

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 x and the average.

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

Property 2. Symmetry

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

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

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 .

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

Property 2. Missingness

The Missingness property enforces that missing features get a Shapley value of 0.

This property is not among the properties of the *normal* Shapley values. So why do we need it for SHAP?

A missing feature could – in theory – have an arbitrary Shapley value without hurting the local accuracy property, since it is multiplied with $x'_j = 0$.

Properties Uniquely Determine **Additive Feature Attributions**

Property 3. Consistency

The consistency property says that if a model changes so that the marginal contribution of a feature value increases or stays the same (regardless of other features), the Shapley value also increases or stays the same.

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

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.

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

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

SHAP Plots: *Waterfall Plot*

Plot 6. Waterfall Plot.

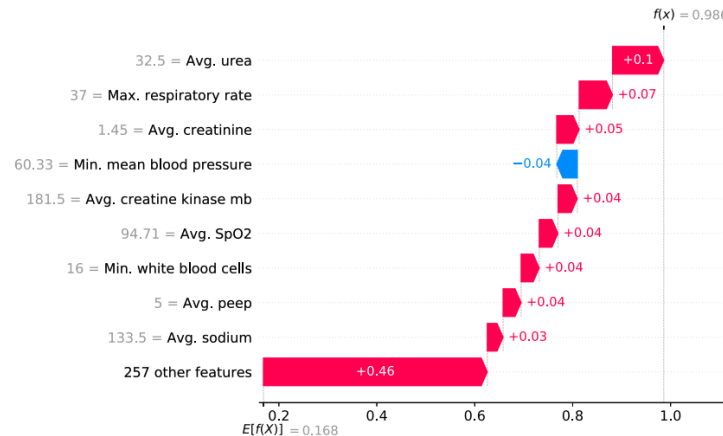
The waterfall plot is designed to visually display how the SHAP values of each feature move the model output from our prior expectation under the background data distribution, to the final model prediction given the evidence of all the features.

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.

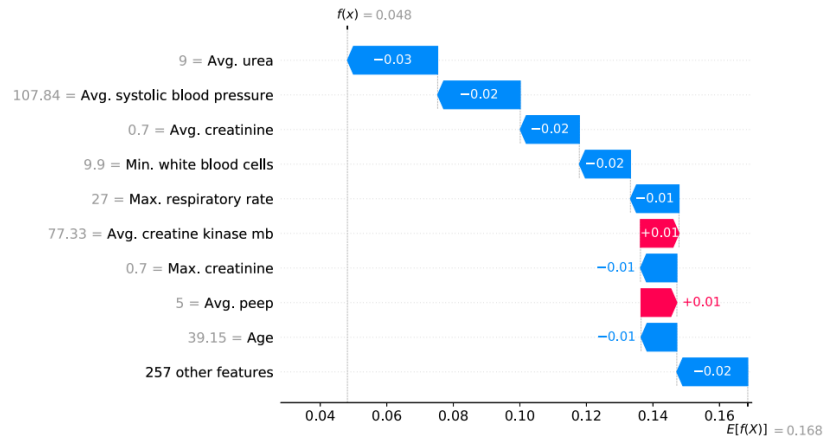
SHAP Plots: *Waterfall Plot*

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



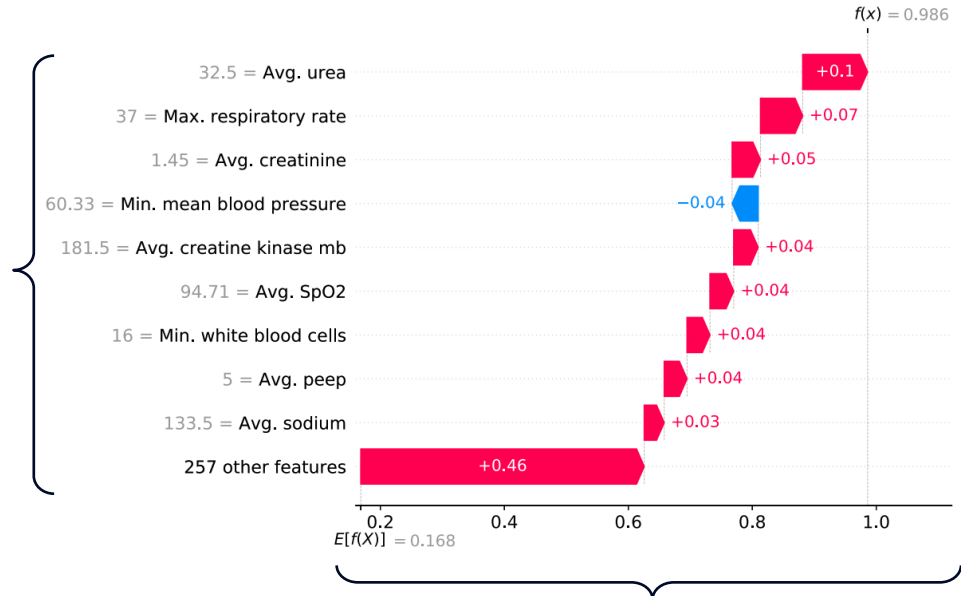
(d) Man with a low risk

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