Inherent Inconsistencies of Feature Importance

Anonymous Author(s)
Affiliation
Address
email

Abstract

The rapid advancement and widespread adoption of machine learning-driven technologies have underscored the practical and ethical need for creating interpretable artificial intelligence systems. Feature importance, a method that assigns scores to the contribution of individual features on prediction outcomes, seeks to bridge this gap as a tool for enhancing human comprehension of these systems. Feature importance serves as an explanation of predictions in diverse contexts, whether by providing a global interpretation of a phenomenon across the entire dataset or by offering a localized explanation for the outcome of a specific data point. Furthermore, feature importance is being used both for explaining models and for identifying plausible causal relations in the data, independently from the model. However, it is worth noting that these various contexts have traditionally been explored in isolation, with limited theoretical foundations.

This paper presents an axiomatic framework designed to establish coherent relationships among the different contexts of feature importance scores. Notably, our work unveils a surprising conclusion: when we combine the proposed properties with those previously outlined in the literature, we demonstrate the existence of an inconsistency. This inconsistency highlights that certain essential properties of feature importance scores cannot coexist harmoniously within a single framework.

1 Introduction

Feature Importance scores gauge the contribution of each feature to an outcome of a model. Most model-agnostic feature importance scores use a two-step process: in the first step, value is assigned to subsets of the features. In the second step, the score of individual features is derived from the values of subsets. This two-step process allows for a discussion about the expected behavior of the value function and the feature importance score. Many feature importance scores have been proposed in the literature: the bivariate-association [1] evaluates a feature's importance based on its conditional attributes, independent of other features, ablation-studies [2, 3, 4] quantify a feature's significance by assessing its contribution when removed from the entire feature set, SHAP [5] computes feature importance as the mean of its contributions across various subsets of features, and MCI [6] determines importance as the maximal contribution among all possible feature subsets (see Table 1). SHAP and MCI use an axiomatic approach in which the expected behaviors are defined as properties, and the functions are derived to satisfy these properties.

Feature importance scores can be categorized by two main attributes: the scope, i.e. *local* vs *global*, and the objective, i.e. *data* vs *model*. Methods focusing on local interpretations seek to explain individual predictions (e.g., the role of each feature in a patient's diagnosis [7]).

Submitted to the Workshop "XAI in Action: Past, Present, and Future Applications" at the 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

Table 1: Examples of common feature importance scores. φ denotes the importance function, which takes ν, the value function, and a feature $f \in \mathcal{F}$ as inputs, and assigns an importance score.

Name	**Feature importance score: $\varphi(\nu, f)$**
Bivariate	$\nu(\{f\})$
Ablation	$\nu(\mathcal{F}) - \nu(\mathcal{F} \setminus \{f\})$
Shapley	$\sum_{\mathcal{S}\subseteq\mathcal{F}\setminus\{f\}} \frac{\lvert\mathcal{S}\rvert!(\lvert\mathcal{F}\rvert-\lvert\mathcal{S}\rvert-1)!}{\lvert\mathcal{F}\rvert!} \cdot (\nu(\mathcal{S} \cup \{f\}) - \nu(\mathcal{S}))$
MCI	$\max_{\mathcal{S}\subseteq\mathcal{F}\setminus\{f\}}(\nu(\mathcal{S} \cup \{f\}) - \nu(\mathcal{S}))$

Conversely, methods focusing on global interpretation try to understand how each feature affects a phenomenon (e.g., the role of each gene in a particular disease [8, 9]). Along the second axis, the data and the model are distinguished by the type of conclusion required. The objective of explaining the data is to infer conclusions about the world that are encoded in the data, as the scientist does in his research [10, 11, 12]. The objective of explaining the model, however, is to use an explanation to monitor and debug a model, to ensure it is working as intended (e.g., as the engineer does for security purposes [13, 14]).

Table 2 maps feature importance research according to the local vs. global and data vs. model settings. Most feature importance scores thus far have focused on explaining models, although the data scenario has also been gaining increased attention in recent years. However, the quadrant of the data-local setting is still unexplored in the field of explainable AI. Perhaps this is due to the challenge of providing an accurate explanation as to why a specific outcome (rather than an average result) came into being (rather than being calculated by a model). For example, which characteristic of John Doe is responsible for the fact that he did, or did not, suffer a stroke? These types of questions pertain to individual causal effects that are notoriously difficult to estimate [15, 16].

Several studies have examined the relations between the different settings: Lundberg et al. [17] presented a global score that is computed by combining local scores, hence indicating that at least the local and the global settings are not independent. Covert et al. [1] proposed a method of assigning global importance to features, which draws a connection with the local feature importance score of *SHAP* [5]. Chen et al. [18] defined distinctions between the data and the model and argued that the nature of an explanation depends on what one seeks to explain – the data or the model. Nevertheless, most studies focus solely on one setting. The studies that do consider multiple settings, often do not present an explicit set of expectations for the relations between importance scores under the different settings.

In this work, we establish the expected behavior of feature importance scores across diverse contexts. Our objective is to formalize a set of properties that capture the anticipated consistency between local and global interpretations, as well as the alignment between data-driven and model-based assessments. An intriguing extension of this framework is the introduction of the data-local scenario, which, in theory, can be achieved by integrating our properties with axiomatic methods, wherein expected behaviors are rigorously defined, and functions are derived accordingly. In the global-data scenario, we employ a set of properties introduced by Catav et al. [6], implying the MCI importance score. Leveraging our proposed local-global relation, one can derive the expected data-local importance score. Conversely, in the local-model scenario, Lundberg and Lee [5] present the SHAP importance score as the only function that satisfies their proposed set of properties. Here, utilizing our proposed data-model relation, one can similarly obtain the expected data-local importance score. However, to our surprise, in both cases, even a modest set of requirements leads to contradictions.

The paper is structured as follows: Section 2 consists of a formulation of the framework that generalizes the two-step process of feature importance to all settings. In Section 3 we focus on the local-global consistency: we present its properties and then demonstrate that they are incompatible with a previous result that defined the data-global setting. At the end of the section, we provide a brief discussion of the nature of the contradiction. Section 4 follows a similar structure as the latter, except addressing the case of the data-model consistency, which

Table 2: Examples of feature importance scores and their categorization according to the global/local and data/model settings.

	Global	Local
Model	Additive-Importance-Measures [1] Bivariate-Association [1] Ablation-Studies [2, 3, 4] FIRM [19] Tree-Shap [17]	SHAP [5] Lime [20] Gradient-Based-Localization [21] Relevance-Propagation [22] TreeExplainer [17]
Data	True-To-Data [18] MCI [6] UMFI [23]	

contradicts a previous result that defined the model-local setting. Due to space constraints, we provide supporting proofs for our claims in the Appendix, along with an extended version of Theorem 1 that allows a clear demarcation between local and global importance scores.

This work makes two key contributions: (1) We introduce a unified axiomatic framework that encompasses feature importance analysis in diverse settings, including global vs. local and model vs. data contexts. (2) We rigorously demonstrate inconsistencies within these settings, shedding light on disparities between global and local interpretations and between model-based and data-centric evaluations. These findings enhance our understanding of the nuances and challenges in the theory of feature importance analysis within machine learning interpretability.

2 Framework

We begin by introducing some notation: the setting consists of an input space $\mathcal{X}$, an output space $\mathcal{Y}$, so that given a pair $(x, y) \sim (\mathcal{X} \times \mathcal{Y})$, the learning task is to predict y by observing x. Without loss of generality, $\mathcal{X} \subseteq \mathbb{R}^{|\mathcal{F}|}$, where $\mathcal{F}$ is the set of available features. The explanation task is aimed to assign a score to each feature, based on its contribution to prediction. It consists of a two-step process: a **value function** is a function $\nu : \{x, \mathcal{X}\} \times 2^{\mathcal{F}} \to \mathbb{R}$ that assigns a scalar to each subset of features $\mathcal{S} \subseteq \mathcal{F}$, where $\nu(x, \mathcal{S})$ denotes the local value of a subset of features $\mathcal{S}$ for a given pair (x, y) and $\nu(\mathcal{X}, \mathcal{S})$ denotes the global value of this subset. A **feature importance function** is a function $\varphi : 2^{\mathcal{F}} \times \mathcal{F} \to \mathbb{R}$ that receives the output of a value function and assigns a feature importance score to each feature. For simplicity, we denote the importance of the feature f for both the global and the local importance functions as $\varphi(\nu(z), f)$ for $z \in \{x, \mathcal{X}\}$, where the actual input for the value function differs between them. An elaborated version of this notation appears in Section C.1 of the Appendix. For the data-model discussion, we add several notations: the data itself, $\mathcal{D}$, which is a probability measure over $\mathcal{X} \times \mathcal{Y}$; $\mathcal{M}$, which is a predictor over $\mathcal{D}$; and $\nu^{\mathcal{D}}, \nu^{\mathcal{M}}$ which are indicators for the current mode of evaluation in the value function.

We say that ν is a valid value function if it satisfies: $\nu(\mathcal{X}, \emptyset) = 0$, and in the global setting, we further require that monotonicity, i.e. for any subset $\mathcal{S} \subseteq T \Rightarrow \nu(\mathcal{X}, \mathcal{S}) \leq \nu(\mathcal{X}, T)$. This reflects the intuition that adding features to a model can not decrease the amount of information regarding the target variable, and thus can not decrease the prediction ability of a model. These two properties imply that $\forall \mathcal{S} \subseteq \mathcal{F}$, $\nu(\mathcal{X}, \mathcal{S}) \geq 0$ in the global setting. We do not assume these conditions generally hold for the local setting, implying that $\nu(x, \mathcal{S})$ may be negative for some $x \in \mathcal{X}$. Finally, $\mathbb{E}$ denotes the expected value function with respect to $\mathcal{D}$.

3 The Local-Global relation

It is natural to anticipate that a global phenomenon is an aggregate of local phenomena. This anticipated consistency can be illustrated intuitively: we find it confusing if a model that predicts loan repayment by lenders considers the age of the lenders to be a crucial factor in the global sense, yet at the same time declares that age is not a factor in predicting repayment for any specific lender. To avoid such scenarios, we require a small set of properties that ensure a meaningful relation between local and global settings in the framework of feature importance.

3.1 Expected properties

In this section, we formulate two consistency properties that we require to hold between the local and the global settings. We use these properties to prove the first inconsistency theorem.

Property 1 (Value Consistency). *ν is* Value Consistent *if*

$$\forall \mathcal{S} \subseteq \mathcal{F}, \quad \nu\left(\mathcal{X}, \mathcal{S}\right) = \mathbb{E}\left[\nu\left(x, \mathcal{S}\right)\right]$$

In property 1, to establish the relation between the local and the global value functions, the global value of each subset is constrained to be the expectancy taken over the inputs of the local value on this subset.

Property 2 (Importance Consistency). *A tuple $\{\nu, \varphi\}$ is* Importance Consistent *if*

$$\forall f \in \mathcal{F}, \quad \varphi(\nu(\mathcal{X}), f) = \mathbb{E}[\varphi(\nu(x), f)]$$

In property 2, to establish the local-global relations of the importance score, a consistency requirement analogous to the one above is made for the feature importance function: the global importance of a feature is the expected value of the local feature importance of this feature.

The two properties above define the expected relations between feature importance in local and global settings. We say that a tuple $\{\nu, \varphi\}$ is local-global consistent to denote that the *Value Consistency* and *Importance Consistency* properties hold.

3.2 The local-global inconsistency

We use the MCI function [6] to demonstrate the discrepancy between local and global settings. This function relies on a pre-defined set of properties which the importance score is expected to maintain. Apparently, the only function that satisfied these properties is the MCI function, defined as follows:

$$\text{MCI}(\nu, f) = \max_{\mathcal{S} \subseteq \mathcal{F} \setminus \{f\}} (\nu(\mathcal{S} \cup \{f\}) - \nu(\mathcal{S}))$$

Remarkably, the MCI score is the only function that uniquely satisfies the MCI properties, detailed in Section A.1. Our analysis leads us to demonstrate the following inconsistency:

Theorem 1. *properties 1,2, and MCI properties do not hold simultaneously.*

Proof sketch. Let $\{\nu, \varphi\}$ be local-global consistent tuple such that ν is non-decreasing. Assume that MCI properties hold, i.e. φ is the MCI function. From the local-global consistency, we get that $\forall f \in \mathcal{F}$:

$$MCI(\mathbb{E}[(\nu(x)], f) = MCI(\nu(\mathcal{X}), f) = \mathbb{E}[MCI(\nu(x), f)]$$

This leads to a contradiction since MCI uses the max operator and therefore is a non-linear function of the value function. A proof by counterexample is attached in Section B.1. $\square$

The proof sketch presented here is a simplified version, in which the local importance function and the global importance function are identical. A more detailed version of the proof, which does not assume that, can be found in Section C.

3.3 Discussion of the local-global relation

While the global-data setting is defined by Marginal Contribution Importance (MCI), the local-data setting (as the fourth quadrant in Table 2 demonstrates) is much harder to interpret and define. To tackle this issue, the approach adopted in this study was to use MCI's definition of global-data and define the local-global expected relation. However, this led to an inconsistency theorem. The source of inconsistency lies in the different considerations of ambiguous information: MCI ensures that meaningful information is not missed by attributing the maximum contribution to each feature, regardless of the contribution of other correlated features. This differs from methods such as SHAP [5], where contributions are split between correlated features.

4 The Data-Model relation

When explaining data, the focus is on understanding the underlying process generating them; while when explaining the model, the focus is on understanding how the model is making predictions based on the data. However, these settings are intertwined – models are often used as proxies by which nature can be explored. In cases where the model predictions are identical to the data, we expect conclusions reached from analyzing the model to hold with regard to the data. Therefore, we expect that the data and model will agree on each feature's importance. Nonetheless, this expected property implies a degenerate case where $\nu^{\mathcal{D}} \equiv 0$, which implies that for any importance function, the importance score of all features becomes zero, rendering them insignificant.

4.1 Expected properties

In this section, we formulate another consistency property, that expresses the expected relations between the data and the model settings. Then, we show that fulfillment of this property, along with known previous results, is only possible in a degenerate case.

Property 3 (Data-Model Consistency)**.** *Let $\mathcal{M}$ be a model that predicts over $\mathcal{D}$. A tuple $\{\mathcal{D}, \mathcal{M}, \nu, \varphi\}$ is Data-Model Consistent if $\forall x, y \sim \mathcal{D}, \mathcal{M}(x) = y$ and $\forall z \in \{x, \mathcal{X}\}$ it holds that*

$$\forall f \in \mathcal{F}, \quad \varphi(\nu^{\mathcal{D}}(z), f) = \varphi(\nu^{\mathcal{M}}(z), f)$$

The *Data-Model Consistency* property 3 states that if a model predicts the target perfectly, then the data and model importance scores of each feature are identical.

4.2 The data-model inconsistency

We use the SHAP function [5] to demonstrate the discrepancy between model and data settings. This function relies on a pre-defined set of properties which the importance score is expected to maintain. Apparently, the only function that satisfied these properties is the SHAP function, defined as follows:

$$\text{SHAP}(\nu, f) = \sum_{\mathcal{S} \subseteq \mathcal{F} \setminus \{f\}} \frac{|\mathcal{S}|!(|\mathcal{F}| - |\mathcal{S}| - 1)!}{|\mathcal{F}|!} \cdot (\nu(\mathcal{S} \cup \{f\}) - \nu(\mathcal{S}))$$

Notably, the SHAP score implies additional properties, detailed in Section A.2. This introduction leads us to the following inconsistency:

Theorem 2. *If a tuple $\{\mathcal{D}, \mathcal{M}, \nu, \varphi\}$ satisfies Data-Model Consistency (property 3) and SHAP properties, then $\nu^{\mathcal{D}} \equiv 0$.*

Our proof is based on a difference between models and the real world. Specifically, when the data contain correlated features, e.g. height measured in centimeters and inches, a model may learn based on only one of the features, resulting in different feature importance scores for each feature in the model. However, in the real world, both features are equally important. A detailed proof of the theorem is attached in Section B.2.

4.3 Discussion of the data-model relation

The need to link the data and model settings is not only theoretical. It is motivated by the need to use models to understand how the world works. Feature importance is often used, even if not stated explicitly, as a proxy for causal analysis. Unfortunately, the known limitations of trying to establish causal relations from observational data apply to feature importance too. The example we used to prove the inconsistency of the data-model often appears in real-world problems. Two features can be highly similar because a common, unobserved variable, caused them, or one of them caused the other. For example, when continuously measuring a variable of interest but only recording its mean and maximum values as observed variables. This problem of lacking information to disentangle the effect of two variables is known as unidentifiability.

To illustrate this in our context, consider two penalized regression models that are trained on two identical features. The first model employs an L1 regularization (lasso regression), and the second model employs an L2 regularization (ridge regression). The predictions of the two models are identical. However, assigning feature importance may lead to different results between the models - lasso regression will result in assigning all the importance to one of the features, whereas ridge regression will result in assigning equal importance to both features.

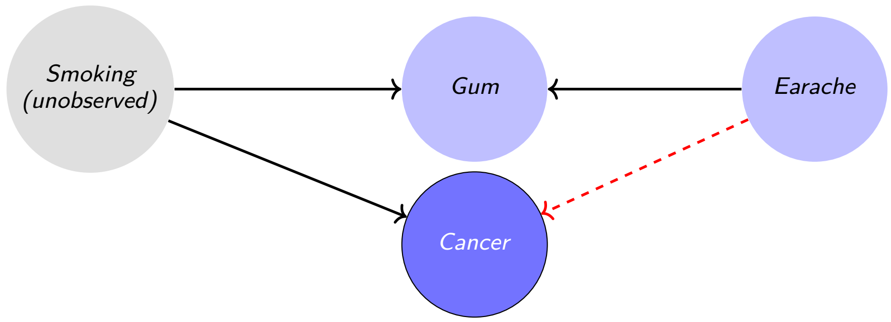

Figure 1: An example of a directed acyclic graph with a collider variable *Gum*.

Another situation that may lead to unexpected outcomes from feature importance scores is when a collider (also known as an inverted fork) exists in the data [15, 16]. For example consider the situation illustrated in Figure 1: Smoking cigarettes (*Smoking*) causes cancer (*Cancer*), but also increases chewing gum consumption (*Gum*). Assume also that doctors recommend people with earaches (*Earache*) to chew gum. Now, imagine scenarios in which a researcher is developing a model to predict *Cancer* using different subsets of the features *Gum* and *Earache*, but lacks information on *Smoking*. In the first scenario, the researcher uses only the *Earache* feature. Since earache and cancer are independent, any value-based feature importance score will assign zero importance to *Earache*. In the second scenario, where only the *Gum* feature is present, the researcher will conclude that *Gum* is an important feature since it is correlated with *Cancer*. In the third scenario, where a model that contains both *Earache* and *Gum* is considered, the researcher will infer that *Earache* has non-zero importance. This results from conditioning on *Gum*, creating an association between *Earache* and *Cancer* due to the presence of a collider. Intuitively, a person who chews gum and does not have an earache is more likely to be a smoker (notice that the smoking feature is unobserved), and hence at high risk of cancer. Therefore, the feature importance score might mislead a naïve researcher into thinking that earaches are predictive of cancer and that gum chewing is a cure for the disease.

The situations described here have been studied in the causality literature and there is no recipe for overcoming them that does not involve additional information about the world [15, 16].

5 Conclusion

In this work, we investigated the possibility to create a unified framework of feature importance scores, by defining their expected properties. Surprisingly, we found that it is impossible to define feature importance scores that are consistent between different settings. Specifically, the expected consistency between local and global scores contradicts properties of the data-global setting. Furthermore, there is no guarantee that feature importance scores of a model that perfectly predicts the data will reflect the feature importance of the data themselves.

Our inconsistency result is reminiscent of Kleinberg [24], which proves a similar result for clustering. Analogously, we do not argue that we have defined the only possible set of relevant properties for the various settings. We did, however, attempt to define a set of properties that we believe are essential. Yet, even these requirements led to inconsistencies. Future research can tackle which further assumptions can be made about feature importance scores, or other explainability methods, that are meaningful and yet can still be consistent.

In the meantime, our results show that feature importance scores should be used cautiously, aligning with recent research that has attempted to measure the quality and usefulness of explainability tools for different applications [25, 26, 27]. As such, our work tries to promote substantive discussions and accurate definitions of explainability, as previously advocated, for example, by Lipton [28] and Kumar et al. [29]. Hence, we hope that our work will contribute to stimulating additional research that will result in a solid theoretical foundation for explainable AI.

References

[1] Ian Covert, Scott Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, 33, 2020.

[2] VA Casagrande and IT Diamond. Ablation study of the superior colliculus in the tree shrew (tupaia glis). *Journal of Comparative Neurology*, 156(2):207–237, 1974.

[3] Eric Bengtson and Dan Roth. Understanding the value of features for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 294–303, Honolulu, Hawaii, October 2008. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/D08-1031`.

[4] Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[5] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.

[6] Amnon Catav, Boyang Fu, Yazeed Zoabi, Ahuva Libi Weiss Meilik, Noam Shomron, Jason Ernst, Sriram Sankararaman, and Ran Gilad-Bachrach. Marginal contribution feature importance-an axiomatic approach for explaining data. In *International Conference on Machine Learning*, pages 1324–1335. PMLR, 2021.

[7] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*, pages 126–137, 2015.

[8] Sarah Tan, Rich Caruana, Giles Hooker, Paul Koch, and Albert Gordo. Learning global additive explanations for neural nets using model distillation. *arXiv preprint arXiv:1801.08640*, 2018.

[9] Juan M Fontana, Muhammad Farooq, and Edward Sazonov. Estimation of feature importance for food intake detection based on random forests classification. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6756–6759. IEEE, 2013.

[10] Padideh Danaee, Reza Ghaeini, and David A Hendrix. A deep learning approach for cancer detection and relevant gene identification. In *Pacific symposium on biocomputing*, pages 219–229. World Scientific, 2017.

[11] Brett A McKinney, David M Reif, Marylyn D Ritchie, and Jason H Moore. Machine learning for detecting gene-gene interactions. *Applied bioinformatics*, 5(2):77–88, 2006.

[12] Kadri Haljas, Azmeraw T Amare, Behrooz Z Alizadeh, Yi-Hsiang Hsu, Thomas Mosley, Anne Newman, Joanne Murabito, Henning Tiemeier, Toshiko Tanaka, Cornelia Van Duijn, et al. Bivariate genome-wide association study of depressive symptoms with type 2 diabetes and quantitative glycemic traits. *Psychosomatic medicine*, 80(3):242, 2018.

[13] Bojan Karlaš, David Dao, Matteo Interlandi, Bo Li, Sebastian Schelter, Wentao Wu, and Ce Zhang. Data debugging with shapley importance over end-to-end machine learning pipelines. *arXiv preprint arXiv:2204.11131*, 2022.

[14] Wenbo Guo, Dongliang Mu, Jun Xu, Purui Su, Gang Wang, and Xinyu Xing. Lemna: Explaining deep learning based security applications. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 364–379, 2018.

[15] Judea Pearl. *Causality*. Cambridge university press, 2009.

[16] Miguel A Hernán and James M Robins. Causal inference, 2010.

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

[18] Hugh Chen, Joseph D Janizek, Scott Lundberg, and Su-In Lee. True to the model or true to the data? *arXiv preprint arXiv:2006.16234*, 2020.

[19] Alexander Zien, Nicole Krämer, Sören Sonnenburg, and Gunnar Rätsch. The feature importance ranking measure. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 694–709. Springer, 2009.

[20] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " 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, 2016.

[21] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[22] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[23] Joseph Janssen and Vincent Guan. Ultra marginal feature importance. *arXiv preprint arXiv:2204.09938*, 2022.

[24] Jon Kleinberg. An impossibility theorem for clustering. *Advances in neural information processing systems*, 15, 2002.

[25] Cynthia Rudin and Joanna Radin. Why are we using black box models in ai when we don't need to? a lesson from an explainable ai competition. 2019.

[26] Chun-Hao Chang, Sarah Tan, Ben Lengerich, Anna Goldenberg, and Rich Caruana. How interpretable and trustworthy are gams? In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 95–105, 2021.

[27] Giorgio Visani, Enrico Bagli, Federico Chesani, Alessandro Poluzzi, and Davide Capuzzo. Statistical stability indices for lime: Obtaining reliable explanations for machine learning models. *Journal of the Operational Research Society*, 73(1):91–101, 2022.

[28] Zachary C Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018.

[29] I Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning*, pages 5491–5500. PMLR, 2020.

[30] Eric J Friedman. Paths and consistency in additive cost sharing. *International Journal of Game Theory*, 32(4):501–518, 2004.

[31] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

A Additional Properties

In this section, we present the additional properties mentioned in the local-global section and the data-model section.

A.1 MCI properties

Catav et al. [6] introduced the following three properties to define the expected behavior of feature importance scores in the global-data setting:

Property 4 (Marginal Contribution)**.** *A tuple $\{\nu, \varphi\}$ satisfies the Marginal Contribution property when the importance of a feature is equal to or higher than the increase in the value function when adding it to all the other features, i.e. $\varphi(\nu(\mathcal{X}, f)) \geq \nu(\mathcal{X}/, \mathcal{F}) - \nu(\mathcal{X}, \mathcal{F} \setminus f))$.*

The *Marginal Contribution* property states that the importance of a feature is at least its contribution to the value function when adding the latter to the set of all other features.

Property 5 (Elimination)**.** *A tuple $\{\nu, \varphi\}$ satisfies the Elimination property when eliminating features from $\mathcal{F}$ can only decrease the importance of each feature. i.e., if $T \subseteq \mathcal{F}$ and $\bar{\nu}$ is the value function which is obtained by eliminating T from $\mathcal{F}$ then*

$$\forall f \in \mathcal{F} \setminus T, \quad \varphi(\nu(\mathcal{X}, f)) \geq \varphi(\bar{\nu}(\mathcal{X}, f))$$

The *Elimination* property states that the importance of a feature does not become smaller when other features are removed from the calculation. The process of elimination is defined as follows:

Definition. ***(Elimination operation)*** *Let $\mathcal{F}$ be a set of features and ν be a value function. Eliminating the set $T \subset \mathcal{F}$ creates a new set of features $\mathcal{F}' = \mathcal{F} \setminus T$ and a new value function $\nu' : 2^{\mathcal{F}'} \to \mathbb{R}$ such that*

$$\forall \mathcal{S} \subseteq \mathcal{F}', \quad \nu'(\mathcal{S}) = \nu(\mathcal{S})$$

Property 6 (Minimalism)**.** *A tuple $\{\nu, \varphi\}$ satisfies the Minimalism property when for every function $\bar{\varphi} : \mathbb{R}^{2^{\mathcal{F}}} \to \mathbb{R}^{\mathcal{F}}$ for which properties 4 and 5 hold, then*

$$\forall f \in \mathcal{F}, \quad \varphi(\nu(\mathcal{X}, f)) \leq \bar{\varphi}(\nu(\mathcal{X}, f))$$

The *Minimalism* property states that among all the functions that satisfy properties 4 and 5, the feature importance scoring function should be minimal.

Using these properties, Catav et al. [6] prove the following:

Theorem 3. *The MCI feature importance score (see Table 1) is the only score for which the Marginal Contribution property, the Elimination property, and the Minimalism property (Properties 4,5,6) hold simultaneously.*

A.2 SHAP properties

The following properties stem naturally from the SHAP function, which is the only function that satisfies the SHAP properties proposed in Lundberg and Lee [5]:

Property 7 (Triviality)**.** *A tuple $\{\nu, \varphi\}$ satisfies the Triviality Property if the following conditions hold:*

1. *For all $\mathcal{S} \subseteq \mathcal{F}$, if $\nu(x, \mathcal{S}) \neq 0$, then there exists a feature $f \in \mathcal{S}$ such that $\varphi(\nu(x), f) \neq 0$.*
2. *If $\varphi(\nu(x), f) \neq 0$, then there exists a subset $\mathcal{S} \subseteq \mathcal{F}$ such that $\nu(x, \mathcal{S} \cup \{f\}) \neq \nu(x, \mathcal{S})$.*

The Triviality Property establishes a non-trivial relationship between the value and the importance functions. It requires that if a subset of features has any value, it will be reflected in the importance of at least one feature from this subset. Conversely, it demands that if any feature is important (i.e., has non-zero importance), it must be included in some valuable subset. Notably, if a feature f satisfies $\nu(x, \mathcal{S}) = \nu(x, \mathcal{S} \cup \{f\})$ for any subset of features, then f has zero importance.

Property 8 (Dummy Feature)**.** *Let $\mathcal{M}$ be a model that predicts over $\mathcal{D}$. A tuple $\{\nu, \varphi\}$ satisfies the Dummy Feature Property if, for all $f \in \mathcal{F}$ and for all $x, x' \in \mathcal{X}$ such that x differs from x' only by the f'th feature $\mathcal{M}(x) = \mathcal{M}(x')$, then*

$$\varphi(\nu^{\mathcal{M}}, f) = 0$$

The Dummy Feature Property implies that if changing the value of a feature has no effect on a model's output, then the importance of that feature is zero. This property also had been recognized in previous works such as Friedman [30] and Sundararajan et al. [31].

B Inconsistencies Proofs

In this section, we present proofs for the inconsistency theorems.

B.1 Theorem 1

Let $\{\nu, \varphi\}$ be local-global consistent tuple such that ν is non-decreasing. Assume that MCI properties hold, i.e. φ is the MCI function. From the local-global consistency, we get that $\forall f \in \mathcal{F}$:

$$MCI(\mathbb{E}[(\nu(x)], f) = MCI(\nu(\mathcal{X}), f) = \mathbb{E}[MCI(\nu(x), f)]$$

This leads to a contradiction since MCI uses the max operator, and therefore is a non-linear function of the value function.

Now, we aim to demonstrate, by way of a counter-example, that the MCI function is not linear. This will lead to a contradiction between the properties of the data-global, as defined in [6] setting and $\{\nu, \varphi\}$ Consistency properties. Formally, we contradict the following equality:

For any ν which is a valid value function,

$$\alpha \cdot MCI(\nu(x_0)) + (1 - \alpha) \cdot MCI(\nu(x_1)) = MCI(\alpha \cdot \nu(x_0) + (1 - \alpha) \cdot \nu(x_1)) \tag{1}$$

Counter-example. Let $\mathcal{X}$ be a dataset consisting of two samples: x_0 and x_1, over the feature space $\mathcal{F} = \{f_0, f_1\}$. We define the value function ν as follows:

$$\nu = \begin{pmatrix} & x_0 & x_1 & \mathcal{X} \\ \{\emptyset\}: & 0 & 0 & 0 \\ \{f_0\}: & 0 & 1 & 1.5 \\ \{f_1\}: & 1 & 1 & 1 \\ \{f_0, f_1\}: & 2 & 1 & 1.5 \end{pmatrix}$$

Now, let $\alpha = \frac{1}{2}$. We will evaluate the left-hand side of equation (1) and the right-hand side separately.

Left-hand side evaluation:

$$\begin{aligned} &\alpha \cdot MCI(\nu(x_0)) + (1 - \alpha) \cdot MCI(\nu(x_1)) \\ &= \frac{1}{2} \cdot MCI\left(\begin{pmatrix} 0 \\ 0 \\ 1 \\ 2 \end{pmatrix}\right) + \frac{1}{2} \cdot MCI\left(\begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \end{pmatrix}\right) \\ &= \frac{1}{2} \cdot \begin{pmatrix} 1 \\ 1.5 \end{pmatrix} + \frac{1}{2} \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1.25 \end{pmatrix} \end{aligned}$$

Right-hand side evaluation:

$$\begin{aligned} &MCI\left(\alpha \cdot \nu(x_0) + (1 - \alpha) \cdot \nu(x_1)\right) \\ &= MCI\left(\nu(\mathcal{X})\right) \\ &= MCI\left(\begin{pmatrix} 0 \\ 0.5 \\ 1 \\ 1.5 \end{pmatrix}\right) = \begin{pmatrix} 0.5 \\ 1 \end{pmatrix} \end{aligned}$$

Hence, we have found a counter-example for equation (1), which contradicts the claimed linearity of the MCI function. This concludes the proof. □

B.2 Theorem 2

We establish that any tuple $\{\nu, \varphi\}$ satisfying the Triviality, Dummy-Feature, and Data-Model Consistency properties (Properties 7, 8, 3) inevitably encounters a scenario where $\nu^{\mathcal{D}} \equiv 0$. This scenario implies that for any importance function that considers the value, the importance score of all features becomes zero, rendering them insignificant.

Proof. Let $\mathcal{D}$ be a probability measure over $\mathcal{X}$ such that its feature space contains two duplicate features of a random variable, which solely dictate the target. Formally, $\rho \in [0,1]$, $\{f_0, f_1\} \subseteq \mathcal{F},$ and $\forall x \in \mathcal{X}, \quad f_0(x) = f_1(x) = \rho$. The target is defined as $\mathcal{D}(x) = h(\rho)$, where h is some function of ρ. Let $\mathcal{M}_0, \mathcal{M}_1$ be two models s.t each model focuses on one feature and neglects the other:

$$\forall i \in \{0,1\}, \quad \mathcal{M}_i(x) = h(f_i(x))$$

Let the tuple $\{\nu, \varphi\}$ satisfy Triviality, Dummy-Feature and Data-Model Consistency (properties 7, 8, 3). By definition, $\mathcal{M}_0, \mathcal{M}_1$ predict the data perfectly, and therefore by the *Data-Model Consistency*, it holds that

$$\forall i \in \{0,1\} \text{ and } \forall f \in \mathcal{F}, \quad \varphi(\nu^{\mathcal{D}}, f) = \varphi(\nu^{\mathcal{M}_i}, f)$$

The *Dummy Feature* Axiom implies that

$$\forall i \in \{0,1\}, \quad \varphi\left(\nu^{\mathcal{M}_i}, f_{1-i}\right) = 0$$

Combining the last two implies that

$$\forall f \in \mathcal{F}, \quad \varphi\left(\nu^{\mathcal{D}}, f\right) = 0$$

By the *Triviality* Axiom, the only value function that satisfies the above is $\nu^{\mathcal{D}} \equiv 0$. □

C Distinguish between the local and global importance functions

In this section, we will reformulate our properties to distinguish between the local and global importance functions.

C.1 Framework

Before proceeding, we will provide a more detailed definition of the value and importance functions to ensure precision in describing these functions:

A value function is represented as $\nu : (\mathcal{D} \times \{x, \mathcal{X}\} \times 2^{\mathcal{F}}) \to \mathbb{R}$. It assigns a scalar to each subset of features $\mathcal{S} \subseteq \mathcal{F}$. Here, $\nu(\mathcal{D}, x, \mathcal{S})$ signifies the value of a feature subset $\mathcal{S}$ for the local instance x, drawn from a probability measure $\mathcal{D}$. Additionally, $\nu(\mathcal{D}, \mathcal{X}, \mathcal{S})$ represents the value of the same feature subset over the entire sample space.

On the other hand, a feature importance function is denoted as $\varphi : (\{\text{local}, \text{global}\} \times 2^{\mathcal{F}}) \to \mathbb{R}^{\mathcal{F}}$. It takes an indicator specifying whether it operates in the local or global context and the output of the value function. This function assigns feature importance scores to individual features. Specifically, $\varphi(\text{local}, \nu(\mathcal{D}, x), f)$ indicates the importance of feature f for the instance x, while $\varphi(\text{global}, \nu(\mathcal{D}, \mathcal{X}), f)$ signifies the importance of the same feature across the entire sample space.

Note that the monotonicity property of ν holds in the global setting, but not necessarily in the local setting. For example, consider the case where the prediction target is whether a person has cancer and one of the features is whether the person carries a lighter in their pocket. This feature may be globally important, since it may correlate with smoking. However, it is possible that some people carry lighters but do not smoke, in which case this feature might lead to an erroneous prediction and hence has a negative local contribution. Globally admissible denotes a case where an instance x has only a non-negative contribution. Formally, $\nu(\mathcal{D}, x)$ is globally admissible if $\nu(\mathcal{D}, x)$ is monotonic non-decreasing and $\nu(\mathcal{D}, x, \emptyset) = 0$.

Figure 2: **consistency diagram:** In local-global consistency the global value is the expectation of the local values, while the global importance is the expected value of local importances.

$$\begin{tikzcd}
\nu(x) \arrow[r] \arrow[d, "\mathbb{E}"] & \varphi(x, \nu(x)) \arrow[d, "\mathbb{E}"] \\
\nu(\mathcal{X}) \arrow[r] & \varphi(\mathcal{X}, \nu(\mathcal{X}))
\end{tikzcd}$$

C.2 Expected properties

Property 9 (Value Consistency)**.** *ν is* Value Consistent *if for every $\mathcal{D}$*

1. $\forall \mathcal{S} \subseteq \mathcal{F}, \quad \nu\left(\mathcal{D}, \mathcal{X}, S\right) = \mathbb{E}\left[\nu\left(\mathcal{D}, x, \mathcal{S}\right)\right]$
2. *$\exists x^*$ and $\exists \mathcal{D}^*$ such that $\mathcal{D}^*$ is a Dirac measure and $\nu(\mathcal{D}^*, x^*) = \nu(\mathcal{D}, x)$*

To establish the relation between the local and the global value functions, two complementary conditions are required: First, the global value of each subset is constrained to be the expectancy taken over the inputs of the local value on this subset. Second, the local value is constrained to be able to realize the global value.

Property 10 (Importance Consistency)**.** *A tuple $\{\nu, \varphi\}$ is* Importance Consistent *for every $\mathcal{D}$*

1. $\forall f \in \mathcal{F}, \quad \varphi\left(\textit{global}, \nu(\mathcal{D}, \mathcal{X}), f\right) = \mathbb{E}\left[\varphi\left(\textit{local}, \nu\left(\mathcal{D}, x\right), f\right)\right]$
2. *ν is Value Consistent.*

To establish the local-global relations of the importance score, a consistency requirement analogous to the one above is made for the feature importance function: The global importance of a feature is the expected value of the local feature importance of this feature.

Consistency implies a commutative diagram, which is presented in Figure 2.

C.3 Detailed proof for Theorem 1

The proof uses the following two lemmas. Combining these lemmas implies that φ is a linear function, and the rest of the proof is identical to the abbreviated version that appears above.

Lemma 1. *Let $\mathcal{D}$ be a probability measure over $\mathcal{X}$. If $\{\nu, \varphi\}$ is importance consistent and $x \in \mathcal{X}$ such that $\nu(\mathcal{D}, x)$ is globally admissible, then*

$$\varphi(\textit{global}, \nu(\mathcal{D}, x)) = \varphi(\textit{local}, \nu(\mathcal{D}, x))$$

Proof. [of Lemma 1] Let $\{\nu, \varphi\}$ be important consistent tuple. Let $\mathcal{D}$ be a probability measure over $\mathcal{X}$ and let $x \in \mathcal{X}$ be a globally admissible instance. Denote $\mathcal{D}'$ as the corresponding probability measure, i.e $\nu(\mathcal{D}, x) = \nu(\mathcal{D}', \mathcal{X}')$. By the *Value Consistency* property there exist a Dirac measure $\mathcal{D}^*$ such that $\nu(\mathcal{D}', \mathcal{X}') = \nu(\mathcal{D}^*, x^*)$. Hence,

$$\begin{aligned}
\varphi\left(\text{global}, \nu(\mathcal{D}, x)\right) &= \varphi\left(\text{global}, \nu(\mathcal{D}', \mathcal{X}')\right) && (2)\\
&= \varphi\left(\text{global}, \nu(\mathcal{D}^*, \mathcal{X}^*)\right) && (3)\\
&= \varphi\left(\text{local}, \nu(\mathcal{D}^*, x^*)\right) && (4)\\
&= \varphi\left(\text{local}, \nu(\mathcal{D}', x')\right) && (5)\\
&= \varphi\left(\text{local}, \nu(\mathcal{D}, x)\right) && (6)
\end{aligned}$$

where (3) is from the global admissibility of $\nu(\mathcal{D}, x)$, (4) follows from the consistency and from the fact that $\mathcal{D}^*$ is Dirac, and the following equations follow from the definition of $\mathcal{D}'$. ☐

Lemma 2. *Let $\mathcal{D}$ be a probability measure over $\mathcal{X}$ and let ν be a monotonic non-decreasing value function (i.e. $\forall x \in \mathcal{X}, \quad \nu(\mathcal{D}, x)$ is globally admissible). If $\{\nu, \varphi\}$ is local-global consistent then*

$$\varphi\left(\textit{global}, \mathbb{E}[(\nu(\mathcal{D}, x)])\right) = \mathbb{E}[\varphi\left(\textit{global}, \nu(\mathcal{D}, x)\right)]$$

Proof. [of Lemma 2] Let $\{\nu, \varphi\}$ be a local-global consistent tuple and let $\mathcal{D}$ be such that $\nu(\mathcal{D}, x)$ is globally admissible for every x in the support of $\mathcal{D}$. Therefore,

$$
\begin{aligned}
&\mathbb{E}[\varphi\left(\text{global}, \nu(\mathcal{D}, x)\right))] && (7)\\
&= \mathbb{E}[\varphi\left(\text{local}, \nu(\mathcal{D}, x)\right))] && (8)\\
&= \varphi\left(\text{global}, \nu(\mathcal{D}, \mathcal{X})\right) && (9)\\
&= \varphi\left(\text{global}, \mathbb{E}[\nu(\mathcal{D}, x)]\right) && (10)
\end{aligned}
$$

where (8) is valid by Lemma 1,and (9) and (10) are valid by the importance consistency 2. □