A Unified Approach to Interpreting Model Predictions

Authors:

Scott M. Lundberg, Su-In Lee

Presented by:

Amirehsan Davoodi

AmirKabir University of Technology

01.06.2024

Quotes

- “The **best explanation** of a simple model is the **model itself**; it perfectly represents itself and is easy to understand.”
- Viewing any **explanation of a model’s prediction** as a model itself, which we term the ***explanation model***.
- ***Explanation model:*** defined as any interpretable approximation of the original model.

Outlines

- Shapley Value: Game theory definition
- Additive Feature Importance
- Additive feature attribution methods

Definition

- SHAP (SHapley Additive exPlanations) assigns an importance value to each feature for a particular/single prediction.
- The SHAP value for a feature i in the context of a prediction x is denoted as $\phi_i(x)$
- $\underline{Shapley\ Value = Expected\ marginal\ contribution}$
- $\underline{Shapley\ Value = weighted\ average\ of\ a\ player's\ contribution\ of\ all\ the\ coalitions\ which\ the\ player\ could\ join}$
- $\underline{Shapley\ Value = Fair\ way\ to\ divide\ a\ game\ prize\ amongst\ it's\ players.}$

Shapley Value

Competition Prize

First	Second	Third
$10,000	$7,500	$5,000

Player 1

Player 2

Coalition Values

$$C_{12} = 10{,}000$$
$$C_{1} = 7{,}500$$
$$C_{2} = 5{,}000$$
$$C_{0} = 0$$

Shapley Value (Marginal Contribution)

Coalition Values

$C_{12} = 10{,}000$
$C_1 = 7{,}500$
$C_2 = 5{,}000$
$C_0 = 0$

$C_{12} - C_2 = 5{,}000$
$C_1 - C_0 = 7{,}500$

$$\frac{(5{,}000 + 7{,}500)}{2} = 6{,}250$$

Marginal Contribution:
The increase in a coalition's value due to a player joining that coalition

$C_{12} - C_1 = 2{,}500$
$C_2 - C_0 = 5{,}000$

$$\frac{(2{,}500 + 5{,}000\ )}{2} = 3{,}750$$

Shapley Value (Marginal Contribution)

$$C_{12} - C_2 = 5{,}000$$
$$C_1 - C_0 = 7{,}500$$
$$\frac{(5{,}000 + 7{,}500)}{2} = 6{,}250$$

Compute Probability instead −> Expected Marginal Contribution

$P(C_{12} - C_2)$= Probability that player 1 makes a marginal contribution to a coalition of player 2

Shapley Value Formulation

Fair value for player i in a p player game

Weight

Marginal contribution of player i to coalition S

$$\phi_i = \sum_{S \subseteq \{1,\ldots,p\} \backslash \{i\}} \frac{|S|!\,(p - |S| - 1)!}{p!} \left[val(S \cup \{i\}) - val(S)\right]$$

Probabilities

All the coalitions player i can join

Expected Marginal Contribution

Shapley Value Formulation

$$\phi_i = \sum_{S \subseteq \{1,\ldots,p\}\{i\}} \frac{|S|!\,(p - |S| - 1)!}{p!}\left[val(S \cup \{i\}) - val(S)\right]$$

- p = number of players
- $p!$ = number of ways to form a coalition of p players
- S = the coalition of players
- $|S|$= number of players in the coalition S
- $|S|!$ = number of ways coalition S can form
- $(p - |S| - 1)!$ = number of ways players can join after player i joins a coalition S

Shapley Value for Explainable Model

$$\phi_i = \sum_{S \subseteq \{1,\ldots,p\}\{i\}} \frac{|S|!\,(p - |S| - 1)!}{p!}\left[val(S \cup \{i\}) - val(S)\right]$$

$$val_x(S) = \int f(x_1, \ldots, x_p)\, dP_{x \notin S}$$

Value Predicted by a model

- p = number of ~~players~~ features in the model
- $p!$ = number of ways to form a coalition of p ~~players~~ features
- S = the coalition of ~~players~~ feature values
- f = model prediction
- $|S|$= number of ~~players~~ features in the coalition S
- $|S|!$ = number of ways coalition S can form
- $(p - |S| - 1)!$ = number of ways ~~players~~ features can join after ~~player~~ feature i joins a coalition S

Shapley Value for Explainable Model

x_1 (Age): 20

x_2 (Degree): 1

y (Predicted Income): $5,000

$$f(x_1, x_2) = 200x_1 + 1000x_2$$

$$age \ \rightarrow x_1 \in [18, 60]$$

$$degree \ \rightarrow x_2 \in \{0, 1\}$$

What is the marginal Contribution of degree {2} to the coalition of age {2}?

$$\begin{aligned} val_x(\{1,2\}) &= f(20,1) \\ &= 200(20) + 100(1) \\ &= 5000 \end{aligned}$$

$$\begin{aligned} val_x(\{1\}) &= \int f(20, x_2)\, dP_{x_2} \\ &= \sum_{i=0}^{1} f(20, x_2) P(x_2 = i) \\ &= \big(200(20) + 1000\ (0)\big)(0.5) \\ &+ \big(200(20) + 1000(1)\big)(0.5) \\ &= 4500 \end{aligned}$$

$$val_x(\{1,2\}) - val_x(\{1\}) = 500$$

Shapley Value for Explainable Model

$$\phi_i = \sum_{S \subseteq \{1,\ldots,p\}\{i\}} \frac{|S|!\,(p-|S|-1)!}{p!}\left[val(S \cup \{i\}) - val(S)\right]$$

$$val_x(S) = \int f(x_1, \ldots, x_p)\,dP_{x \notin S}$$

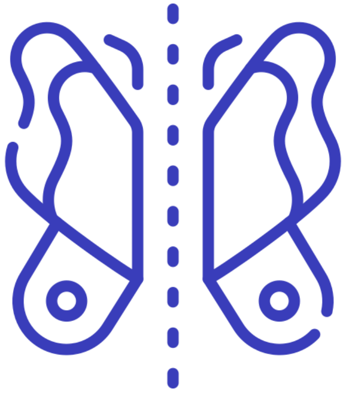
Value Predicted by a model

- p = number of ~~players~~ features in the model
- $p!$ = number of ways to form a coalition of p ~~players~~ features
- S = the coalition of ~~players~~ feature values
- f = model prediction
- $|S|$= number of ~~players~~ features in the coalition S
- $|S|!$ = number of ways coalition S can form
- $(p-|S|-1)!$ = number of ways ~~players~~ features can join after ~~player~~ feature i joins a coalition S

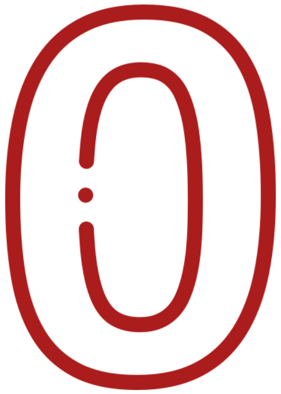
Why Shapley value is fair?

Efficiency

Symmetric



Dummy



Additivity

Consistency

Why Shapley value is fair?

Efficiency

$$\boxed{f(x) = \overbrace{\sum_{i=1}^{p} \phi_i}^{\text{Sum of all Shapley Values}} + \overbrace{E_X[f(X)]}^{\text{Average Predicted Value}}}$$

Prediction is divided among the features

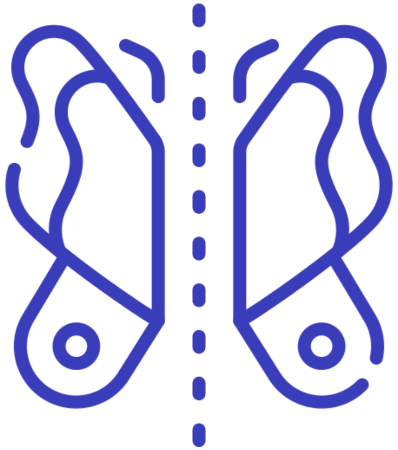
Why Shapley value is fair?

Efficiency

- **LIME** is not necessarily Efficient
- With LIME we know which feature is most important to that prediction

Why Shapley value is fair?

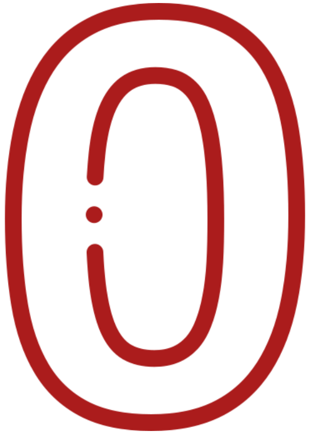
Symmetric



- Two features have same Shapley Value if they have same contributions to all coalitions.

Why Shapley value is fair?

Dummy



- A feature will have zero/null Shapley Value if it never changes the prediction.
- Features that are not used in a model will not have Shapley Value.

Why Shapley value is fair?

Additivity

- Relevant for Ensemble models.
- Overall Shapley Value is the weighted average of the Shapley values of all the models in the Ensemble model

Why Shapley value is fair?

Consistency

- If we change a model and the marginal contribution of the feature changes then the feature’s Shapley Value will change in the same direction
- Reliably compare the Shapley Values of different models

Shapley Value for Explainable Model

- p = number of ~~players~~ features in the model
- $p!$ = number of ways to form a coalition of p ~~players~~ features
- S = the coalition of ~~players~~ feature values
- f = model prediction
- $|S|$= number of ~~players~~ features in the coalition S
- $|S|!$ = number of ways coalition S can form
- $(p - |S| - 1)!$ = number of ways ~~players~~ features can join after ~~player~~ feature i joins a coalition S

Additive Feature Importance

- SHAP introduces a new class of additive feature importance measures.
- The SHAP value for a feature i in a prediction x can be expressed as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right].$$

- $\phi i(x)$: This represents the SHAP value for feature i in the context of prediction x. It quantifies the impact of feature i on the model's prediction for the input x.
- S: This is a subset of the set of features $\{1, \dots, M\}$ excluding feature i. It represents the set of features that do not include feature i in the calculation of the SHAP value for feature i.
- M: This denotes the total number of features in the model.
- $f(x_S)$: This represents the model's prediction when considering only the features in subset S, excluding feature i
- $f(x_{S\cup\{i\}})$: This represents the model's prediction when including feature i in addition to the features in subset S.
- $|S|$: This denotes the cardinality (number of elements) of subset S.
- $|S|!(M - |S| - 1)!/M!$: This term is the Shapley value, which is a weighted average of the marginal contributions of feature i across all possible subsets of features.

Additive feature attribution methods

- M: This denotes the total number of features in the model.
- $f(x)$: machine learning model
- $g(x)$: explanation model
- $z' \in \{0,1\}^M$: simplified input features

$$\boldsymbol{z' \cong x'}$$

$$\boldsymbol{g(z') \cong f(h_{x(z')})}$$

Additive feature attribution methods

- Explanation model as a linear function of binary variables

$$\boldsymbol{g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i}$$

- $\phi_i(x) \in R$: This represents the value for feature i in the context of prediction z'. It quantifies the impact of feature i on the model's prediction for the input z'.

Additive feature attribution methods

LIME

$$\xi = \underset{g \in \mathcal{G}}{\arg\min}\ L(f, g, \pi_{x'}) + \Omega(g).$$

- $L(f, g, \pi_{x'})$: Loss function
- $\pi_{x'}$: Local kernel
- Ω: penalty for the complexity of g
- It quantifies the impact of feature i on the model's prediction for the input z'.

Additive feature attribution methods

DeepLIFT

$$\sum_{i=1}^{n} C_{\Delta x_i \Delta o} = \Delta o,$$

$$\phi_i = C_{\Delta x_i \Delta_o}$$

$$\phi_o = f(r)$$

- $C_{\Delta x_i \Delta_y}$: It attributes to each input x_i a value $C_{\Delta x_i \Delta_y}$ that represents the effect of that input being set to a reference value as opposed to its original value.
- $o = f(x)$
- r: reference input
- $\Delta_o = f(x) - f(r)$
- $\Delta_{x_i} = x_i - r_i$

Additive feature attribution methods

Layer-Wise Relevance Propagation

- This method is equivalent to DeepLIFT with the reference activations of all neurons fixed to zero.

Computationally Expensive

- The number of possible coalitions between features increase exponentially with respect to the number of features.
- Approximate Shapley Values:
  - Monte Carlo Sampling

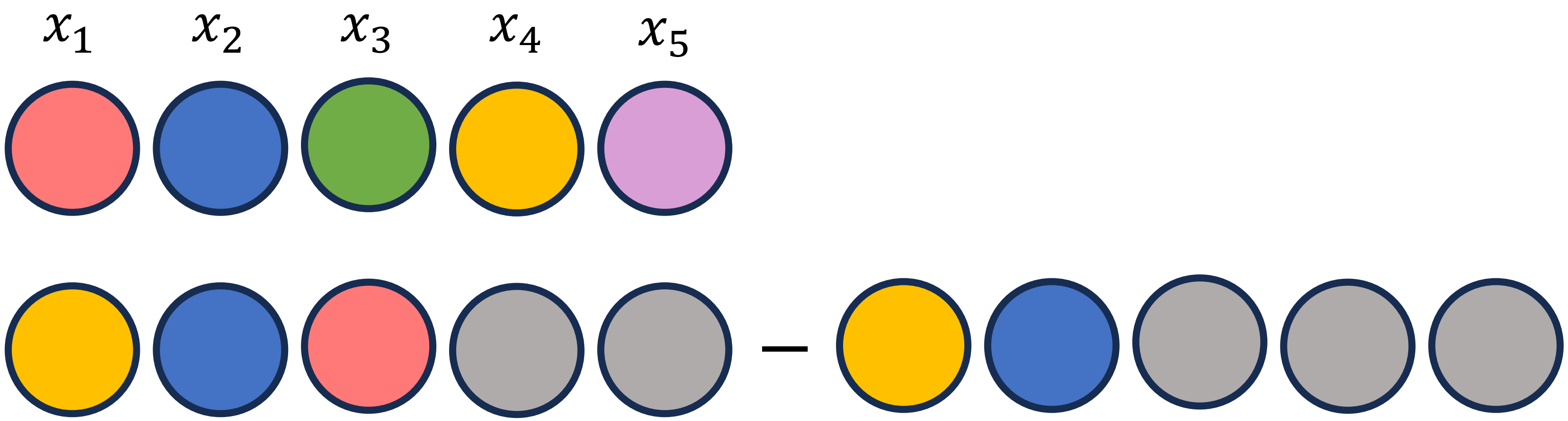
$$\hat{\phi}_i = \frac{1}{M}\sum_{m=1}^{M}\left(f(x_{+i}^{m}) - f(x_{-i}^{m})\right)$$

  - KernelSHAP
    - Linear Regression Estimation
  - TreeSHAP
    - Take advantage of the structure of individual trees in Ensemble models
    - Only for tree-based algorithms (Random Forest, Xgboost, …)

Computationally Expensive

- Monte Carlo Sampling

$$\hat{\phi}_i = \frac{1}{M}\sum_{m=1}^{M}\left(f(x_{+i}^{m}) - f(x_{-i}^{m})\right)$$

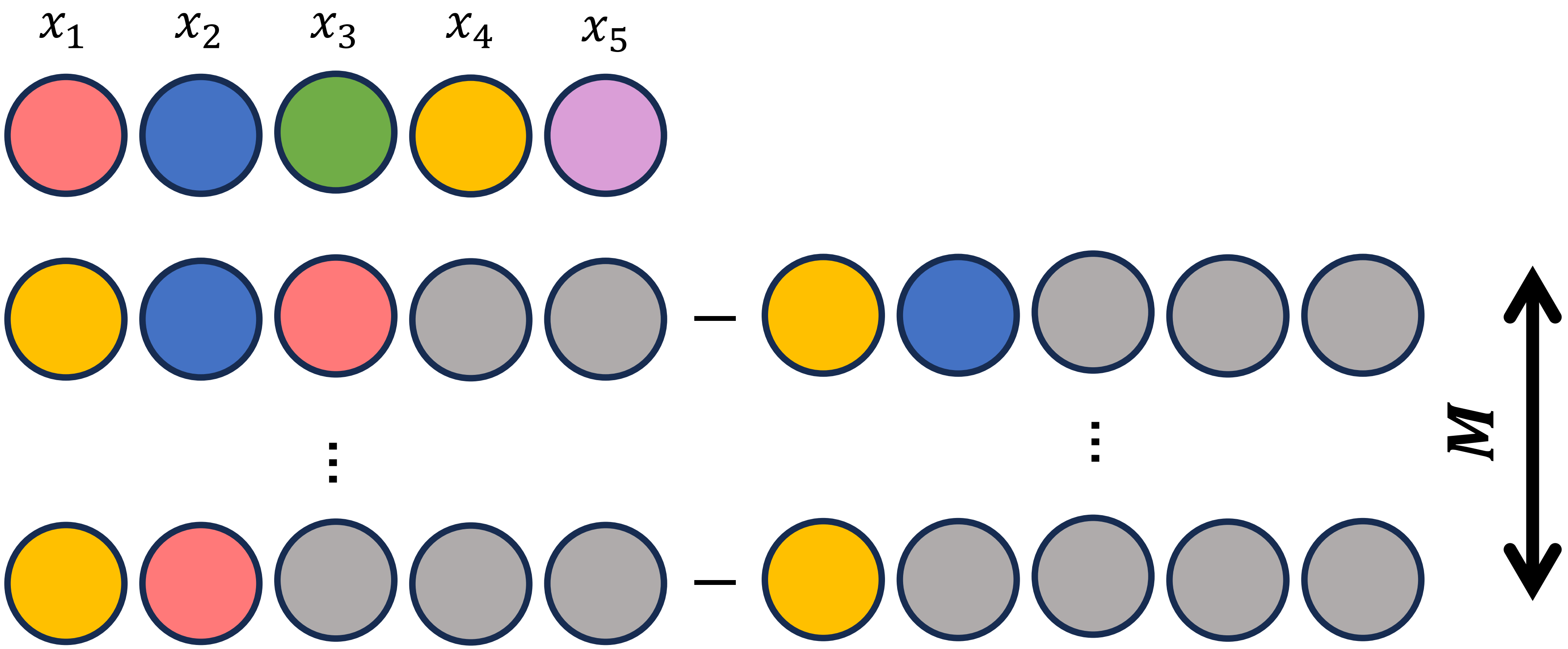


$$f(x_{+i}^{m}) \quad - \quad f(x_{-i}^{m})$$

Computationally Expensive

- Monte Carlo Sampling

$$\hat{\phi}_i = \frac{1}{M}\sum_{m=1}^{M}\left(f(x_{+i}^m) - f(x_{-i}^m)\right)$$



Computationally Expensive

- KernelSHAP