shapley vs. coalitional values

Applying the Shapley values in Figure 3.2 it is simple to see that it assigns $\phi_1 = \phi_2 = \frac{1}{6}$, due to symmetry, and $\phi_3 = \frac{2}{3}$ and due to additivity the value for the images is $\frac{1}{3}$. The coalitional value, for this simple case the Owen value, takes the coalition structure into account by "removing the bottom nodes". The game between coalitions, on the right, now gives $\phi_4 = \frac{1}{2}$ and $\phi_3 = \frac{1}{2}$ and $\phi_1 = \phi_2 = \frac{1}{4}$. We would argue that the coalitional value for this model is more consistent with the model's functioning and intuition than the Shapley values. Moreover, investigating the monotonicity of such solution concepts Young (1985) conclude that most other them do not diverge far from the Shapley values.

Now, consider providing an explanation via Shapley values practitioners put all features regardless of type, on a level playing field. Calculating the marginal contribution of each feature to all other possible feature coalitions is shown in Figure 3.3.

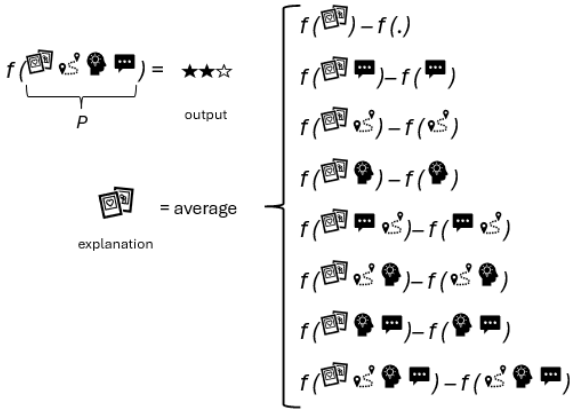


Figure 3.3: The computation of the Shapley value for the dating app model

Not only can these perturbations create non-sensical, off-manifold data instances. But most machine learning models capture a complex set of feature relations, independence of features is the exception, and this is even more likely to be the case between different types of data. Providing explanations via Shapley values then would result in biased explanations. Consider, for instance the image data in this case may be undervalued as its marginal contribution is likely to be small with respect to non-image data resulting in a lower value.

Now, let us consider Owen values with coalition structure first shown in 3.4. With a simple partition of the feature space, for example, we may know that images are processed via a surrogate model. Using this information we can reason that most marginal values for the overall value result in very similar values, so there is no need to consider those marginals.

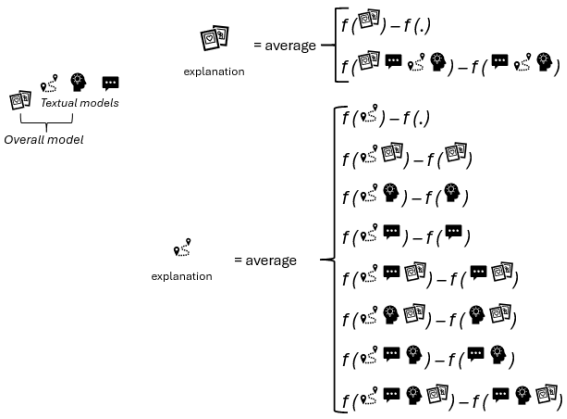


Figure 3.4: Computation for the Owen values for two features

We can think of the computation of the images as modifying the weights on the marginals so as to consider only relevant marginal effects. Not only does this reduce the number of calculations but the permutations used are more likely to "make sense" that is, be closer to the base distribution used for training.

For levels structures, we are considering two subsets, all the sibling nodes on the path, and all the sibling coalitions/features of the feature/coalition of interest. This can be neatly illustrated on a n-ary tree as shown in Besner (2022)

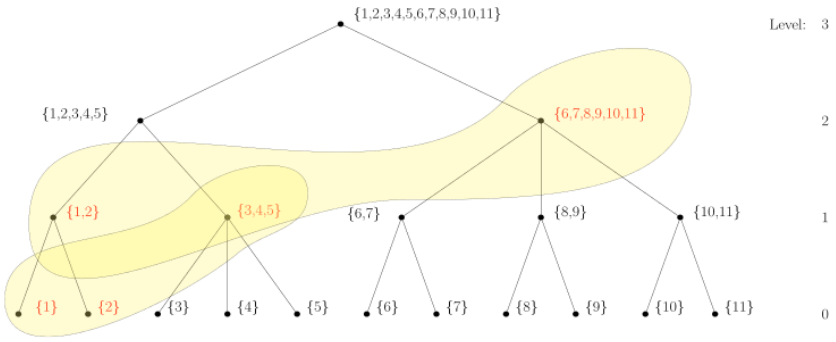


Fig. 5. Components needed for player 1 to form relevant coalitions to calculate the nested Owen Levels value.

Figure 3.5: The relevant coalitions for nested Owen Levels values, that is Winter values Besner (2022)

In our example, for the location data, this is considering the images as out of coalition features and the text and tabular data as local coalition features, therefore calculating the marginals of the product between the power sets of these two groups.

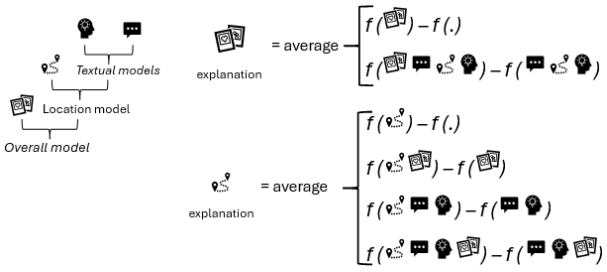


Figure 3.6: Computation for the Winter values for two features

3.5 Computation of the clustering method

To illustrate the working of the clustering method we will examine the Simple California Demo one of the model agnostic SHAP example notebooks using the PartitionExplainer on the scikit learn dataset of the median house prices of districts in California Pace and Barry (1997). The dataset has 8 features longitude, latitude, median income, population, average number of rooms, bedrooms, and occupancy of districts. The PartitionExplainer using the default option uses the correlation distances to create a binary tree of coalitions Figure 3.7.

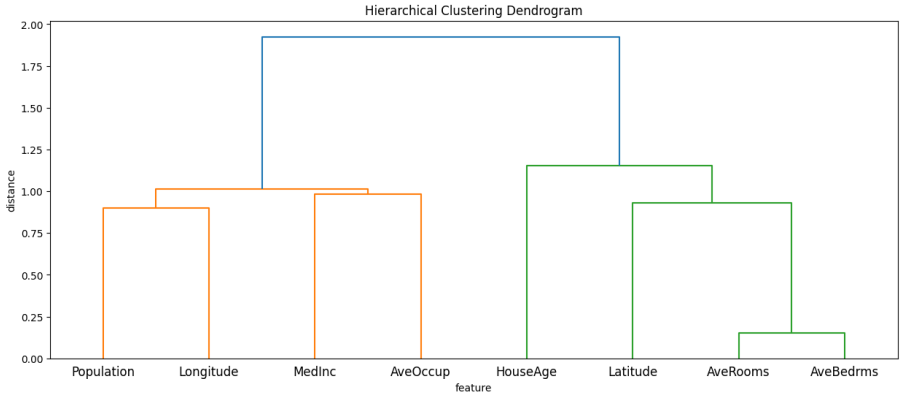


Figure 3.7: The binary clustering tree of the California housing features constructed by pdist from scipy

The clustering method for which the pseudo-code is shown in Programs **??** traverses the tree and calculates the Winter values. The recursive function traverses the binary tree creating the masks in batches by adding the left and right children's masks to the mask of the current node and the base array of all features turned off.

We take a look at the first three iterations of the clustering method to understand how the value gets calculated. In all cases we start the loop with all features turned off, we call this the null mask.

$$\begin{bmatrix} MedInc & HouseAge & AveRooms & AveBedrms & Population & AveOccup & Latitude & Longitude \\ o & o & o & o & o & o & o & o \end{bmatrix}$$

Then we turn on the features on the left and right of the node. See these printed on the right and left side of the matrix, x's corresponding to the active features and o's to the masked features.

$$\left[\begin{array}{cccccccc|cccccccc} x & o & o & o & x & x & o & x & o & x & x & x & o & o & x & o \end{array}\right]$$

As the clustering method uses a binary tree proceeding for every node on the null mask and the left and right children will generate all possible combinations of masks as the possible permutations are also binary. The binary tree also guarantees that every resulting mask will be unique.

$$\left[\begin{array}{cccccccc|cccccccc} o & x & x & x & x & o & x & x & x & x & x & x & o & x & x & o \\ o & o & o & o & x & o & o & x & x & o & o & o & o & x & o & o \\ x & x & o & o & x & x & o & x & x & o & x & x & x & x & x & x \\ o & x & o & o & o & o & o & o & o & o & x & x & o & o & x & o \end{array}\right].$$

Here the colours correspond to the "lineage" of the masks, red corresponds to the null mask, blue the right and black the left sub-tree. The clustering method is further parameterised by the $fixed\_context$ parameter that restricts recursion to proceed only on the null mask ($fixed\_context = 0$) or the children nodes ($fixed\_context = 1$). This way the resulting values can be considered the feature being "present" and "absent" respectively.

The clustering method uses the distance from the clustering hierarchy to tell if a leaf is reached and calculates the average marginals that sum up to the winter values. Since every level adds 2 nodes the weights for each level correspond to a simple average. Using the clustering distance also allows for early stopping behaviour as users can modify the distance matrix to negative values which terminates the descent to the nodes below. Then the *lower_credit* function is used to spread the attribution equally between constituent features.

As the tree is binary at worst for balanced clustering guarantees quadratic exact runtime, uses batching, parallel processing, and the $fixed\_context$ option the clustering method is highly efficient for large feature sets. Hence it's used as the default explanation method for text and image data. However, it is not possible to calculate Winter values for non-binary levels structures using this method. Moreover, it is tedious to input domain knowledge even in those cases when the levels structures happen to be binary.

3.6 Computation of the partition method

To be able to encode domain knowledge of the structure of the model we will be using a native Python class for n-ary trees populated via nested dictionaries of the coalitions with the features as leaves. We have also decided to encode all level-specific information in the n-ary tree object.

That is each node in the class has four attributes, first its name, second all of its children, third the power set of its sibling nodes, and fourth the weights for every element of the power set i.e. the binomial coefficients. This is done for quick retrieval and in order to save computational resources during tree traversal.

The current implementation of the Winter values utilises Python's itertools and separates computing the weights and the marginals from the model calls. Itertools is used as it allows for efficient computation of Cartesian products. The computation of the values is done through the following steps. For complete pseudo-code please see Programs **??**.

- Depth-first traversal of the n-ary tree. We recursively collect all the paths to the leaves along with all the power sets of sibling nodes and weights on the path.
- Compute the Cartesian product of the sibling power sets and weights. Then add the end leaves to get all the marginals for a feature.
- Evaluate the model on the unique power sets.
- Calculate the Winter values using the formula in 3.6

This formulation has several drawbacks. The biggest opportunity for improvement may be that it doesn't allow for top-down computation, early stopping, and parallelisation unlike the binary tree implementation. This design choice also causes significant memory usage as the product operation is costly for longer paths. Therefore, the clustering method is more efficient for the usually deeper binary clustering trees.

There exist polynomial time algorithms for the general Winter values and similar hierarchical values Besner (2022). However, their formulation utilises sub-games. These sub-games are induced from the original cooperative game played by players inside a coalition to allocate the fraction of the payoff for their coalition from the level above. It is hard to formulate such analogous restricted models that only take into account certain features for a subsection of models' output.

Chapter 4

Experiments

4.1 Coalitional tractability

We will showcase the tractability of Winter values and how they extend Shapley values and the clustering PartitionExplainer. We will utilise some of the example notebooks of the SHAP library. Specifically, we will use one large example the League of Legends Win Prediction with XGBoost notebook in the SHAP library that showcases the TreeExplainer method on the Campanelli (2017) dataset from Kaggle. And the previously discussed Simple California Demo. We will show that the partition method is an extension of the hierarchical clustering method by reproducing the results of the exact and clustering methods. Then we will show how the Winter values represent coalitionally consistent explanations.

In the California notebook, the example model used is an XGBoost regression. We will generate explanations for this model using the ExactExplainer, TreeExplainer, the PartitionExplainer with the binary method, and the custom partition method.

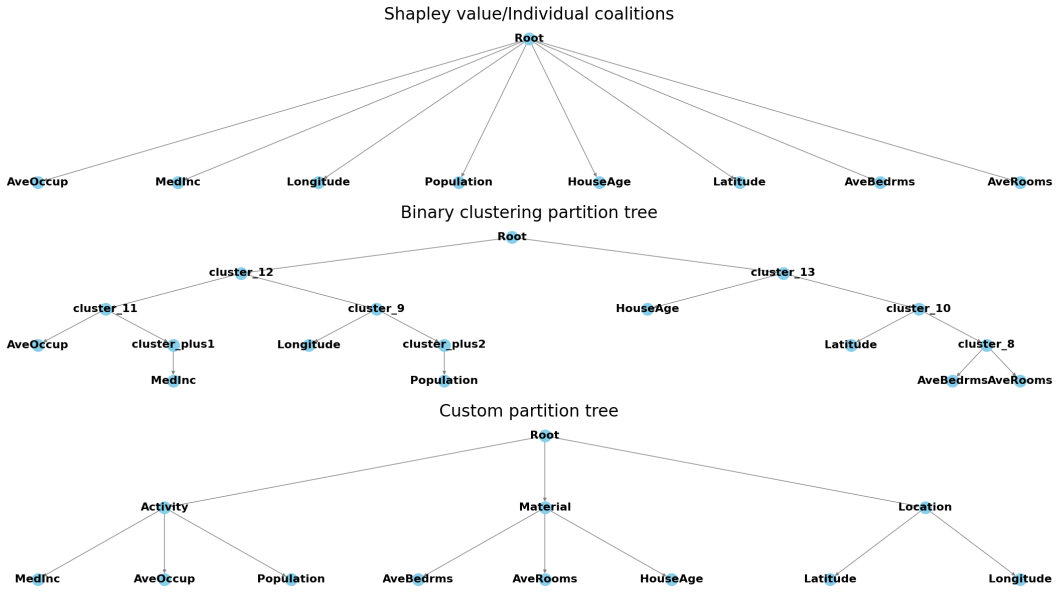
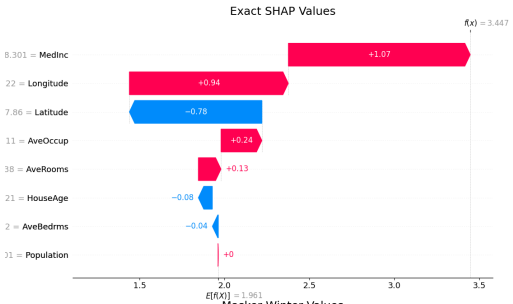
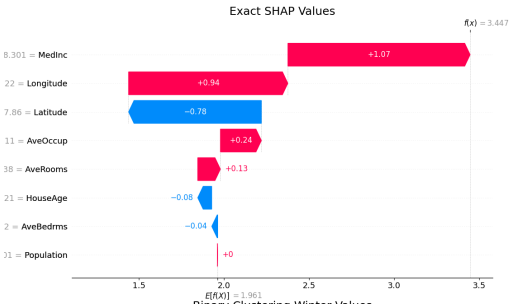
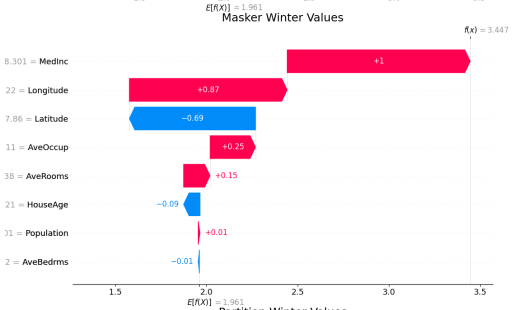


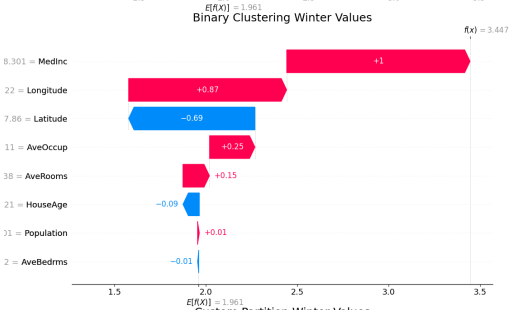
Figure 4.1: Levels structures of California housing features for our explanations

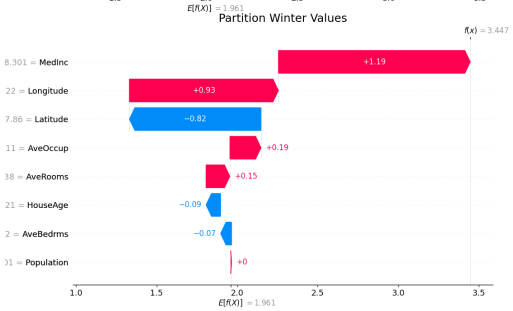
We will therefore generate explanations for the following levels structures of features in Figure 4.1. Figure 4.1 shows the result of explanations for these hierarchies on the left using the original SHAP methods, and the right using the custom partitions with our method. Our expectation is that if the method we developed works, passing the tree structure used by the ExactExplainer and PartitionExplainer the method returns the exact same explanations.

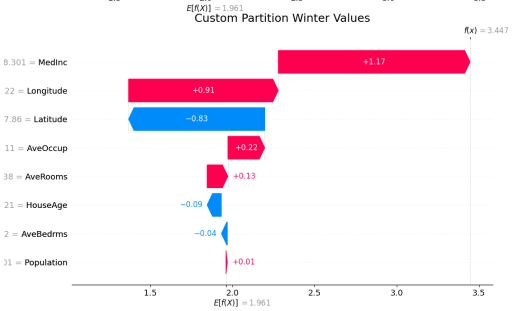












(a) Using ExactExplainer, clustering PartitionExplainer

(b) Using the custom partitions in PartitionExplainer

We can see that the explanations from left to right match exactly and the explanations in the second two rows by the Winter values are close to the exact Shapley values. In fact, the Winter value explanations are closely approximating the exact Shapley values for most samples, see Figure 4.2. Note for this model the TreeExplainer results perfectly approximate the exact values.

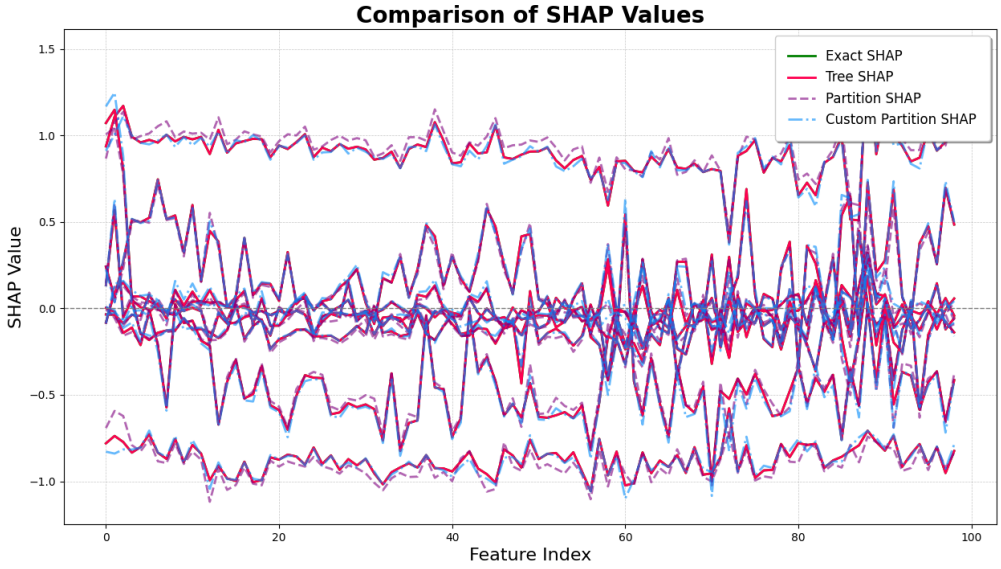


Figure 4.2: Explanations for 100 instances with each line corresponding to a feature.

To showcase how coalitional explanations can be considered more explainable and more useful in many cases we look at the League of Legends Win Prediction with XGBoost notebook as it has a larger feature space with 67 features for over 180-thousand matches in the game. The example notebook uses an XGboost classification model. We complemented it with a simple 3-layer Multi-Layer-Preceptron(MLP) model using Pytorch to showcase the model-agnostic nature of the explanation method. The model is too large to use the ExactExplainer without resorting to sampling, therefore, we will use the KernelExplainer and the clustering method PartitionExplainer. The levels structure we will be using is visualised in Figure 4.3.

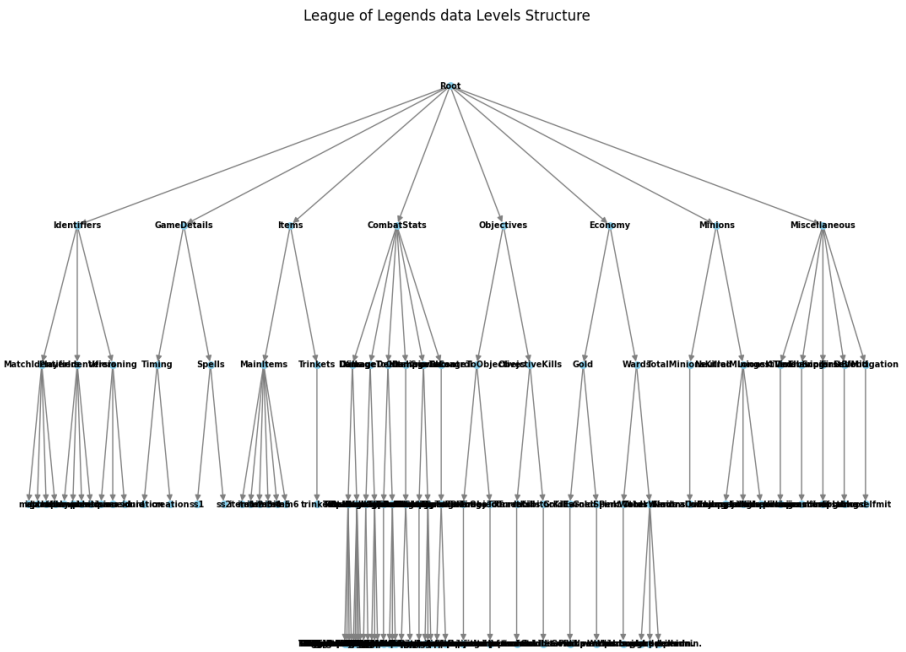
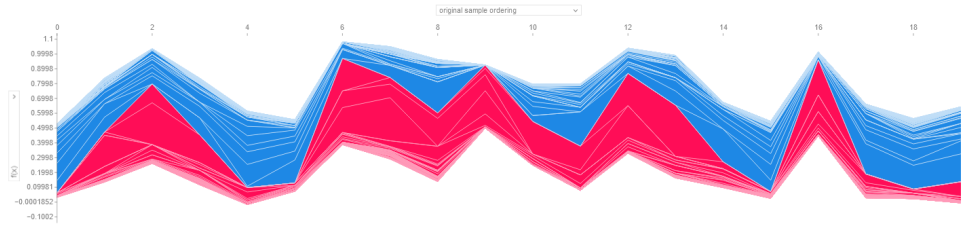
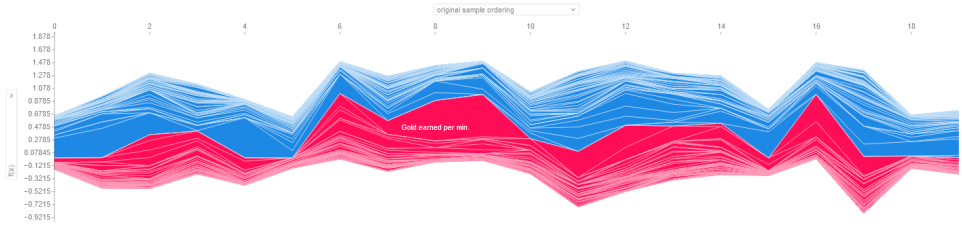


Figure 4.3: Levels Structure used for the League of Legends dataset

To begin we examine the approximate Shapley values and clustering Winter values for the XGB and the MLP models in Figure 4.4. We can see that a wide variety of features contribute to the explanations for the different instances.

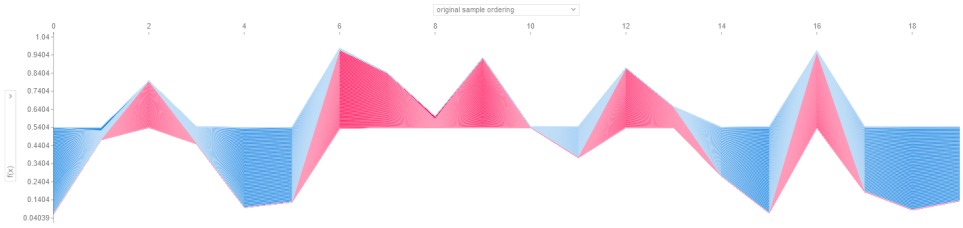


(a) XGB model

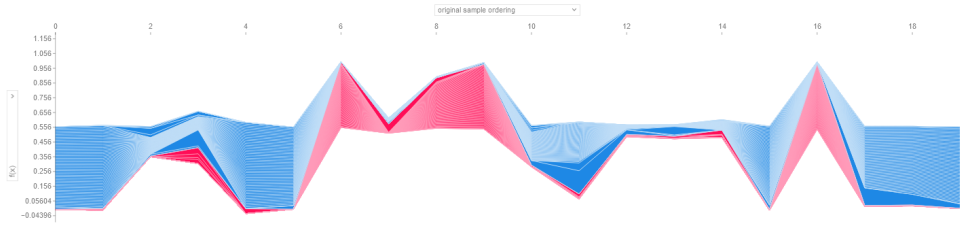


(b) MLP model

Figure 4.4: KernelExplainer explanations for 20 instances presented in the same order



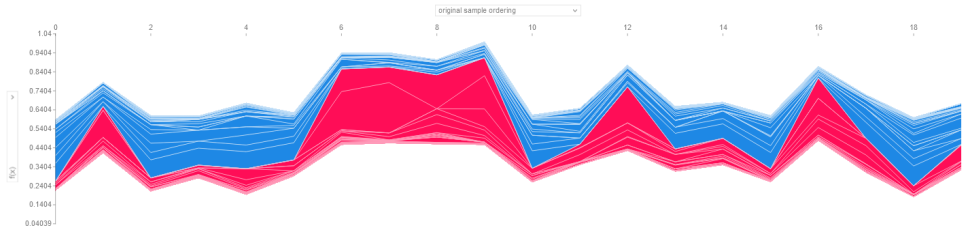
(a) XGB model



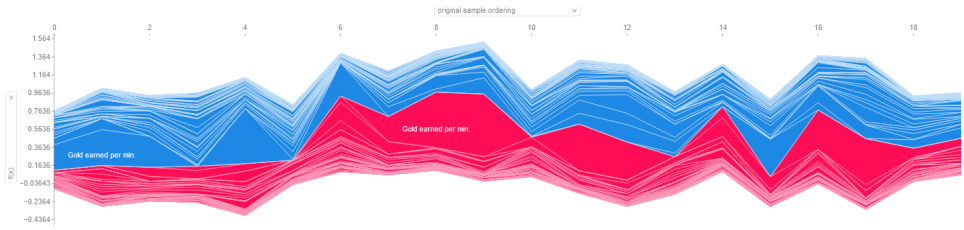
(b) MLP model

Figure 4.5: Clustering PartitionExplainer explanations for 20 instances presented in the same order

The Winter values spread the influence of the marginals over the binary tree therefore we end up with a lot of small feature attributions negating their usefulness for explanations. Most importantly, plotting 69 features on any plot is not human interpretable.



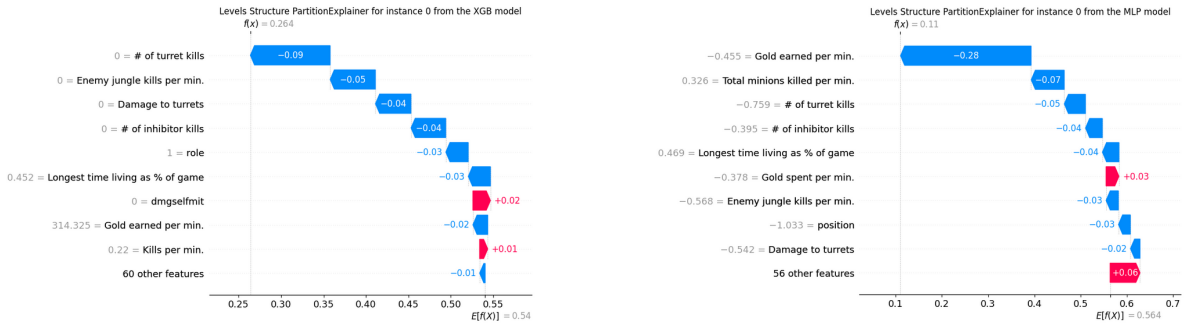
(a) XGB model



(b) MLP model

Figure 4.6: Levels structure PartitionExplainer explanations for 20 instances presented in the same order

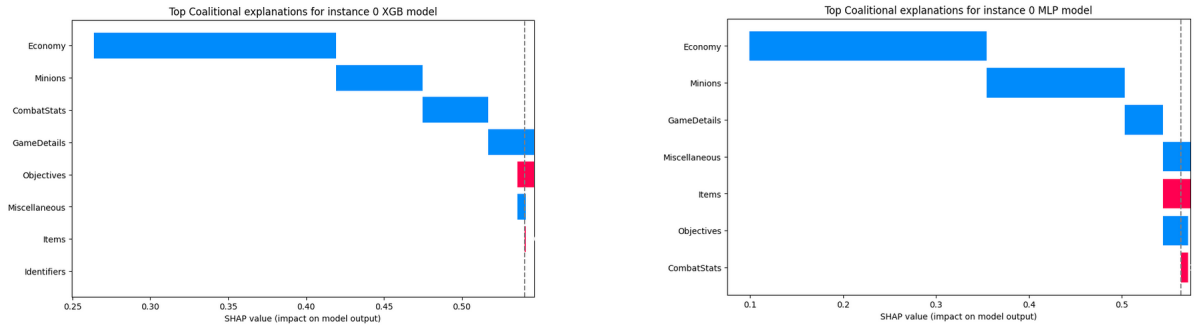
In Figure 4.8 we can see that the Winter values for the levels structure return explanations with more variation. Furthermore, comparing the KernelExplainer results we can see that the explanations are more consistent, especially for the MLP model.



(a) XGB model (b) MLP model

Figure 4.7: Levels structure PartitionExplainer explanations for instance 0

A more user-friendly way to plot feature attributions for the Winter values would be to plot the coalition values. Importantly, these are values consistent with the levels structure and therefore these explanations capture the interactions between these coalitions, disregarding interactions not consistent with the coalitions.



(a) XGB model (b) MLP model

Figure 4.8: Top coalitions of the levels structure PartitionExplainer explanations for instance 0

4.2 Robustness to adversarial attacks

To test the robustness of the Winter values we used the experiments from Slack et al. (2020). We will use two datasets used in the study. The COMPAS dataset of the criminal history and demographics of defendants in Broward County, Florida Angwin et al. (2022). And the German credit dataset from the UCI machine learning dataset of 1000 loan applications including financial and demographic information Blake and Mertz (1999). To replicate the results of the paper we used their Github repository upgrading the SHAP module and then testing the hierarchical clustering and custom partition Winter value methods. Our hypothesis is that since the Winter values are calculated by restricting the perturbations to those consistent with the levels structure these are more likely to be closer to the original background distribution and provide consistent explanations despite the scaffolding applied.

We can easily replicate the findings of Slack et al. (2020). Their adversarial "scaffolding" successfully biases the explanations hiding the behaviour of the model.

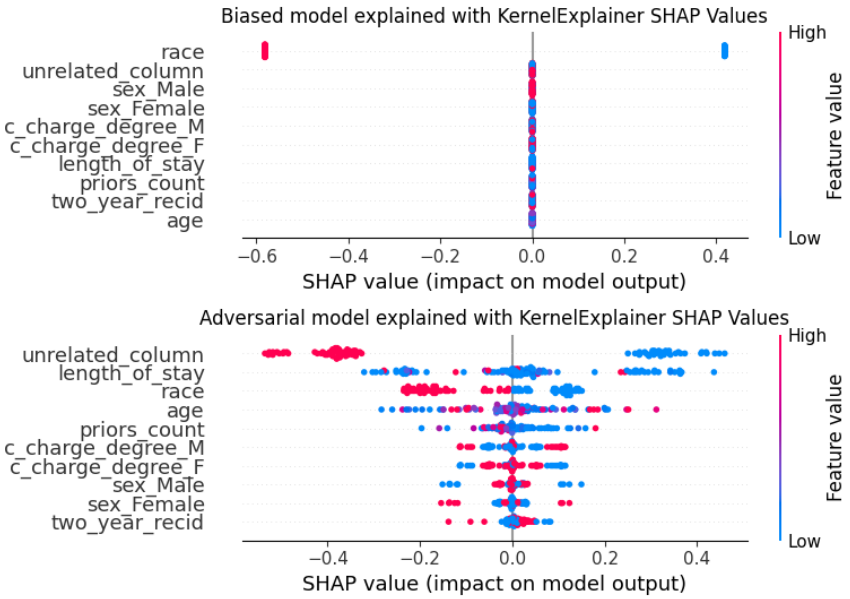


Figure 4.9: 100 Explanations using the KernelExplainer for the biased and adversarial models on the COMPAS dataset

In figure 4.9 we plotted explanations for 100 instances for the biased and the adversarial model. The biased model makes its prediction entirely on the race feature, and this racism is concealed by their method with several features getting larger SHAP values assigned by the KernelExplainer method.

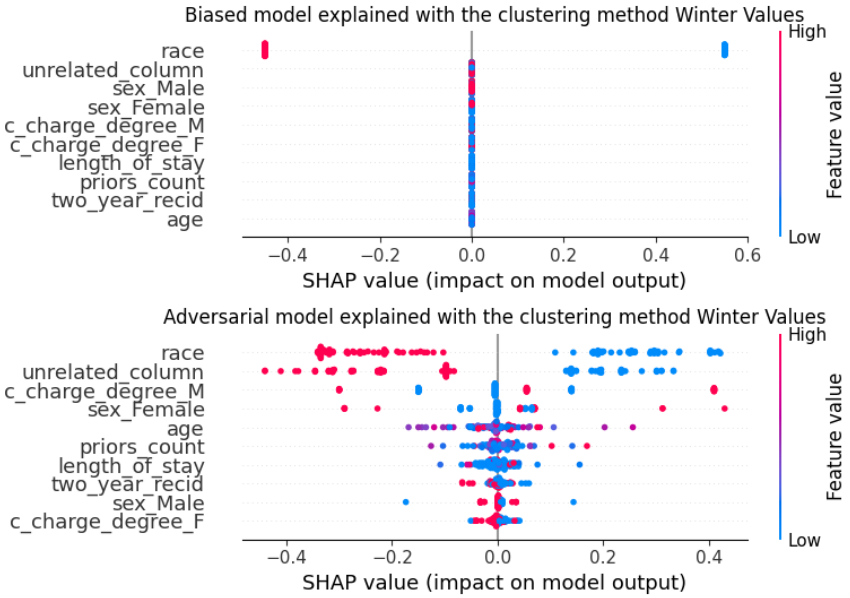


Figure 4.10: 100 Explanations using the clustering PartitionExplainer for the biased and adversarial models on the COMPAS dataset

Using the hierarchical clustering-based partition method we can however identify the bias of the model despite the adversarial scaffolding 4.10. We do see large values attributed to some other features such as the noise.

To see if a more interpretable levels structure may fit the base distribution better we defined the following hierarchy Figure 4.11.

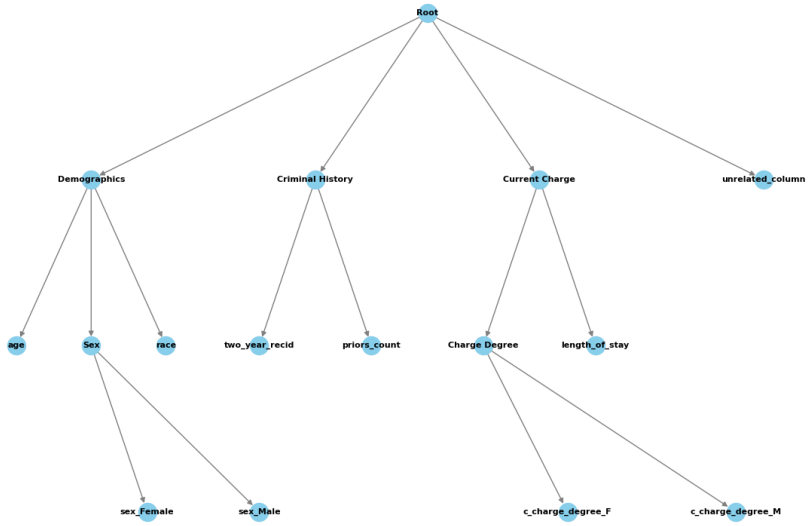


Figure 4.11: Example levels structure for the COMPAS dataset

Using this levels structure we indeed find that the bias of the model is detectable and the unrelated features get smaller values assigned to them. We achieve similar results on the German credit dataset with the custom partitions resulting in the most consistent explanations.

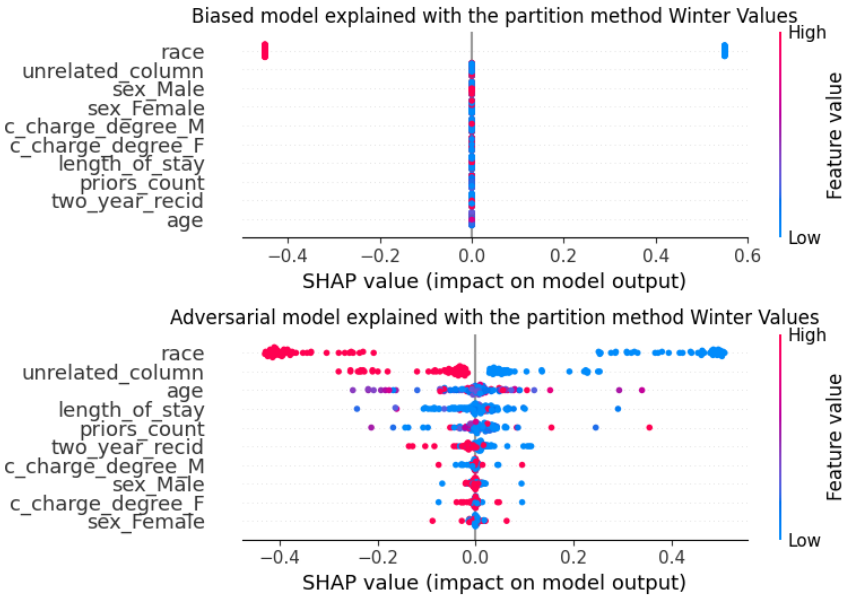


Figure 4.12: 100 Explanations using the levels structure PartitionExplainer for the biased and adversarial models on the COMPAS dataset

This example showcases the robustness of the PartitionExplainer for this particular adversarial attack. This result however by no means guarantees the robustness of the PartitionExplainer against attacks using different methods and further research is required to assess the true robustness of this and other methods such as the GradientExplainer from SHAP.

Chapter 5

Conclusion

We have in this work examined how a solution concept from cooperative game theory Winter values can be applied to explain machine learning models. We have done this by; defining the various exact solution concepts currently implemented in one of the most popular XAI libraries SHAP, and shown with illustrative examples how the Winter values extend these to models with feature coalitions. Then we examined a restricted implementation of the Winter values by the clustering PartitionExplainer, and developed a new implementation to calculate the Winter values for any levels structures of features. To test and showcase the uses of such methods we have used a variety of models and datasets. As a result we saw that using custom levels structures the Winter values closely approximate the Shapley values. They can be applied to models with large feature spaces, are coalitionally consistent, and robust against some adversarial attacks by using more salient perturbations.

The current work does have numerous limitations. First, the implementation of the Winter values could be improved in several ways, such as adopting batching and parallel processing, study into the time-complexities of different implementation and levels structures, and adoption for different data modalities. These changes would allow to study the of Winter values for explaining multi-modal and ensemble models. This extension is particularly interesting as it would allow consistent masking for data types. Going further, it may be of interest to study the different ways the levels structures can be constructed. Firstly, assessing the different distance measures used by the clustering method for images for instance. Or to examine how linguistic partonomies as intermediate concepts can be used to explain language models. There remains further work on assessing the robustness of Winter value explanations to the other adversarial attacks. As an exact method it is robust against attacks utilising the sampling process however the limits to robustness of such game-theory based explanations is not well understood.

Appendix A

Appendix

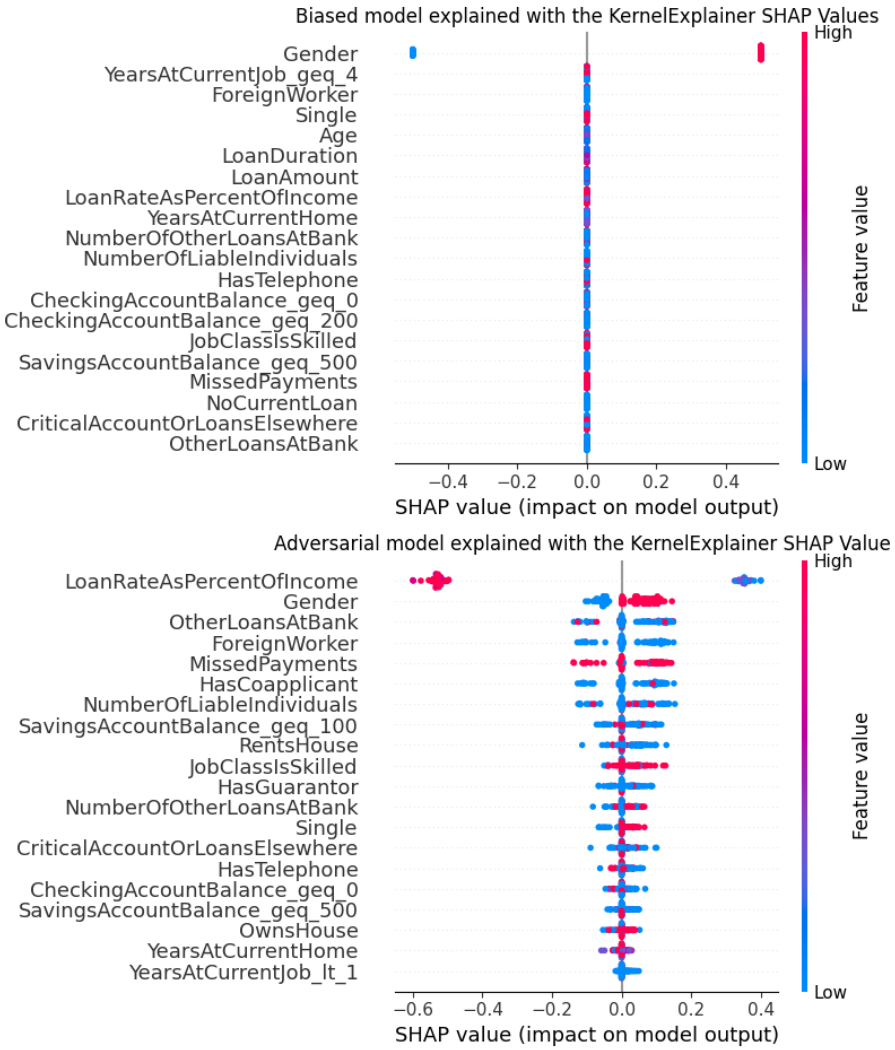


Figure A.1: 100 Explanations for models on the German credit data recreating results from Slack et al. (2020)

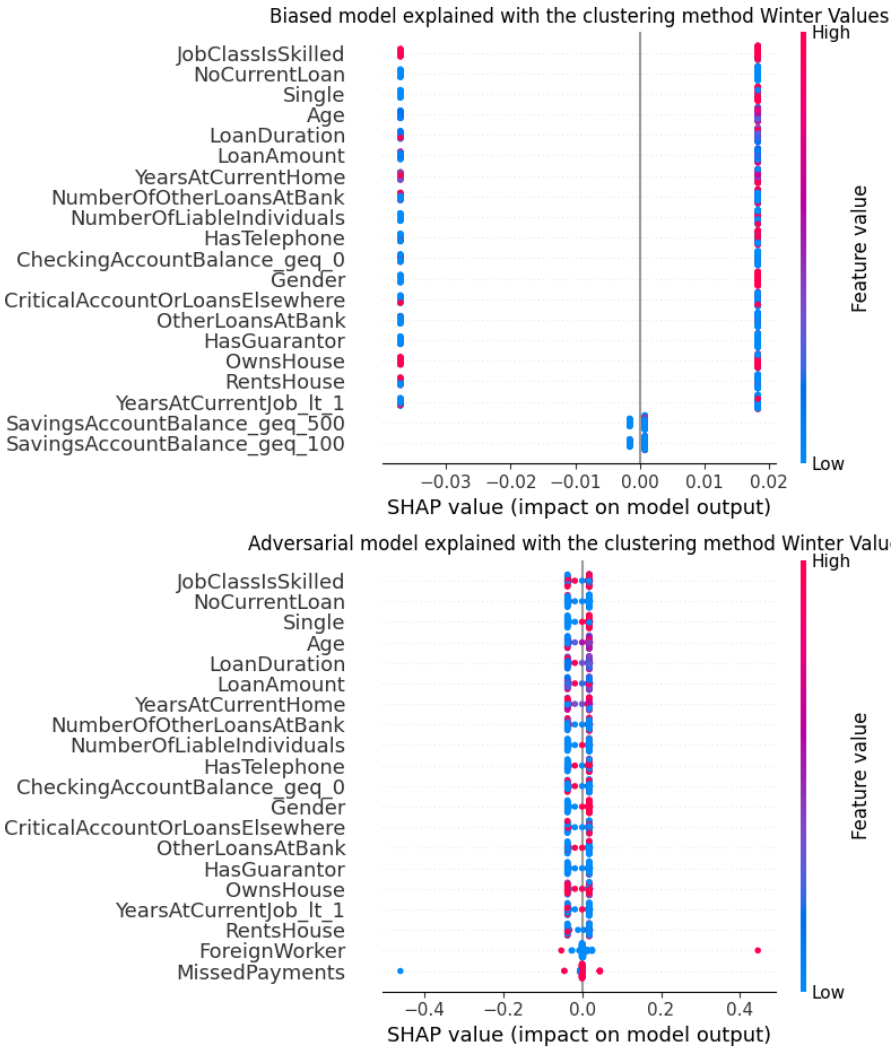


Figure A.2: 100 Explanations for models on the German credit data with the clustering PartitionExplainer

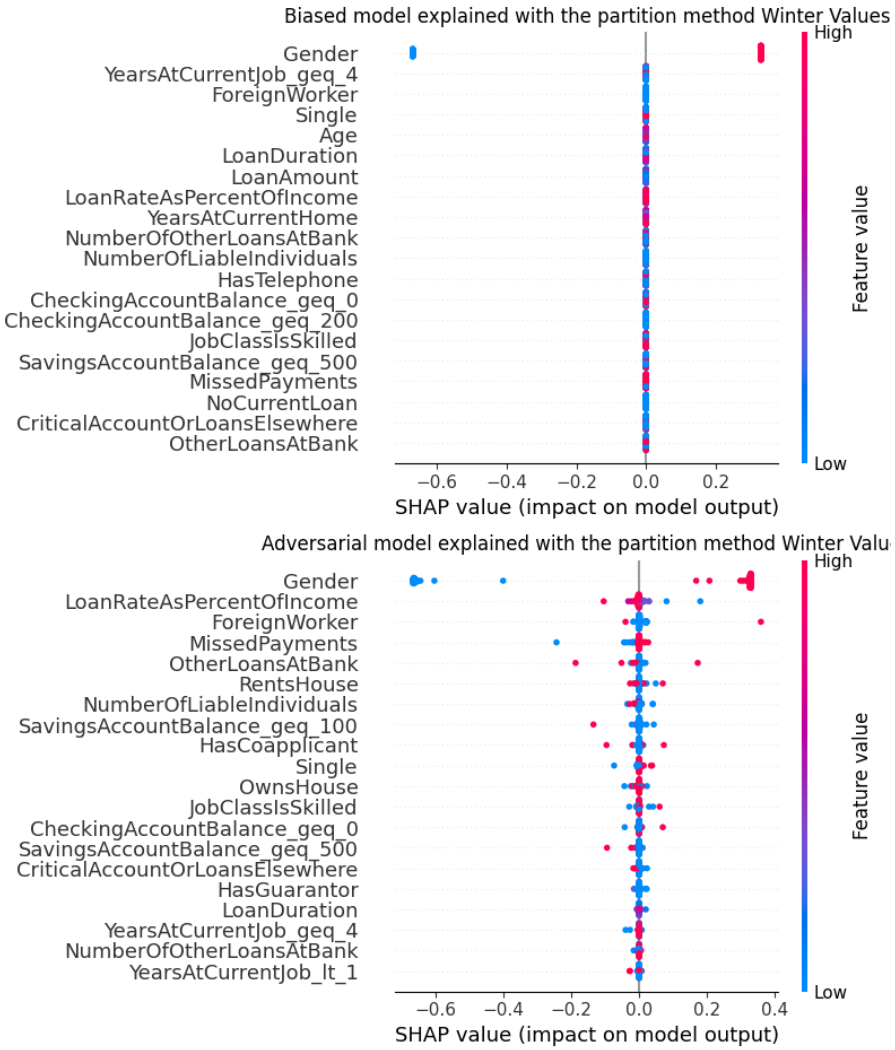


Figure A.3: 100 Explanations for models on the German credit data with a custom levels structure in the PartitionExplainer

Appendix B

Programs

Data: fm, f00, outputs
Result: winter_values
Step 1: Initialize Tree Structure
root ← `Node("Root")`;
`build_tree`(*partition_tree, root*);
Step 2: Generate all marginals for nodes
combinations_list ← `traverse_tree_and_generate_products`(*root*);
masks, keys ← `create_masks`(*root, feature_names*);
masks_dict ← `map_keys_to_masks`(*keys, masks*);
unique_masks ← `get_unique_masks`(*masks_dict*);
Step 2: Evaluate Model on Unique Masks
mask_results ← `evaluate_model_on_masks`(*fm, unique_masks*);
Step 3: Compute SHAP Values
winter_values ← `initialize_winter_values`(*len(fm)*);
mask_mappings ← `map_combinations_to_masks`(*combinations_list, masks_dict, unique_masks*);
foreach *key* ***in*** *mask_mappings* **do**
    off_indexes, on_indexes, weights ← mask_mappings[key];
    foreach *off_idx, on_idx, weight* ***in*** *zip(off_indexes, on_indexes, weights)* **do**
        off_result ← mask_results[unique_masks[off_idx]];
        on_result ← mask_results[unique_masks[on_idx]];
        shap_values[key] ← shap_values[key] + (on_result - off_result) * weight;
    end
end

Algorithm 1: Explain with Partition Tree

```
Data: m00, f00, f11, ind, weight
Result: winter_values
begin
    if maximum evaluations are reached then
        m00, f00, f11, ind, weight ← q.get()[2];
        winter_values[ind] += (f11 - f00) * weight;
        break;
    end
    (lind, rind) ← get_children(current_node);
    if get_distance(current_node) < 0 then
        calculate the marginal;
        winter_values[ind] += (f11 - f00) * weight;
        return;
    end
    (m10, m01) ← create_masks(lind, rind);
    f10 ← evaluate_model(m10);
    f01 ← evaluate_model(m01);
    new_weight ← weight;
    new_weight /= 2
    if fixed_context is None or fixed_context == 0 then
        winter_recursive(m00, f00, f10, lind, new_weight);
        winter_recursive(m00, f00, f01, rind, new_weight);
    end
    if fixed_context is None or fixed_context == 1 then
        winter_recursive(m01, f01, f11, lind, new_weight);
        winter_recursive(m10, f10, f11, rind, new_weight);
    end
end
```

Algorithm 2: Winter Recursive

Bibliography

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2022). Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115.

Aumann, R. J. and Dreze, J. H. (1974). Cooperative games with coalition structures. *International Journal of game theory*, 3:217–237.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

Baniecki, H. and Biecek, P. (2022). Manipulating shap via adversarial data perturbations (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12907–12908.

Baniecki, H., Kretowicz, W., and Biecek, P. (2022). Fooling partial dependence via data poisoning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 121–136. Springer.

Besner, M. (2022). Values for level structures with polynomial-time algorithms, relevant coalition functions, and general considerations. *Discrete Applied Mathematics*, 309:85–109.

Blake, C. and Mertz, C. (1999). Repository of machine learning. *University of California at Irvine*, page 75.

Burton, J. and Al Moubayed, N. (2023). Shap explanations for multimodal text-tabular models.

Campanelli, P. (2017). League of legends ranked matches. `https://www.kaggle.com/datasets/paololol/league-of-legends-ranked-matches`. Accessed: 29 June 2024.

Deng, X. and Papadimitriou, C. H. (1994). On the complexity of cooperative solution concepts. *Mathematics of operations research*, 19(2):258.

Dimanov, B., Bhatt, U., Jamnik, M., and Weller, A. (2020). You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods. In *ECAI 2020*, pages 2473–2480. IOS Press.

European Parliament (2024). Corrigendum to the position of the european parliament adopted at first reading on 13 march 2024 with a view to the adoption of regulation (eu) 2024/..... of the european parliament and of the council laying down harmonised rules on artificial intelligence and amending various regulations and directives (artificial intelligence act). P9_TA(2024)0138, COM(2021)0206 – C9-0146/2021 – 2021/0106(COD), cor01, 19.4.2024.

Ferrettini, G., Escriva, E., Aligon, J., Excoffier, J.-B., and Soulé-Dupuy, C. (2022). Coalitional strategies for efficient individual prediction explanation. *Information Systems Frontiers*, 24(1):49–75.

Frye, C., Rowat, C., and Feige, I. (2019). Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *arXiv preprint arXiv:1910.06358*.

Gücükbel, E. (2023). Evaluating the explanation of black box decision for text classification.

Heskes, T., Sijben, E., Bucur, I. G., and Claassen, T. (2020). Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *Advances in neural information processing systems*, 33:4778–4789.

Laberge, G., Aïvodji, U., Hara, S., Khomh, F., et al. (2022). Fool shap with stealthily biased sampling. *arXiv preprint arXiv:2205.15419*.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Maroni, G., Stojceski, F., Pallante, L., Deriu, M. A., Piga, D., and Grasso, G. (2024). Lightcppgen: An explainable machine learning pipeline for rational design of cell penetrating peptides. *arXiv preprint arXiv:2406.01617*.

Martínez Mora, A., Nilsson, S., Polychronopoulos, D., and Ughetto, M. (2024). Community-aware explanations in knowledge graphs with xp-gnn. *bioRxiv*, pages 2024–01.

Owen, G. (1977). Values of games with a priori unions. In *Mathematical economics and game theory: Essays in honor of Oskar Morgenstern*, pages 76–88. Springer.

Pace, R. K. and Barry, R. (1997). Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297.

Patil, M. S. and Främling, K. (2023). Do intermediate feature coalitions aid explainability of black-box models? In *World Conference on Explainable Artificial Intelligence*, pages 115–130. Springer.

Podgorelec, V., Pečnik, Š., and Vrbančič, G. (2020). Classification of similar sports images using convolutional neural network with hyper-parameter optimization. *Applied Sciences*, 10(23):8494.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Rozemberczki, B., Watson, L., Bayer, P., Yang, H.-T., Kiss, O., Nilsson, S., and Sarkar, R. (2022). The shapley value in machine learning. *arXiv preprint arXiv:2202.05594*.

Schwalbe, G. and Finzel, B. (2023). A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, pages 1–59.

Shapley, L. S. et al. (1953). A value for n-person games.

Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMlR.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020). Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186.

Speith, T. (2022). A review of taxonomies of explainable artificial intelligence (xai) methods. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 2239–2250.

Sun, S., Koch, L. M., and Baumgartner, C. F. (2023). Right for the wrong reason: Can interpretable ml techniques detect spurious correlations? In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 425–434. Springer.

The White House (2022). Blueprint for an ai bill of rights: Making automated systems work for the american people. Office of Science and Technology Policy.

Vamvourellis, D., Toth, M., Desai, D., Mehta, D., and Pasquali, S. (2022). Learning mutual fund categorization using natural language processing. In *Proceedings of the Third ACM International Conference on AI in Finance*, pages 87–95.

Warmuth, C. and Leopold, H. (2022). On the potential of textual data for explainable predictive process monitoring. In *International Conference on Process Mining*, pages 190–202. Springer.

Winter, E. (1989). A value for cooperative games with levels structure of cooperation. *International Journal of Game Theory*, 18:227–240.

Winter, E. (2002). The shapley value. *Handbook of game theory with economic applications*, 3:2025–2054.

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