Explainable Machine Learning with Shapley Value

Gianmarco Midena

Aalto University

12 March 2024

The Attribution Problem[1]

- Distribute the prediction score of a model to its input features for a specific data point.

[1] Ribeiro et al. (2016), Lundberg and Lee (2017), Sundararajan et al. (2017).

The Attribution Problem[1]

- Distribute the prediction score of a model to its input features for a specific data point.

- How each feature affects the prediction for a particular data point.

[1] Ribeiro et al. (2016), Lundberg and Lee (2017), Sundararajan et al. (2017).

The Attribution Problem[1]

- Distribute the prediction score of a model to its input features for a specific data point.
- How each feature affects the prediction for a particular data point.
- Importance of a feature value to a prediction

[1] Ribeiro et al. (2016), Lundberg and Lee (2017), Sundararajan et al. (2017).

The Attribution Problem[1]

- Distribute the prediction score of a model to its input features for a specific data point.
- How each feature affects the prediction for a particular data point.
- Importance of a feature value to a prediction
- Attributions have explanatory value

[1] Ribeiro et al. (2016), Lundberg and Lee (2017), Sundararajan et al. (2017).

The Attribution Problem[1]

- Distribute the prediction score of a model to its input features for a specific data point.
- How each feature affects the prediction for a particular data point.
- Importance of a feature value to a prediction
- Attributions have explanatory value
- What-if analysis

[1] Ribeiro et al. (2016), Lundberg and Lee (2017), Sundararajan et al. (2017).

Example - Probability of Cervical Cancer for a Woman

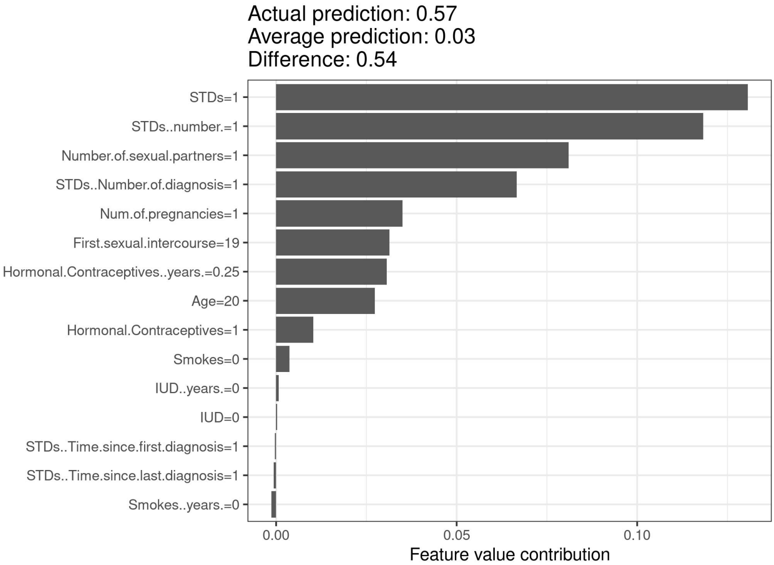

Image credit: *Molnar (2020)*

Risk Factors for Cervical Cancer[2]

- Has patient ever had a sexually transmitted disease (STD) [binary]
- Number of sexual partners
- Number of STD diagnoses
- Number of pregnancies
- First sexual intercourse (age in years)
- Hormonal contraceptives (in years)
- Age in years
- Hormonal contraceptives [binary]
- Smokes (binary)
- Number of years with an intrauterine device (IUD)
- Intrauterine device (IUD) [binary]
- Time since first STD diagnosis
- Time since last STD diagnosis
- Smokes (in years)

[2] Fernandes et al. (2017)

Example - Number of Rented Bikes for a Day

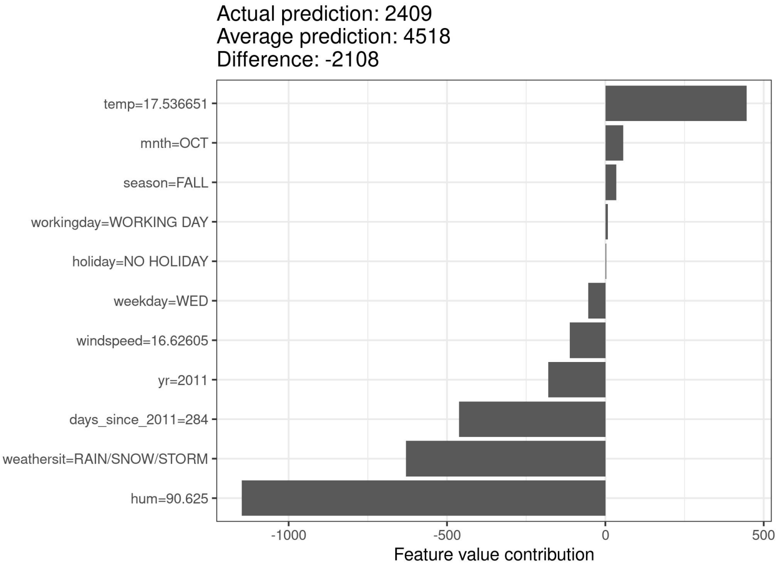

Image credit: *Molnar (2020)*

Bike Rental Features[3]

- Temperature in degrees Celsius
- Season: spring, summer, fall or winter
- Working day or weekend
- Holiday or not
- Wind speed in km per hour
- Year: 2011 or 2012
- Nr. days since 01.01.2011 (the first day in the dataset).
- Weather situation:
  a. clear, few clouds, partly cloudy, cloudy
  b. mist + clouds, mist + broken clouds, mist + few clouds, mist
  c. light snow, light rain + thunderstorm + scattered clouds, light rain + scattered clouds
  d. heavy rain + ice pallets + thunderstorm + mist, snow + mist
- Relative humidity percentage

[3] Fanaee-T (2013)

Linear Model

- Prediction function

$$\hat{f}(\boldsymbol{x}) = w_0 + w_1 x_1 + \cdots + w_p x_p \tag{1}$$

  - x_j: value of feature j
  - w_j: weight corresponding to feature j
    - j-th feature **global** importance
      -Standardized input features
      -(typically) How a **specific** feature value influences the prediction is more interesting!
  - p: nr. features

Linear Model

- Prediction function

$$\hat{f}(\boldsymbol{x}) = w_0 + w_1 x_1 + \cdots + w_p x_p \tag{1}$$

  - x_j: value of feature j
  - w_j: weight corresponding to feature j
    - j-th feature **global** importance
      -Standardized input features
      -(typically) How a **specific** feature value influences the prediction is more interesting!
  - p: nr. features

- No feature interaction → *individual effects* are easy to compute

Linear Model - Feature Contribution

$$
\begin{aligned}
\phi_j^{add}\left(\boldsymbol{x};\hat{f}\right) &= w_j x_j - \mathbb{E}\left[w_j X_j\right] \\
&= w_j\left(x_j - \mathbb{E}\left[X_j\right]\right) \\
&= \hat{f}(\boldsymbol{x}) - \mathbb{E}\left[\hat{f}(x_1,\dots,X_j,\dots,x_n)\right]
\end{aligned}
\tag{2}
$$

where

- $\mathbb{E}(w_j X_j)$: mean effect estimate for feature j

- j-th feature contribution

Linear Model - Feature Contribution

$$\begin{aligned}\phi_j^{add}\left(\boldsymbol{x};\hat{f}\right) &= w_j x_j - \mathbb{E}\left[w_j X_j\right] \\ &= w_j\left(x_j - \mathbb{E}\left[X_j\right]\right) \\ &= \hat{f}(\boldsymbol{x}) - \mathbb{E}\left[\hat{f}(x_1,\ldots,X_j,\ldots,x_n)\right]\end{aligned} \tag{2}$$

where

- $\mathbb{E}(w_j X_j)$: mean effect estimate for feature j

- j-th feature contribution
- j-th feature effect minus average j-th feature effect

Linear Model - Feature Contribution

$$\begin{aligned}\phi_j^{add}\left(\boldsymbol{x}; \hat{f}\right) &= w_j x_j - \mathbb{E}\left[w_j X_j\right] \\ &= w_j\left(x_j - \mathbb{E}\left[X_j\right]\right) \\ &= \hat{f}(\boldsymbol{x}) - \mathbb{E}\left[\hat{f}(x_1, \dots, X_j, \dots, x_n)\right]\end{aligned} \tag{2}$$

where

- $\mathbb{E}(w_j X_j)$: mean effect estimate for feature j

- j-th feature contribution
- j-th feature effect minus average j-th feature effect
- Contribution of feature X_j with value x_j
  minus the expected contribution of feature X_j

Linear Model - Feature Contribution

$$
\begin{aligned}
\phi_j^{add}\left(\boldsymbol{x};\hat{f}\right) &= w_j x_j - \mathbb{E}\left[w_j X_j\right] \\
&= w_j \left(x_j - \mathbb{E}\left[X_j\right]\right) \\
&= \hat{f}(\boldsymbol{x}) - \mathbb{E}\left[\hat{f}(x_1,\ldots,X_j,\ldots,x_n)\right]
\end{aligned}
\tag{2}
$$

where

- $\mathbb{E}(w_j X_j)$: mean effect estimate for feature j

- j-th feature contribution
- j-th feature effect minus average j-th feature effect
- Contribution of feature X_j with value x_j
  minus the expected contribution of feature X_j
- Model prediction
  minus expected prediction if j-th feature value is not known

Linear Model - Feature Contribution

$$\begin{aligned}\phi_j^{add}\left(\boldsymbol{x};\hat{f}\right) &= w_j x_j - \mathbb{E}\left[w_j X_j\right] \\ &= w_j\left(x_j - \mathbb{E}\left[X_j\right]\right) \\ &= \hat{f}(\boldsymbol{x}) - \mathbb{E}\left[\hat{f}(x_1,\ldots,X_j,\ldots,x_n)\right]\end{aligned} \tag{2}$$

where

- $\mathbb{E}(w_j X_j)$: mean effect estimate for feature j

- j-th feature contribution
- j-th feature effect minus average j-th feature effect
- Contribution of feature X_j with value x_j
  minus the expected contribution of feature X_j
- Model prediction
  minus expected prediction if j-th feature value is not known
- Perturbs the value of the j-th feature only

Linear Model - Feature Contribution

$$\begin{aligned}\phi_j^{add}\left(\boldsymbol{x};\hat{f}\right) &= w_j x_j - \mathbb{E}\left[w_j X_j\right] \\ &= w_j\left(x_j - \mathbb{E}\left[X_j\right]\right) \\ &= \hat{f}(\boldsymbol{x}) - \mathbb{E}\left[\hat{f}(x_1,\ldots,X_j,\ldots,x_n)\right]\end{aligned} \tag{2}$$

where

- $\mathbb{E}(w_j X_j)$: mean effect estimate for feature j

- j-th feature contribution
- j-th feature effect minus average j-th feature effect
- Contribution of feature X_j with value x_j
  minus the expected contribution of feature X_j
- Model prediction
  minus expected prediction if j-th feature value is not known
- Perturbs the value of the j-th feature only
- Keeps the value of all other features

Linear Model - Feature Contribution

$$\begin{aligned}\phi_j^{add}\left(\boldsymbol{x};\hat{f}\right) &= w_j x_j - \mathbb{E}\left[w_j X_j\right] \\ &= w_j\left(x_j - \mathbb{E}\left[X_j\right]\right) \\ &= \hat{f}(\boldsymbol{x}) - \mathbb{E}\left[\hat{f}(x_1,\ldots,X_j,\ldots,x_n)\right]\end{aligned} \tag{2}$$

where

- $\mathbb{E}(w_j X_j)$: mean effect estimate for feature j

- j-th feature contribution
- j-th feature effect minus average j-th feature effect
- Contribution of feature X_j with value x_j
  minus the expected contribution of feature X_j
- Model prediction
  minus expected prediction if j-th feature value is not known
- Perturbs the value of the j-th feature only
- Keeps the value of all other features
- Independent w.r.t. the values of all the features but j

Linear Model - Feature Contribution

$$
\begin{aligned}
\phi_j^{add}\left(\boldsymbol{x}; \hat{f}\right) &= w_j x_j - \mathbb{E}\left[w_j X_j\right] \\
&= w_j \left(x_j - \mathbb{E}\left[X_j\right]\right) \\
&= \hat{f}(\boldsymbol{x}) - \mathbb{E}\left[\hat{f}(x_1, \ldots, X_j, \ldots, x_n)\right]
\end{aligned} \tag{2}
$$

where

- $\mathbb{E}(w_j X_j)$: mean effect estimate for feature j

- j-th feature contribution
- j-th feature effect minus average j-th feature effect
- Contribution of feature X_j with value x_j
  minus the expected contribution of feature X_j
- Model prediction
  minus expected prediction if j-th feature value is not known
- Perturbs the value of the j-th feature only
- Keeps the value of all other features
- Independent w.r.t. the values of all the features but j
- Situational importance of $X_j = x_j$ (Achen, 1982)

Linear Model - Total Feature Contribution

$$
\begin{aligned}
\sum_{j=1}^{p} \phi_j^{add}\left(\boldsymbol{x}; \hat{f}\right) &= \sum_{j=1}^{p} (w_j x_j - \mathbb{E}\left[w_j X_j\right]) \\
&= \left(w_0 + \sum_{j=1}^{p} w_j x_j\right) - \left(w_0 + \sum_{j=1}^{p} \mathbb{E}\left[w_j X_j\right]\right) \\
&= \hat{f}(\boldsymbol{x}) - \mathbb{E}\left[\hat{f}(X)\right]
\end{aligned} \tag{3}
$$

where

- $E(w_j X_j)$: mean effect estimate for feature j
- ϕ_j: j-th feature contribution
- $\boldsymbol{x}$: data instance
- x_j: value of feature j
- w_j: weight corresponding to feature j
- p: nr. features

Linear Model - Total Feature Contribution

$$\begin{aligned}\sum_{j=1}^{p}\phi_j^{add}\left(\boldsymbol{x};\hat{f}\right) &= \sum_{j=1}^{p}(w_jx_j - \mathbb{E}\left[w_jX_j\right]) \\ &= \left(w_0 + \sum_{j=1}^{p} w_jx_j\right) - \left(w_0 + \sum_{j=1}^{p}\mathbb{E}\left[w_jX_j\right]\right) \\ &= \hat{f}(\boldsymbol{x}) - \mathbb{E}\left[\hat{f}(X)\right]\end{aligned} \tag{3}$$

where

- $E(w_jX_j)$: mean effect estimate for feature j
- ϕ_j: j-th feature contribution
- $\boldsymbol{x}$: data instance
- x_j: value of feature j
- w_j: weight corresponding to feature j
- p: nr. features

- predicted value minus average predicted value

Feature Contribution in General - Example

- Prediction function: $\hat{f}(x_1, x_2) = x_1 \vee x_2$
- $x_1, x_2 \sim U(\{0, 1\})$, $\mathbb{E}_{X_1,X_2}\left[\hat{f}(X_1, X_2)\right] = \frac{3}{4}$

Feature Contribution in General - Example

- Prediction function: $\hat{f}(x_1, x_2) = x_1 \vee x_2$
- $x_1, x_2 \sim U(\{0, 1\})$, $\mathbb{E}_{X_1,X_2}\left[\hat{f}(X_1, X_2)\right] = \frac{3}{4}$

- E.g., input: $x_1 = 1$, $x_2 = 1$
- Prediction: $\hat{f}(1, 1) = 1$
- Gain: $\hat{f}(1, 1) - \mathbb{E}_{X_1,X_2}\left[\hat{f}(X_1, X_2)\right] = 1 - \frac{3}{4} = \frac{1}{4}$

Feature Contribution in General - Example

- Prediction function: $\hat{f}(x_1, x_2) = x_1 \vee x_2$
- $x_1, x_2 \sim U(\{0, 1\})$, $\mathbb{E}_{X_1,X_2}\left[\hat{f}(X_1, X_2)\right] = \frac{3}{4}$

- E.g., input: $x_1 = 1$, $x_2 = 1$
- Prediction: $\hat{f}(1, 1) = 1$
- Gain: $\hat{f}(1, 1) - \mathbb{E}_{X_1,X_2}\left[\hat{f}(X_1, X_2)\right] = 1 - \frac{3}{4} = \frac{1}{4}$

- Contribution of feature value $x_1 = 1$

$$\begin{aligned} \phi_1^{add}\left(x_1, x_2; \hat{f}\right) &= \hat{f}(1, 1) - \frac{1}{2}[\hat{f}(0, 1) + \hat{f}(1, 1)] \\ &= 1 \vee 1 - \frac{1}{2}[0 \vee 1 + 1 \vee 1] \\ &= 1 - \frac{1}{2}[1 + 1] = 0 \neq \frac{1}{8} \end{aligned} \tag{4}$$

Feature Contribution in General - Example

- Prediction function: $\hat{f}(x_1, x_2) = x_1 \vee x_2$
- $x_1, x_2 \sim U(\{0,1\})$, $\mathbb{E}_{X_1,X_2}\left[\hat{f}(X_1, X_2)\right] = \frac{3}{4}$

- E.g., input: $x_1 = 1$, $x_2 = 1$
- Prediction: $\hat{f}(1,1) = 1$
- Gain: $\hat{f}(1,1) - \mathbb{E}_{X_1,X_2}\left[\hat{f}(X_1, X_2)\right] = 1 - \frac{3}{4} = \frac{1}{4}$

- Contribution of feature value $x_1 = 1$

$$\begin{aligned}\phi_1^{add}\left(x_1, x_2; \hat{f}\right) &= \hat{f}(1,1) - \frac{1}{2}[\hat{f}(0,1) + \hat{f}(1,1)] \\ &= 1 \vee 1 - \frac{1}{2}[0 \vee 1 + 1 \vee 1] \\ &= 1 - \frac{1}{2}[1+1] = 0 \neq \frac{1}{8}\end{aligned} \tag{4}$$

- Same issue with contribution of feature value $x_2 = 1$

Feature Contribution in General - Example

- Prediction function: $\hat{f}(x_1, x_2) = x_1 \vee x_2$
- $x_1, x_2 \sim U(\{0,1\})$, $\mathbb{E}_{X_1,X_2}\left[\hat{f}(X_1, X_2)\right] = \frac{3}{4}$

- E.g., input: $x_1 = 1$, $x_2 = 1$
- Prediction: $\hat{f}(1,1) = 1$
- Gain: $\hat{f}(1,1) - \mathbb{E}_{X_1,X_2}\left[\hat{f}(X_1, X_2)\right] = 1 - \frac{3}{4} = \frac{1}{4}$

- Contribution of feature value $x_1 = 1$

$$\begin{aligned}\phi_1^{add}\left(x_1, x_2; \hat{f}\right) &= \hat{f}(1,1) - \frac{1}{2}[\hat{f}(0,1) + \hat{f}(1,1)] \\ &= 1 \vee 1 - \frac{1}{2}[0 \vee 1 + 1 \vee 1] \\ &= 1 - \frac{1}{2}[1+1] = 0 \neq \frac{1}{8}\end{aligned} \tag{4}$$

- Same issue with contribution of feature value $x_2 = 1$

- Perturbing one feature at a time gives undesirable results!

Feature Contribution in General

- *Can we do the same for any type of model?*
  - Model-agnostic
  - No assumptions on features interactions

Feature Contribution in General

- *Can we do the same for any type of model?*
  - Model-agnostic
  - No assumptions on features interactions

- Perturbing one feature at a time gives undesirable results
- Nonlinear models need a different solution

Feature Contribution in General

- *Can we do the same for any type of model?*
  - Model-agnostic
  - No assumptions on features interactions

- Perturbing one feature at a time gives undesirable results
- Nonlinear models need a different solution

- Possible solution: Shapley value
  - Field: cooperative game theory
  - Considers every subset of features
  - Perturbs all subsets of features

Cooperative Game Theory vs. Machine Learning - Terminology

- Game $\equiv$ (prediction) task
  - Single data point

Cooperative Game Theory vs. Machine Learning - Terminology

- Game $\equiv$ (prediction) task
  - Single data point

- Players $\equiv$ input feature values of a single data point
  - Collaborates

Cooperative Game Theory vs. Machine Learning - Terminology

- Game $\equiv$ (prediction) task
  - Single data point
- Players $\equiv$ input feature values of a single data point
  - Collaborates
- Coalition $=$ subset of players

Cooperative Game Theory vs. Machine Learning - Terminology

- Game $\equiv$ (prediction) task
  - Single data point
- Players $\equiv$ input feature values of a single data point
  - Collaborates
- Coalition $=$ subset of players
- (total) Payout $\equiv$ prediction value
  - Coalition-specific
  - Single data point

Cooperative Game Theory vs. Machine Learning - Terminology

- Game $\equiv$ (prediction) task
  - Single data point
- Players $\equiv$ input feature values of a single data point
  - Collaborates
- Coalition $=$ subset of players
- (total) Payout $\equiv$ prediction value
  - Coalition-specific
  - Single data point
- Gain $=$ specific payout minus average payout
  $\equiv$ Single prediction minus average prediction for all data points

Shapley Value - Intuition

- The average marginal contribution of a feature value over all possible coalitions.

Shapley Value - Intuition

- The average marginal contribution of a feature value over all possible coalitions.

- Shapley value for a feature j: average change in prediction that a subset of features receives when the feature j joins them.

Shapley Value - Feature Contribution

$$\phi_j(\boldsymbol{x}) = \frac{1}{p} \sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}} \binom{p-1}{|S|}^{-1} \left(val_{S\cup\{j\}}(\boldsymbol{x}) - val_S(\boldsymbol{x}) \right) \tag{5}$$

where

- j-th feature value
- $val_{\boldsymbol{x}}(S)$: value of players in S
- S: a subset of features used in the model (*coalition*)
- $\boldsymbol{x}$: vector of feature values of an instance to be explained
- p: nr. features

- Contribution of j-th feature value to the prediction (*payout*)

Shapley Value - Feature Contribution

$$\phi_j(\boldsymbol{x}) = \frac{1}{p} \sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}} \binom{p-1}{|S|}^{-1} \left(val_{S \cup \{j\}}(\boldsymbol{x}) - val_S(\boldsymbol{x}) \right) \tag{5}$$

where

- j-th feature value
- $val_{\boldsymbol{x}}(S)$: value of players in S
- S: a subset of features used in the model (*coalition*)
- $\boldsymbol{x}$: vector of feature values of an instance to be explained
- p: nr. features

- Contribution of j-th feature value to the prediction (*payout*)
- Normalized: weighted and summed over all possible feature combinations

Shapley Value - The Value Function - Example

$$val_{\{1,3\}}\left(\boldsymbol{x};\hat{f}\right) = \int_{\mathbb{R}}\int_{\mathbb{R}} \hat{f}(x_1, X_2, x_3, X_4) d\mathbb{P}_{X_2 X_4} - \mathbb{E}_X\left[\hat{f}(X)\right] \tag{6}$$

where

- $\{1, 3\}$: features in coalition
- $p = 4$: tot. model features
- $\boldsymbol{x} = (x_1, \cancel{x_2}, x_3, \cancel{x_4})$: data instance

Shapley Value - The Value Function

$$val_S\left(\boldsymbol{x};\hat{f}\right)=\int\hat{f}(x_1,\ldots,x_p)d\mathbb{P}_{x\notin S}-\mathbb{E}_X\left[\hat{f}(X)\right] \tag{7}$$

- Payout function for coalitions of players (feature values)
- Predicts feature values in S
- Marginalizes over features that are not in S
- Multiple integrations for each feature that is not in S

Shapley Value - The Value Function

$$val_S\left(\boldsymbol{x};\hat{f}\right)=\int\hat{f}(x_1,\ldots,x_p)d\mathbb{P}_{x\notin S}-\mathbb{E}_X\left[\hat{f}(X)\right] \tag{7}$$

- Payout function for coalitions of players (feature values)
- Predicts feature values in S
- Marginalizes over features that are not in S
- Multiple integrations for each feature that is not in S

- An empty coalition is worth zero

$$\begin{aligned}val_{\{\}}\left(\boldsymbol{x};\hat{f}\right)&=\int\hat{f}(\boldsymbol{x})d\mathbb{P}_{\boldsymbol{x}}-\mathbb{E}_X\left[\hat{f}(X)\right]\\&=\mathbb{E}_X\left[\hat{f}(X)\right]-\mathbb{E}_X\left[\hat{f}(X)\right]=0\end{aligned} \tag{8}$$

Shapley Value - Feature Contribution - Example

- Prediction function: $\hat{f}(x_1, x_2) = x_1 \vee x_2$
- $x_1, x_2 \sim U(\{0, 1\})$, $\mathbb{E}_{X_1, X_2}\left[\hat{f}(X_1, X_2)\right] = \frac{3}{4}$

Shapley Value - Feature Contribution - Example

- Prediction function: $\hat{f}(x_1, x_2) = x_1 \vee x_2$
- $x_1, x_2 \sim U(\{0, 1\})$, $\mathbb{E}_{X_1,X_2}\left[\hat{f}(X_1, X_2)\right] = \frac{3}{4}$

- E.g., input: $x_1 = 1$, $x_2 = 1$
- Prediction: $\hat{f}(1, 1) = 1$
- Gain: $\hat{f}(1, 1) - \mathbb{E}_{X_1,X_2}\left[\hat{f}(X_1, X_2)\right] = 1 - \frac{3}{4} = \frac{1}{4}$

Shapley Value - Feature Contribution - Example

- Prediction function: $\hat{f}(x_1, x_2) = x_1 \vee x_2$
- $x_1, x_2 \sim U(\{0,1\})$, $\mathbb{E}_{X_1,X_2}\left[\hat{f}(X_1, X_2)\right] = \frac{3}{4}$

- E.g., input: $x_1 = 1$, $x_2 = 1$
- Prediction: $\hat{f}(1,1) = 1$
- Gain: $\hat{f}(1,1) - \mathbb{E}_{X_1,X_2}\left[\hat{f}(X_1, X_2)\right] = 1 - \frac{3}{4} = \frac{1}{4}$

- Contribution of feature value $x_1 = 1$

$$\begin{aligned}
\phi_1\left(x_1, x_2; \hat{f}\right) &= \frac{1}{2}\Big[val_{\{1\}}\left(\boldsymbol{x}; \hat{f}\right) - val_{\{\}}\left(\boldsymbol{x}; \hat{f}\right) \\
&\qquad + val_{\{1,2\}}\left(\boldsymbol{x}; \hat{f}\right) - val_{\{2\}}\left(\boldsymbol{x}; \hat{f}\right)\Big] \\
&= \frac{1}{2} val_{\{1,2\}}\left(\boldsymbol{x}; \hat{f}\right) \\
&= \frac{1}{2}\left(\hat{f}(1,1) - \mathbb{E}_{X_1,X_2}\left[\hat{f}(X_1, X_2)\right]\right) = \frac{1}{2}\left[1 - \frac{3}{4}\right] = \frac{1}{8}
\end{aligned} \tag{9}$$

  - Perturbs all subsets of features

Shapley Value - Feature Contribution - Example

- Prediction function: $\hat{f}(x_1, x_2) = x_1 \vee x_2$
- $x_1, x_2 \sim U(\{0,1\})$, $\mathbb{E}_{X_1,X_2}\left[\hat{f}(X_1, X_2)\right] = \frac{3}{4}$

- E.g., input: $x_1 = 1$, $x_2 = 1$
- Prediction: $\hat{f}(1,1) = 1$
- Gain: $\hat{f}(1,1) - \mathbb{E}_{X_1,X_2}\left[\hat{f}(X_1, X_2)\right] = 1 - \frac{3}{4} = \frac{1}{4}$

- Contribution of feature value $x_1 = 1$

$$\begin{aligned}\phi_1\left(x_1, x_2; \hat{f}\right) &= \frac{1}{2}\Big[val_{\{1\}}\left(\boldsymbol{x}; \hat{f}\right) - val_{\{\}}\left(\boldsymbol{x}; \hat{f}\right) \\ &\qquad + val_{\{1,2\}}\left(\boldsymbol{x}; \hat{f}\right) - val_{\{2\}}\left(\boldsymbol{x}; \hat{f}\right)\Big] \\ &= \frac{1}{2} val_{\{1,2\}}\left(\boldsymbol{x}; \hat{f}\right) \\ &= \frac{1}{2}\left(\hat{f}(1,1) - \mathbb{E}_{X_1,X_2}\left[\hat{f}(X_1, X_2)\right]\right) = \frac{1}{2}\left[1 - \frac{3}{4}\right] = \frac{1}{8}\end{aligned} \tag{9}$$

  - Perturbs all subsets of features
- Same contribution for feature value $x_2 = 1$

Shapley Value - Exact Estimation

- All possible subsets (coalitions) of feature values have to be evaluated with and without the j-th feature.

- The number of possible coalitions increases exponentially as the the number of features increases.

Shapley Value - Approximation[4]

$$\hat{\phi}_j \left(\boldsymbol{x}; \hat{f} \right) = \frac{1}{M} \sum_{m=1}^{M} \left(\hat{f} \left(\boldsymbol{x}_{+j}^{m} \right) - \hat{f} \left(\boldsymbol{x}_{-j}^{m} \right) \right) \tag{10}$$

- Monte-Carlo Sampling
- $\hat{f} \left(\boldsymbol{x}_{+j}^{m} \right)$
  - prediction for a data point $\boldsymbol{x}^m$
  - random number of features replaced by feature values from a random data point $\boldsymbol{z}^m$.
  - uses the feature value x_j^m
- $\hat{f} \left(\boldsymbol{x}_{-j}^{m} \right)$
  - like $\hat{f} \left(\boldsymbol{x}_{+j}^{m} \right)$
  - uses the random feature value z_j^m

[4] Štrumbelj and Kononenko (2014)

Shapley Value - Properties

1. Dummy
2. Efficiency
3. Symmetry
4. Additivity

- Axioms
- Fair Payout

Shapley Value - Dummy

If

$$val_{S \cup \{j\}}(\cdot) = val_S(\cdot) \tag{11}$$

for all

$$S \subseteq \{1, \ldots, p\}$$

then

$$\phi_j = 0$$

- A feature j that does not change the predicted value - regardless of which coalition of feature values it is added to - should have a Shapley value of 0.

Shapley Value - Efficiency

$$\sum_{j=1}^{p} \phi_j \left(\boldsymbol{x}; \hat{f} \right) = \hat{f}(\boldsymbol{x}) - \mathbb{E}_X \left[\hat{f}(X) \right] \tag{12}$$

- Feature contributions must *sum* up to prediction for $\boldsymbol{x}$ *minus* average prediction

Shapley Value - Symmetry

If

$$val_{S\cup\{j\}}(\cdot) = val_{S\cup\{k\}}(\cdot) \tag{13}$$

for all

$$S \subseteq \{1,\ldots,p\}\backslash\{j,k\}$$

then

$$\phi_j = \phi_k$$

- The contribution of two feature values j and k should be the same, if they contribute equally to all possible coalitions.

Shapley Value - Additivity

If

$$val_S(\boldsymbol{x} + \boldsymbol{y}) = val_S(\boldsymbol{x}) + val_S(\boldsymbol{y}) \tag{14}$$

for all

$$\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X},\ S \subseteq \{1, \ldots, p\}$$

then

$$\phi(\boldsymbol{x} + \boldsymbol{y}) = \phi(\boldsymbol{x}) + \phi(\boldsymbol{y})$$

- Combined payouts
- Example: Random forest = average of many decision trees
  - Prediction = average prediction in decision trees
  - Feature contribution = average feature contribution in decision trees

Software

- `fastshap` (R) (Jethani et al., 2021)
- `iml` (R) (Molnar et al., 2018)
- `breakDown` (R) (Staniak and Biecek, 2018)
- `Shapley.jl` (Julia) [5]

[5] https://gitlab.com/ExpandingMan/Shapley.jl

Shapley Value in Short

- Permutation-based
- Model-agnostic
- Solid theory
- Full-explanation
  - All the features
  - Non sparse (proper subset of features)
- Model-free
- Data access or generation
- Building block of SHAP (Lundberg and Lee, 2017)

References I

Achen, Christopher H (1982). *Interpreting and using regression*. Vol. 29. Sage.

Fanaee-T, Hadi (2013). *Bike Sharing Dataset*. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5W894.

Fernandes, Kelwin, Jaime Cardoso, and Jessica Fernandes (2017). *Cervical cancer (Risk Factors)*. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5Z310.

Jethani, Neil et al. (2021). "Fastshap: Real-time shapley value estimation". In: *International Conference on Learning Representations*.

Lundberg, Scott M and Su-In Lee (2017). "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30.

Molnar, Christoph (2020). *Interpretable machine learning*. Lulu. com.

Molnar, Christoph, Giuseppe Casalicchio, and Bernd Bischl (2018). "iml: An R package for interpretable machine learning". In: *Journal of Open Source Software* 3.26, p. 786.

References II

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). “”' Why should i trust you?” Explaining the predictions of any classifier”. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144.

Staniak, Mateusz and Przemyslaw Biecek (2018). “Explanations of model predictions with live and breakDown packages”. In: *arXiv preprint arXiv:1804.01955*.

Štrumbelj, Erik and Igor Kononenko (2014). “Explaining prediction models and individual predictions with feature contributions”. In: *Knowledge and information systems* 41, pp. 647–665.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan (2017). “Axiomatic attribution for deep networks”. In: *International conference on machine learning*. PMLR, pp. 3319–3328.