SHAP Analysis:

Shapley Additive Explanation Analysis

How does prediction change when feature is removed from the model?

Romen Samuel Wabina, MSc.
MSc Data Science & Artificial Intelligence
Ongoing PhD Data Science *in Healthcare and Clinical Informatics*

Recall: Properties of **Shapley Values**

General Properties.
The Shapley Value is the only attribution method that satisfies the properties **Efficiency, Symmetry, Dummy,** and **Additivity**, which together can be considered a definition of a fair payout.

Property 1. ***Efficiency***
The feature contributions must add up to the difference of prediction for $\boldsymbol{x}$ and the average.

$$\sum_{j=1}^{p} \phi_j = \hat{f}(x) - \mathbb{E}_X\left(\hat{f}(X)\right)$$

Property 2. ***Symmetry***
The contributions of two feature values $\boldsymbol{j}$ and $\boldsymbol{k}$ should be the same if they contribute equally to all possible coalitions.

$$\sum_{j=1}^{p} \phi_j = \hat{f}(x) - \mathbb{E}_X\left(\hat{f}(X)\right)$$

Properties Uniquely Determine **Additive Feature Attributions**

Lundberg and Lee (2017) suggest that methods not based on Shapley values violates the unique properties that can determine additive feature attributions (i.e., *Local accuracy, Missingness, and Consistency*).

Therefore, the authors proposed a unified approach that improves existing interpretability methods that violates these properties.

Definition 1. Additive feature attribution methods have an explanation model that is a **linear function** of binary variables:

$$g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x_i'$$

where $x' \in \{0,1\}^M$, M is the number of simplified input features, and $\phi_i \in \mathbb{R}$.

Theorem 1. Only one possible explanation model follows Definition 1 and satisfies Properties 1, 2, and 3.

$$\phi_i(f,x) = \sum_{z' \in x'} \frac{|z'|!\,(M - |z'| - 1)!}{M!} \left[f_x(z') - f_x\left(\frac{z'}{i}\right) \right]$$

Properties Uniquely Determine **Additive Feature Attributions**

Property 1. **Local accuracy.**
It requires the explanation model g to at least match the output of the original model f for the simplified input x' when approximating f for a specific input x.

$$g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x_i'$$

$$\hat{f}(x) = g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x_i' = \mathbb{E}_X\left(\bar{f}(X)\right) + \sum_{j=1}^{M} \phi_j$$

It means that the explanation of model $g(x')$ should match the original model $f(x)$ in the situation when $x = h_x(x')$ and $f\left(h_x(0)\right)$ represent the model output for which all the inputs are missing.

The sum of all the contributions should be equal [or approximately equal] to the prediction. In other words, the SHAP's local accuracy property that ensures that the contributions of the features should add up to the difference between the prediction and the average prediction of the model.

Properties Uniquely Determine **Additive Feature Attributions**

Property 2. **Missingness**
The Missingness property enforces that the missing features get a Shapley value of 0. A missing feature could have an *arbitrary* Shapley Value without hurting the local accuracy property since it is multiplied with $x'_j = 0$

$$x'_j = 0 \Rightarrow \phi_j = 0$$

In other words, $x'_j = 0$ have no attributed contribution on the model.

Properties Uniquely Determine **Additive Feature Attributions**

Property 3. **Consistency**

The consistency property states that if a model changes in such a way that the *marginal contribution* of a feature value increases the same [or stays the same], regardless of other features, the Shapley Value also increases [or stays the same.

Consistency indicates that if a model changes such that some feature's contribution increases or stays the same regardless of the other inputs, that input's attribution should not decrease.

Let $f_x(x') = f\big(h_x(x')\big)$ and $z_{j'}$ indicate that $z_{j'} = 0$. For any two models f and f' that satisfy:

$$f'_x(z') - f'_x\big(z'_j\big) \geq f_x(z') - f_x\big(z'_j\big)$$

for all inputs $z' \in \{0,1\}^M$, then

$$\phi_j(f', x) \geq \phi_j(f, x)$$

"How does prediction change when feature is removed from the model?"

SHAP Plots: ***Waterfall Plot***

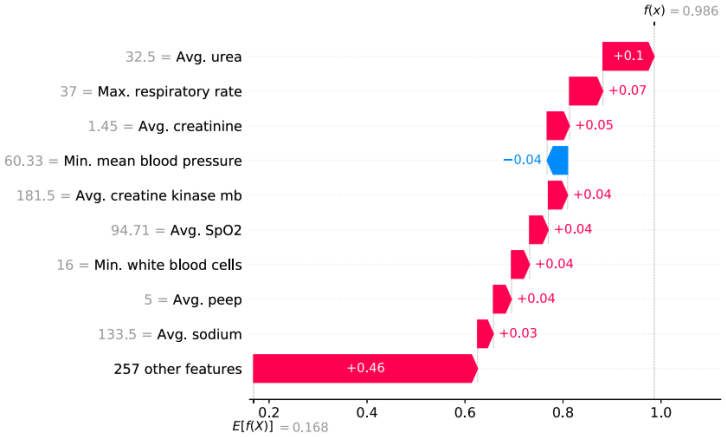
Plot 6. **Waterfall Plot.**

The waterfall plot is designed to visually display how the SHAP values of each feature *move* the model output from our prior expectation under the background data distribution, to the final model prediction given the evidence of all the features.

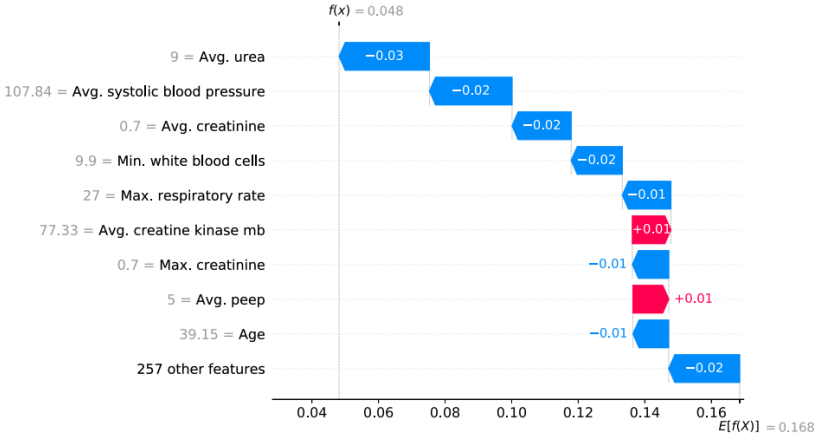
The purpose of the waterfall plot is to *visualize how the SHAP value for each feature in the output model changes from the previous prediction* on the background data set to the final model prediction when the evidence for all features is presented.

SHAP Plots: *Waterfall Plot*

Plot 6. **Waterfall Plot.**
The waterfall plot is designed to visually display how the SHAP values of each feature *move* the model output from our prior expectation under the background data distribution, to the final model prediction given the evidence of all the features.



(c) Man with a high risk



(d) Man with a low risk

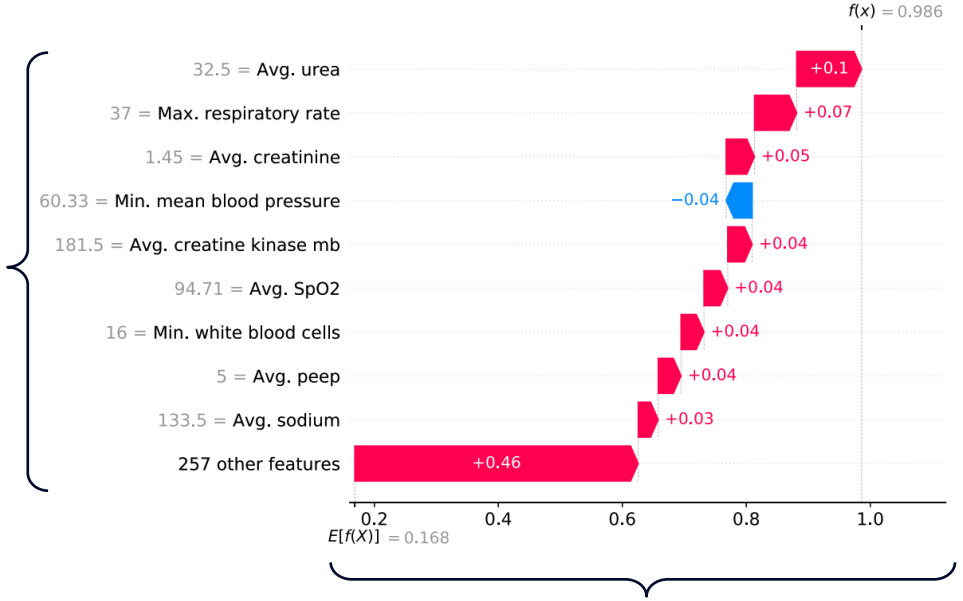
Vazquez, B., Fuentes-Pineda, G., Garcia, F., Borrayo, G., & Prohias, J. (2021). Risk markers by sex for in-hospital mortality in patients with acute coronary syndrome: a machine learning approach. Informatics in Medicine Unlocked, 27, 100791.

SHAP Plots: *Waterfall Plot*

The rows in the y-axis are the most important features (ranked in descending order from top to bottom).

Each subsequent row up the y-axis shows how the positive (red) or negative (blue) contribution of each feature moves the value from the expected model output over the background dataset to the model output for this prediction.

The SHAP values push the output to the left or right of $\mathbb{E}[f(x)]$ over the x-axis and increase or decrease the model's output value.



The x-axis of a waterfall plot displays the expected value of the model output $E[f\ (X)]$. The models' output is presented in logarithm of odds (log-odds) units

Vazquez, B., Fuentes-Pineda, G., Garcia, F., Borrayo, G., & Prohias, J. (2021). Risk markers by sex for in-hospital mortality in patients with acute coronary syndrome: a machine learning approach. Informatics in Medicine Unlocked, 27, 100791.