Sampling, Intervention, Prediction, Aggregation: A Generalized Framework for Model Agnostic Interpretations

Christian A. Scholbeck (✉), Christoph Molnar, Christian Heumann, Bernd Bischl, Giuseppe Casalicchio

Department of Statistics, Ludwig-Maximilians-University Munich,
Ludwigstr. 33, 80539 Munich, Germany
christian.scholbeck@stat.uni-muenchen.de

Abstract. Non-linear machine learning models often trade off a great predictive performance for a lack of interpretability. However, model agnostic interpretation techniques now allow us to estimate the effect and importance of features for any predictive model. Different notations and terminology have complicated their understanding and how they are related. A unified view on these methods has been missing. We present the generalized SIPA (Sampling, Intervention, Prediction, Aggregation) framework of work stages for model agnostic interpretation techniques and demonstrate how several prominent methods for feature effects can be embedded into the proposed framework. We also formally introduce pre-existing marginal effects to describe feature effects for black box models. Furthermore, we extend the framework to feature importance computations by pointing out how variance-based and performance-based importance measures are based on the same work stages. The generalized framework may serve as a guideline to conduct model agnostic interpretations in machine learning.

Keywords: Interpretable Machine Learning | Explainable AI | Feature Effect | Feature Importance | Marginal Effects | Partial Dependence | ALE | Permutation Feature Importance | Model Agnostic | Framework

1 Introduction and Related Work

There has been an ongoing debate about the lacking interpretability of machine learning (ML) models. As a result, researchers have put in great efforts to develop techniques for creating insights into the workings of predictive black box models. The novel field of interpretable machine learning (IML) [19] serves as an umbrella term for all interpretation methods in machine learning research. They may be distinguished in different ways:

(i) *Feature effects or feature importance*: Predictive models are interpreted with two main goals in mind. First, feature effects indicate the direction and magnitude of change in the predicted outcome when a feature value

changes. Prominent methods include the individual conditional expectation (ICE) [10] & partial dependence (PD) [9], accumulated local effects (ALE) [1] and Shapley values [23]. Second, the feature importance is defined as the contribution of a feature to the predictive performance of the fitted model. This includes variance-based measures like the feature importance ranking measure (FIRM) [11,24] and performance-based measures like the permutation feature importance (PFI) [8], individual conditional importance (ICI) and partial importance (PI) curves [4], as well as the Shapley feature importance (SFIMP) [4]. Input gradients were proposed by [12] as a model agnostic tool for both effects and importance that essentially equals marginal effects (ME) [16], which have a long tradition in statistics. They also define an average input gradient which corresponds to the average marginal effect (AME).

(ii) *Intrinsic or post hoc interpretability*: Linear models (LM), generalized linear models (GLM), classification and regression trees (CART) or rule lists [21] are examples for intrinsically interpretable models, while random forests (RF), support vector machines (SVM), neural networks (NN) or gradient boosting (GB) models can only be interpreted post hoc. Here, the interpretation process is detached from and takes place after the model fitting process. e.g. with ICE & PD plots or ALEs.

(iii) *Model specific or model agnostic interpretations*: Interpreting model coefficients of GLMs or deriving a decision rule from a classification tree is a model specific interpretation. Model agnostic methods like ICE & PD plots or ALEs can be applied to any model and are suited for model comparisons.

(iv) *Local or global explanations*: Local explanations like ICE curves evaluate the model behavior when predicting the target variable for one specific observation. Global explanations like the PD curve interpret the model for the entire input space. Furthermore, it is possible to explain model predictions for a group of observations, e.g. on intervals. In a lot of cases, local and global explanations can be transformed into one another via (dis-)aggregation, e.g. with ICE & PD curves.

Motivation: Different terminologies and the variety of available tools complicate the understanding, discussion and research of interpretation techniques in ML. It turns out that deconstructing model agnostic methods into sequential work stages reveals striking similarities. This might not be surprising, considering that the motive for using any model agnostic interpretation technique is to explain the prediction of black box models. Gaining insights into the workings of complex and non-linear prediction functions seems impossible. Instead, all model agnostic techniques work according to the same principle, i.e. they evaluate the model predictions when inputs are changing. However, a unified view on these methods has been missing so far. The motivation for this research paper is to establish a common terminology for model agnostic interpretation methods and to reveal similarities in their computation.

Contributions: In section 3, we formally introduce marginal effects (ME) from statistics as a model agnostic interpretation tool for ML. In section 4, we present

the generalized SIPA (Sampling, Intervention, Prediction, Aggregation) framework of work stages for model agnostic techniques as our main contribution. We proceed to demonstrate how several methods to estimate feature effects (MEs, ICE & PD, ALEs and the Shapley value) can be embedded into the proposed framework. Furthermore, in section 5 and 6 we extend the framework to the feature importance by pointing out how variance-based (FIRM) and performance-based (ICI & PI, PFI and SFIMP) importance measures are based on the same work stages. The SIPA framework reduces the diverse set of available techniques to a single methodology. It may therefore establish a common ground to discuss model agnostic interpretations in terms of notation and terminology and inspire the development of novel methods in IML.

2 Notation and Preliminaries

Assume a p-dimensional feature space $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_p$ with index set $P = 1, \ldots, p$ and a target space $\mathcal{Y}$. We assume an unknown functional relationship f between $\mathcal{X}$ and $\mathcal{Y}$ plus a random error, i.e. $f(\mathcal{X}) + \varepsilon = \mathcal{Y}$. A supervised learning model $\hat{f}$ learns relationships from an i.i.d. training sample with m observations that was drawn from the joint space $\mathcal{X} \times \mathcal{Y}$. The corresponding random variables from the feature space are denoted by $X_1, \ldots, X_p$. The random variable from the target space is denoted by Y. The vector $x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_p^{(i)})^\top$ corresponds to the i-th observation that is associated with the observed target value $y^{(i)} \in \mathcal{Y}$. The vector $x_j = (x_j^{(1)}, x_j^{(2)}, \ldots, x_j^{(m)})^\top$ represents the realized values of X_j. Every feature is assigned a unique index value j with $j \in P$. We apply a variety of model agnostic interpretation techniques to a subset of features with index set S ($S \leq P$). The complement C denotes unselected features ($C = P \setminus S$). If not denoted otherwise, S contains exactly one element. The corresponding random variables and observed feature values are denoted by X_S, X_C and x_S, x_C respectively. The generalization error $GE(\hat{f}, \mathcal{T})$ is measured by the loss function $\mathcal{L}$ on unknown test data $\mathcal{T}$ and estimated with a sample of test data $\mathcal{D}$.

$$GE(\hat{f}, \mathcal{T}) = \mathbb{E}\left[\mathcal{L}(\hat{f}(X), Y)\right]$$
$$\widehat{GE}(\hat{f}, \mathcal{D}) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\hat{f}(x^{(i)}), y^{(i)})$$

The partial derivative of the trained model $\hat{f}(x_S, x_C)$ with respect to x_S at point $x_S = x_0$ is numerically approximated with a symmetric difference quotient [16].

$$\lim_{h \to 0} \frac{\hat{f}(x_0 + h, x_C) - \hat{f}(x_0, x_C)}{h} \approx \frac{\hat{f}(x_0 + h, x_C) - \hat{f}(x_0 - h, x_C)}{2h}, \quad h > 0$$

A term of the form $\hat{f}(x_0+h, x_C) - \hat{f}(x_0-h, x_C)$ on the interval $x_S \in [x_0-h, x_0+h]$ is called a finite difference (FD) with respect to x_S at point x_0.

$$FD_S(x_0, x_C) = \hat{f}(x_0 + h, x_C) - \hat{f}(x_0 - h, x_C)$$

The step size h needs to be as small as possible to approximate the differential quotient, but too small values of h will result in cancelation errors.

The prediction function $\hat{f}(X)$ can be decomposed into a sum of lower-dimensional terms. A functional decomposition of $\hat{f}(X)$ is defined as:

$$\hat{f}(X) = g_0 + \sum_{i=1}^{p} g_i(X_i) + \sum_{j<k} g_{jk}(X_j, X_k) + \sum_{j<k<l} g_{jkl}(X_j, X_k, X_l) + \ldots + g_{12\ldots p}(X_1, X_2, X_3, \ldots, X_p)$$

There is a constant term g_0. The functions $g_1(X_1), \ldots, g_p(X_p)$ can be regarded as main effects (first order effects), the functions $g_{jk}(X_j, X_k), j \neq k$ as two-way interactions between the random variables X_j and X_k (second order effects), the functions $g_{jkl}(X_j, X_k, X_l), j < k < l$ as three-way interactions between X_j, X_k and X_l (third order effects), continuing up to the p-th order. The functional decomposition of $\hat{f}(X)$ is only unique if we assign additional constraints to its components, e.g. orthogonality constraints and zero means (see [1], [13], [14] or [22]).

3 Feature Effects

Partial dependence (PD) & individual conditional expectation (ICE): First suggested by [9], the PD is defined as the dependence of the true model $f(X_S, X_C)$ on one or multiple selected features X_S after all remaining features X_C have been marginalized out [10].

$$PD(X_S) = \mathbb{E}_{X_c}\left[f(X_S, X_C)\right] = \int f(X_S, X_C)\, dP(X_C) \tag{1}$$

Every subset of features S has its own PD function that returns the expected value of the target variable for multiple values of X_S while the vector of complementary features X_C varies over its marginal distribution $dP(X_C)$. Unfortunately, neither the true model, nor the marginal distribution $dP(X_C)$ are usually known. We are using the fitted model as an approximation to the real model and the sampling distribution as a substitute for the marginal distribution. The PD is estimated via Monte Carlo integration.

$$\widehat{PD}(x_S) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}(x_S, x_C^{(i)})$$

Its graphical visualization is called the partial dependence plot (PDP). [9] suggests using the PD as a useful feature effect measure when features are not

interacting. Otherwise it can obfuscate the relationships in the data [4]. [10] instead propose the individual conditional expectation (ICE). The i-th ICE curve corresponds to the expected value of the target for the i-th observation as a function of x_S, conditional on the observed vector $x_C^{(i)}$.

$$\widehat{ICE}^{(i)}(x_S) = \hat{f}(x_S, x_C^{(i)})$$

ICE curves disaggregate the global effect estimates of the PD curve to local effect estimates of single observations. ICE and PD suffer from extrapolation when features are correlated, because the permutations used to predict are located in regions without any training data [1].

Accumulated local effects (ALE): [1] develop ALEs as a feature effect measure for correlated features that does not extrapolate. The idea of ALEs is to take the integral with respect to X_S of the first derivative of the prediction function with respect to X_S. This creates an accumulated partial effect of X_S on the target variable while simultaneously blocking out effects of other features. The main advantage of not extrapolating stems from integrating with respect to the conditional distribution of X_C instead of the marginal distribution of X_C [1]. Consider a single feature S. The minimum of the distribution of the random variable X_S is denoted by Z_0, and the minimum of the observed marginal distribution x_S by z_0. The corresponding maximum values are denoted by Z_k and z_k, respectively. The true first order ALE is defined as:

$$\begin{aligned} ALE(X_S) &= \int_{Z_0}^{X_S} \mathbb{E}_{X_C|X_S} \left[\frac{\partial \hat{f}(X_S, X_C)}{\partial X_S} \middle| X_S = z_S \right] dz_S - constant \\ &= \int_{Z_0}^{X_S} \left[\int \mathcal{P}(X_C|z_S) \frac{\partial \hat{f}(z_S, X_C)}{\partial z_S} dX_C \right] dz_S - constant \end{aligned} \tag{2}$$

A constant is substracted in order to center the plot. We estimate the first order ALE for the whole value range of x_S in three steps:

(i) The continuous function of the partial derivative $\frac{\partial f(X_S, X_C)}{\partial X_S}$ is generally unknown and has to be estimated interval-wise. It is approximated by computing finite differences (FD) on a suitable discretization of the observed value range $[min(x_S), max(x_S)]$ into the set of interval boundaries $\{z_0, z_1, \ldots, z_j, \ldots, z_k\}$. The estimation at point $x_S = z_j$ represents the approximation within the interval $[z_{j-1}, z_j]$. The local effect within the interval $[z_{j-1}, z_j]$ corresponds to the integral of the partial derivative from z_{j-1} to z_j. As we will integrate the partial derivative with respect to the same variable, dividing FDs by interval widths is not necessary.

(ii) The conditional expected value $\mathbb{E}_{X_C|X_S} \left[\frac{\partial \hat{f}(X_S, X_C)}{\partial X_S} \middle| X_S = z_j \right]$, i.e. the expectation with respect to the marginal distribution of X_C, conditional on $X_S = z_j$, is estimated via Monte Carlo integration within the preceding interval $[z_{j-1}, z_j]$. This replaces the inner integral in eq. (2).

(iii) The accumulation of all estimated local effects up to $x_S = z_K$ replaces the outer integral in eq. (2), i.e. the estimated local effects within the intervals $\{[z_0, z_1], [z_1, z_2], \ldots, [z_{j-1}, z_j], \ldots, [z_{k-1}, z_k]\}$ are summed up.

[1] suggests partitioning the value range of x_S according to the quantiles of the sampling distribution. The number of quantiles is determined by a manually specified number of k intervals. Regions with a large concentration of observations receive a smaller partitioning than regions with a low concentration of observations. The partial derivative is approximated by first taking FDs of predictions within each interval. For each i-th observation within the interval $[z_{j-1}, z_j]$, the observational value $x_S^{(i)}$ is substituted by the right interval boundary z_j and by the left one z_{j-1}. This corresponds to an FD of predictions with a forward step h_{forw} and a backward step h_{backw}.

$$\begin{aligned} FD_S(x_S^{(i)}, x_C^{(i)}) &= \hat{f}(x_S^{(i)} + h_{forw}^{(i)}, x_C^{(i)}) - \hat{f}(x_S^{(i)} - h_{backw}^{(i)}, x_C^{(i)}) \\ &= \hat{f}(z_j, x_C^{(i)}) - \hat{f}(z_{j-1}, x_C^{(i)}) \end{aligned}$$

The second order ALE is the extension of the first order ALE to a bivariate X_S. It is important to note that first order effect estimates are substracted from the second order estimates. [1] further lays out the computations necessary for higher order ALEs.

The important property of ALEs to block out effects of other features stems from using FDs. [1] calls this additive unbiasedness because additively linked effects of other features in the prediction function are blocked out. Consider a prediction function with a sole main effect of feature S, denoted by $g_S(X_S)$, and arbitrary effects of other features. The FD of predictions with observed values x within the interval $[z_{j-1}, z_j]$ results in the FD of $g_S(x_S)$ only, which further leads to an estimated ALE of the main effect only:

$$\begin{aligned} &\hat{f}(z_j, x_C) - \hat{f}(z_{j-1}, x_C) \\ &= (g_0 + g_S(z_j) + \sum_{j \in C} g_j(x_j) + \sum_{i \neq j} g_{ij}(x_i, x_j) + \ldots) \\ &\quad - (g_0 + g_S(z_{j-1}) + \sum_{j \in C} g_j(x_j) + \sum_{i \neq j} g_{ij}(x_i, x_j) \ldots) \\ &= g_S(z_j) - g_S(z_{j-1}) \end{aligned}$$

Marginal effects (ME): MEs are an established technique in statistics and commonly applied in econometrics [15]. They are often used to interpret nonlinear functions of coefficients in GLMs like logistic regression by approximating their first derivatives. We are extending this concept to any differentiable prediction function. Although there is extensive literature on MEs, this concept was suggested by [12] as a novel method for ML and referred to as the input gradient. Consider a prediction function with a constant, as well as first and second order

effects of two features. Taking the FD of predictions in the nominator blocks out the constant and additively linked main effect of the second feature in eq. (3).

$$\begin{aligned}
&\frac{\hat{f}(x_1 + h, x_2) - \hat{f}(x_1 - h, x_2)}{2h} \\
&= \frac{[g_0 + g_1(x_1 + h) + g_2(x_2) + g_{12}(x_1 + h, x_2)]}{2h} \\
&\quad - \frac{[g_0 + g_1(x_1 - h) + g_2(x_2) + g_{12}(x_1 - h, x_2)]}{2h} \\
&= \frac{g_1(x_1 + h) - g_1(x_1 - h)}{2h} + \frac{g_{12}(x_1 + h, x_2) - g_{12}(x_1 - h, x_2)}{2h}
\end{aligned} \tag{3}$$

MEs are defined locally and the feature space usually is multi-dimensional. There are three established variations of how to specify the locations of derivatives in the feature space [16].

Average marginal effects (AME): The AME of x_S on the target corresponds to the average of all MEs at observed values of the feature space. It is a scalar value that may be interpreted as a global feature effect estimate. It is important to take into consideration that aggregating MEs is connected with a loss of information when the MEs have a large variance, i.e. when the prediction function has a strong curvature.

Marginal effects at the mean (MEM): For MEMs, all feature values in x_S are substituted by their sampling distribution means before estimating MEs. The effects may be averaged. When averaged, the MEM is a variation of the AME on a modified dataset. MEMs can be misleading as the multi-dimensional sample mean of X may not be observed or even observable, especially for dummy variables [2].

Marginal effects at representative values (MER): MERs evaluate the prediction function on a modified dataset in a similar way to MEMs. The only difference being that the modified values are specified manually and do not depend on the sampling distribution. MERs can be considered to be conditional MEs. This is especially useful for evaluating counterfactuals. Representative values should be chosen inside the training data space because otherwise, the fitted model might extrapolate.

Shapley value: Originating in coalitional game theory [23], the Shapley value is a local feature effect measure that is based on a set of desirable axioms. In coalitional games, a set of p players, denoted by P, play games and join coalitions. They are rewarded with a payout. The characteristic function $v : 2^p \to \mathcal{R}$ maps all player coalitions to their respective payouts [4]. The Shapley value is a player's average contribution to the payout, i.e. the marginal increase in payout for the coalition of players, averaged over all possible coalitions. For Shapley values as feature effects, predicting the target for a single observation corresponds to the game and a coalition of features represents the players. Shapley regression values were first developed for linear models with multicollinear features [17]. A model agnostic Shapley value was first introduced in [23]. Consider the expected

prediction for an observed vector of feature values x, conditional on only observing feature values of subset K, i.e. features in C are marginalized out. This essentially equals a point (or a line, surface etc. depending on the power of K) on the PD from eq. (1).

$$\begin{aligned} &\mathbb{E}\left[\hat{f} \mid X_j = x_j, \forall j \in K\right] \\ &= \mathbb{E}_{X_C}\left[\hat{f}(x_K, X_C)\right] = \int \hat{f}(x_K, X_C)\, dP(X_C) \\ &= \widehat{PD}(x_K) \end{aligned} \tag{4}$$

Eq. (4) is shifted by the mean prediction and used as a payout function $v(x_K)$, such that an empty set of features ($K = \emptyset$) results in a payout of zero [4].

$$v_{PD}(x_K) = \mathbb{E}_{X_C}\left[\hat{f}(x_K, X_C)\right] - \mathbb{E}_{X_S \cup X_C}\left[\hat{f}(X_S, X_C)\right] = \widehat{PD}(x_K) - \widehat{PD}(x_\emptyset) \tag{5}$$

From eq. (5), it follows that the marginal contribution $\Delta_S(x_K)$ of feature values x_S joining the coalition of values x_K is:

$$\Delta_S(x_K) = v_{PD}(x_{K \cup S}) - v_{PD}(x_K) = \widehat{PD}(x_{K \cup S}) - \widehat{PD}(x_K) \tag{6}$$

The exact notation of the Shapley value for a single feature S with observational values x is:

$$\begin{aligned} \widehat{Shapley}_S(x) &= \sum_{K \subseteq (P \setminus S)} \frac{|K|!(|P| - |K| - 1)!}{|P|!}\, \Delta_S(x_K) \\ &= \sum_{K \subseteq (P \setminus S)} \frac{|K|!(|P| - |K| - 1)!}{|P|!} \left[\widehat{PD}(x_{K \cup S}) - \widehat{PD}(x_K)\right] \end{aligned}$$

Shapley values are computationally expensive because the PD function has a complexity of $\mathcal{O}(N^2)$. Computations can be sped up by Monte Carlo sampling [23] or taking advantage of possible graph structures in the data [5]. Furthermore, [18] proposes a distinct variant to compute Shapley values called SHapley Additive exPlanations (SHAP).

4 Generalized Framework

All demonstrated techniques are based on the same principles. Instead of trying to inspect the inner workings of a non-linear black box model, we are evaluating its predictions when changing inputs. The SIPA work stages (Sampling, Intervention, Prediction, Aggregation) serve as a framework for thinking about model agnostic techniques. The software package *iml* [20] contains implementations of all demonstrated techniques and was inspired by the SIPA framework.

Often, the amount of available data is large and the utilized algorithms computationally expensive. We first sample a subset (**sampling stage**) to reduce computational costs, e.g. selecting a random set of available observations to evaluate as ICE curves. In order to change the predictions made by the black box model, the data has to be manipulated. Feature values can be set to values from the observed marginal distributions (ICE & PD curves or Shapley values), or to to unobserved values (FD based methods such as MEs and ALEs). This crucial step is called the **intervention stage**. During the **prediction stage**, we predict on previously intervened data. This requires an already fitted model, which is why model agnostic techniques are always post hoc. The predictions are further aggregated during the **aggregation stage**. Often, the predictions resulting from the prediction stage are local effect estimates, and the ones resulting from the aggregation stage are global effect estimates. Optionally, results can be visualized. In fig. 1, we demonstrate how all demonstrated techniques for feature effects are based on the SIPA framework. For every method, we optionally sample a random subset of observations from the test data (sampling stage) and then proceed to the intervention stage.

ICE & PD: For each observation, we create copies of the observed vector and create an identical set with permuted values of x_S (intervention stage). We predict for each permutation (prediction stage). The predictions for one observation correspond a single ICE curve. All estimated ICE curves may then be averaged point-wise to the PD (aggregation stage).

ALE: We first partition the value range of x_S into a set of intervals and select a subset of observations in each interval. We iterate over every interval. For every observation inside an interval, we set its value of x_S to the right and left interval boundary (intervention stage). We predict at both boundaries (prediction stage). All FDs inside an interval are averaged to an interval-wise average FD. We integrate the partial derivative to the ALE by summing up the interval-wise average FDs (aggregation stage).

MEs: For MEMs, we first set all feature values in x_C for all observations to the sampling distribution means. For MERs, we set some feature values in x_C for some observations to manually specified values. For AMEs, we use the observed data. In all variants, we construct FDs for each observation, i.e. we take a forward and backward step with size h (intervention stage). We predict at both interval boundaries (prediction stage). In the aggregation stage, we compute the FD of predictions and divide by the interval width $2h$ in order to get ME estimates. For AMEs and optionally MEMs, ME estimates can be averaged.

Shapley value: We first construct all possible feature sequences without feature S. For each sequence, we iterate over all observations and construct permutations for ICE curves, once with and once without feature S (intervention stage). We predict for all permutations (prediction stage). In the aggregation stage, we average the ICE curves of both groups to their PD functions. Next, we plug in the observed vector x and take the difference of PD values, i.e. the marginal contribution to the PD of feature S being included in the sequence.

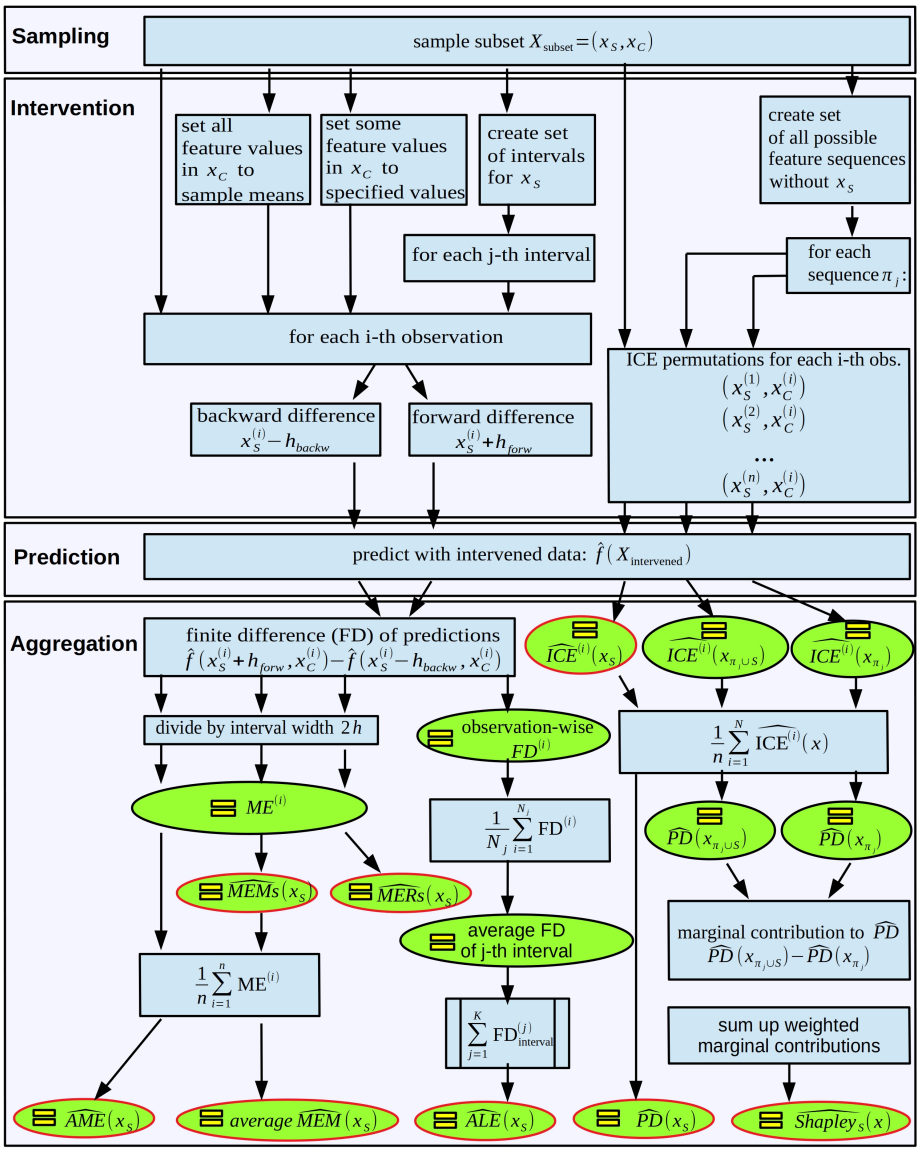

Fig. 1. All demonstrated techniques for feature effects are based on the SIPA framework. Note that one point of each ICE curve corresponds to observed feature values. For reasons of simplicity, we refer to all points on ICE curves as intervened data.

Lastly, we sum up the weighted marginal contributions to the local Shapley value (aggregation stage).

5 Feature Importance

Methods to determine feature effects (direction and magnitude) are tightly intertwined with those determining the importance (contribution to the predictive performance) of a feature. Choosing an appropriate set of features plays an important role in the model building process (feature selection). We are instead interested in interpreting a single, already trained model with a fixed set of features.

In linear models, the quotient of the absolute coefficient and the corresponding standard error serves as a natural feature importance metric. This corresponds to the absolute value of the t-statistic [11]. In the case of black box models, we neither know the analytical form of coefficients nor their standard errors. We categorize importance measures in two groups: variance-based and performance-based.

Variance-based: A mostly flat trajectory of a single ICE curve implies that in the underlying predictive model, varying x_S does not affect the prediction for this specific observation. If all ICE curves are shaped similarly, the PD can be used instead. [11] propose a measure for the curvature of the PD as a feature importance metric. Let the average value of the PD of feature S be denoted by $\overline{\widehat{PD}}(x_S) = \frac{1}{n}\sum_{i=1}^{n} \widehat{PD}^{(i)}(x_S^{(i)})$. The importance $I(x_S)$ of feature S corresponds to the estimated standard deviation of the feature's PD function. The flatter the PD, the smaller its standard deviation and therefore the feature importance. For categorial features, the range of the PD is divided by 4. This is supposed to represent an approximation to the estimate of the standard deviation for small to medium sized samples [11].

$$I(x_S) = \begin{cases} \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left[\overline{\widehat{PD}}(x_S^{(i)}) - \overline{\widehat{PD}}(x_S)\right]^2} & x_S \text{ continuous} \\ \frac{1}{4}\left[max_i\left\{\overline{\widehat{PD}}(x_S^{(i)})\right\} - min_i\left\{\overline{\widehat{PD}}(x_S^{(i)})\right\}\right] & x_S \text{ categorial} \end{cases}$$

[24] propose the feature importance ranking measure (FIRM). They define a conditional expected score (CES) function for feature S.

$$CES_S(v) = \mathbb{E}_{X_C}\left[\hat{f}(x_S, X_C)|x_S = v\right] \tag{7}$$

It turns out that eq. (7) is equivalent to eq. (1), conditional on $x_S = v$. It follows that the CES is a conditional PD.

$$\begin{aligned} CES_S(v) &= \mathbb{E}_{X_C}\left[\hat{f}(v, X_C)\right] \\ &= PD(v) \end{aligned}$$

The FIRM corresponds to the standard deviation of the CES function with all values of x_S used as conditional values. This in turn is equivalent to the standard deviation of the PD curve and therefore to the feature importance metric proposed by [11].

$$FIRM_S = \sqrt{Var(CES_S(x_S))} = \sqrt{Var(PD(x_S))}$$

Performance-based: The permutation feature importance (PFI), originally developed by [3] as a model specific tool for random forests, was described as a model agnostic one by [7]. If feature values are shuffled in isolation, the relationship between the feature and target is broken up. If the feature was important for the predictive performance, the shuffling should result in an increased loss [4]. The model agnostic PFI of a feature S measures the difference between the GE on data with permuted and non-permuted values. Permuting x_S corresponds to drawing from a new random variable $\tilde{X}_S$ that is distributed like X_S but independent of X_C [4].

$$PFI_S = \mathbb{E}\left[\mathcal{L}(\hat{f}(\tilde{X}_S, X_C), Y)\right] - \mathbb{E}\left[\mathcal{L}(\hat{f}(X), Y)\right]$$

Consider the sample of test data $\mathcal{D}_S$ where the columns of S have been permuted and the non-permuted sample $\mathcal{D}$. The PFI estimate is given by the difference between GE estimates with permuted and non-permuted data.

$$\widehat{PFI}_S = \widehat{GE}(\hat{f}, \mathcal{D}_S) - \widehat{GE}(\hat{f}, \mathcal{D}) \tag{8}$$

Extending the idea of measuring differences in performance, [4] propose individual conditional importance (ICI) and partial importance (PI) curves as visualization techniques that disaggregate the global PFI estimate. They are based on the same principles as ICE and PD. The ICI visualizes the influence of a feature on the predictive performance for a single observation, while the PI visualizes the average influence of a feature for all observations. Consider the prediction for the i-th observation with observed values $\hat{f}(x_S^{(i)}, x_C^{(i)})$ and the prediction $\hat{f}(x_S^{(l)}, x_C^{(i)})$ where $x_S^{(i)}$ was replaced by a value $x_S^{(l)}$ from the marginal distribution of observed values x_S. The change in loss is given by:

$$\Delta\mathcal{L}^{(i)}(x_S^{(l)}) = \mathcal{L}(\hat{f}(x_S^{(i)}, x_C^{(i)})) - \mathcal{L}(\hat{f}(x_S^{(l)}, x_C^{(i)}))$$

The ICI curve of the i-th observation plots the value pairs $(x_S^{(l)}, \Delta\mathcal{L}^{(i)}(x_S^{(l)}))$ for all l values of x_S. The PI curve is the point-wise average of all ICI curves at all l values of x_S. It plots the value pairs $(x_S^{(l)}, \frac{1}{n}\sum_{i=1}^{n}\Delta\mathcal{L}^{(i)}(x_S^{(l)})$ for all l values of x_S. Estimating the ICI is nearly identical to estimating the ICE with only two differences. First, we not only predict on an intervened dataset during the prediction stage, but instead we predict on both intervened and non-intervened data. Second, in the aggregation stage, we use the loss functions instead of the absolute predictions and compute the difference between both losses. Substituting values of x_S essentially resembles shuffling them. The authors demonstrate

how integrating the PI curve with respect to x_S results in an estimation of the global PFI.

$$\widehat{PFI_S} = \frac{1}{n}\sum_{i=1}^{n} \widehat{PFI_S}^{(i)} = \frac{1}{n}\sum_{l=1}^{n}\frac{1}{n}\sum_{i=1}^{n} \Delta\mathcal{L}^{(i)}(x_S^{(l)})$$

Furthermore, [4] propose a feature importance measure they call Shapley feature importance (SFIMP). Shapley importance values were first introduced by [6] for feature selection. This requires refitting the model with distinct sets of features, which changes the behavior of the learning algorithm and is not helpful to evaluate a single model, as noted by [4]. The SFIMP is based on the same principle as deriving ICI from ICE curves, i.e. using the same computations as the Shapley value but replacing the payout function with one that is sensitive to the model performance. The authors define a new payout $v_{GE}(S)$ that substitutes the estimated PD with the estimated generalization error (GE). This is equivalent to the expected PFI from eq. (8).

$$v_{GE}(S) = \widehat{GE}(\hat{f}, \mathcal{D}_S) - \widehat{GE}(\hat{f}, D) = \widehat{PFI}_S = v_{PFI}(S)$$

We can therefore refer to $v_{GE}(S)$ as $v_{PFI}(S)$ and regard the SFIMP as an extension to the PFI [4].

6 Extending the Framework to the Feature Importance

It may not surprise that feature importance measures are also based on the SIPA framework. Variance-based methods simply measure the variance of some kind of feature effect estimate, which we already demonstrated to be based on the SIPA framework. Performance-based techniques measure some form of loss, i.e. there are two possible modifications. First, we predict on non-intervened or intervened data (prediction stage). Second, we aggregate predictions to the loss (aggregation stage).

In fig. 2, we demonstrate how feature importance computations are based on the same work stages as feature effect computations.

7 Conclusion

Model agnostic interpretation techniques recently garnered widespread attention. The lack of interpretability of most ML models proved to be an obstacle for their adoption, e.g. in high-stakes decision making. Modularity is a central motive in using ML, i.e. the ability to algorithmically train and substitute models according to certain performance measurements. Therefore, model agnostic interpretation methods represent a major stepping stone in advancing the field, because they allow us to substitute interpretation techniques as well. A variety of model agnostic methods has been developed. Different terminologies complicated the understanding of methods and how they are related to each other. By

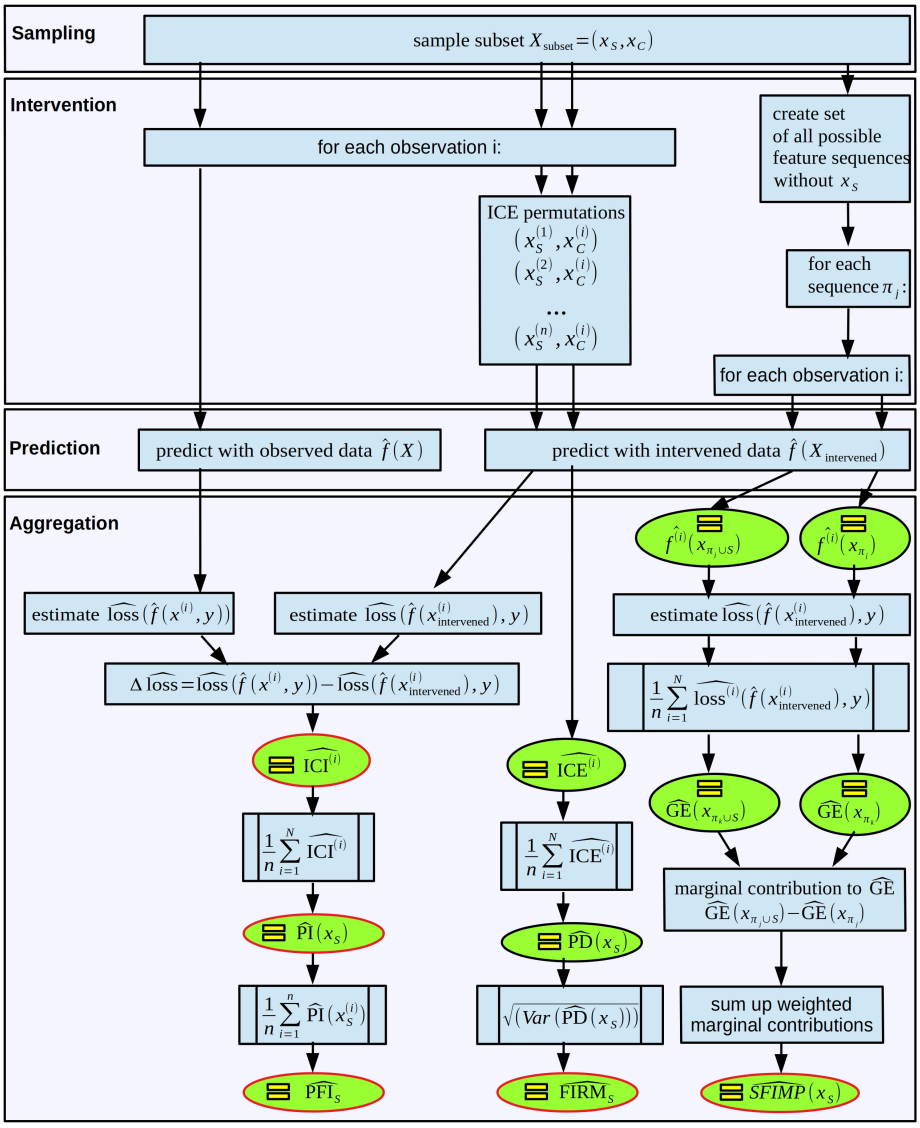

Fig. 2. Extending the framework to the feature importance. Model agnostic importance measures are based on the same work stages as methods for feature effects. Variance based performance measures evaluate the variance of effect measures. Performance based measures evaluate the loss function.

deconstructing them into sequential work stages, one discovers striking similarities in their methodologies. With our contributions, we hope to work towards a unified view on model agnostic techniques and to facilitate their understanding and discussion.

In our first contribution, we formally added marginal effects from statistics to the range of black box interpretation techniques in ML. Our main contribution was to present the generalized SIPA framework of work stages for model agnostic interpretations. Essentially, all model agnostic methods are based on four work stages. First, there is a sampling stage to reduce computational costs. Second, we intervene in the data in order to change the behavior of the black box model. Third, we predict on intervened or non-intervened data. Fourth, we aggregate predictions. Optionally, we can visualize the results. We embedded multiple methods to estimate the effect (ICE & PD, ALEs, MEs and Shapley values) and importance (FIRM, PFI, ICI & PI and the SFIMP) of features into the framework.

Realizing that model agnostic interpretation techniques are essentially operating according to the same principle, may provide researchers and applicants of ML with more clarity about the workings of methods they are using. The SIPA framework may therefore serve as a guideline to conduct model agnostic interpretations and inspire the development of novel methods.

Acknowledgments

This work is supported by the Bavarian State Ministry of Science and the Arts as part of the Centre Digitisation.Bavaria (ZD.B) and by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A. The authors of this work take full responsibilities for its content.

References

1. Apley, D.W.: Visualizing the effects of predictor variables in black box supervised learning models. ArXiv e-prints (Dec 2016)
2. Bartus, T.: Estimation of marginal effects using margeff. The Stata Journal **5**(3), 309 – 329 (2005)
3. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (Oct 2001), `https://doi.org/10.1023/A:1010933404324`
4. Casalicchio, G., Molnar, C., Bischl, B.: Visualizing the feature importance for black box models. ArXiv e-prints (Apr 2018)
5. Chen, J., Song, L., Wainwright, M.J., Jordan, M.I.: L-shapley and C-shapley: Efficient model interpretation for structured data. CoRR **abs/1808.02610** (2018), `http://arxiv.org/abs/1808.02610`
6. Cohen, S., Dror, G., Ruppin, E.: Feature selection via coalitional game theory. Neural Computation **19**(7), 1939–1961 (2007), `https://doi.org/10.1162/neco.2007.19.7.1939`
7. Fisher, A., Rudin, C., Dominici, F.: Model class reliance: Variable importance measures for any machine learning model class, from the "Rashomon" perspective. ArXiv e-prints (Jan 2018)

8. Fisher, A., Rudin, C., Dominici, F.: All models are wrong but many are useful: Variable importance for black-box, proprietary, or misspecified prediction models, using model class reliance. arXiv e-prints arXiv:1801.01489 (Jan 2018)
9. Friedman, J.H.: Greedy function approximation: A gradient boosting machine. Ann. Statist. **29**(5), 1189–1232 (10 2001), https://doi.org/10.1214/aos/1013203451
10. Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E.: Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. ArXiv e-prints (Sep 2013)
11. Greenwell, B.M., Boehmke, B.C., McCarthy, A.J.: A simple and effective model-based variable importance measure. ArXiv e-prints (May 2018)
12. Hechtlinger, Y.: Interpretation of prediction models using the input gradient. arXiv e-prints arXiv:1611.07634 (Nov 2016)
13. Hooker, G.: Discovering additive structure in black box functions. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 575–580. KDD '04, ACM, New York, NY, USA (2004), http://doi.acm.org/10.1145/1014052.1014122
14. Hooker, G.: Generalized functional anova diagnostics for high-dimensional functions of dependent variables. Journal of Computational and Graphical Statistics **16**(3), 709–732 (2007), https://doi.org/10.1198/106186007X237892
15. Kleiber, C., Zeileis, A.: Applied econometrics with R. Springer-Verlag, New York (2008), https://CRAN.R-project.org/package=AER, ISBN 978-0-387-77316-2
16. Leeper, T.J.: margins: Marginal effects for model objects (2018), R package version 0.3.23
17. Lipovetsky, S., Conklin, M.: Analysis of regression in game theory approach. Applied Stochastic Models in Business and Industry **17**(4), 319–330 (October 2001), https://ideas.repec.org/a/wly/apsmbi/v17y2001i4p319-330.html
18. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 4765–4774. Curran Associates, Inc. (2017), http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf
19. Molnar, C.: Interpretable Machine Learning. https://christophm.github.io/interpretable-ml-book/ (2019)
20. Molnar, C., Bischl, B., Casalicchio, G.: iml: An R package for interpretable machine learning. JOSS **3**(26), 786 (2018), http://joss.theoj.org/papers/10.21105/joss.00786
21. Rudin, C., Ertekin, Ş.: Learning customized and optimized lists of rules with mathematical programming. Mathematical Programming Computation **10**(4), 659–702 (Dec 2018), https://doi.org/10.1007/s12532-018-0143-8
22. Stone, C.J.: The use of polynomial splines and their tensor products in multivariate function estimation. The Annals of Statistics pp. 118–171 (1994)
23. Štrumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems **41**(3), 647–665 (Dec 2014), https://doi.org/10.1007/s10115-013-0679-x
24. Zien, A., Krämer, N., Sonnenburg, S., Rätsch, G.: The feature importance ranking measure. In: Buntine, W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J. (eds.) Machine Learning and Knowledge Discovery in Databases. pp. 694–709. Springer Berlin Heidelberg, Berlin, Heidelberg (2009)