- **Model interpretation: SHAP value**

The SHAP value is a method to decompose the prediction into additive feature attribution:

$$y_i = y_0 + f(x_{i1}) + f(x_{i2}) + \cdots + f(x_{ik})$$

where y_i is the prediction for sample i. y_0 is the baseline value which equals to the average prediction of all training samples. The linear equation enables users to calculate the contributions of the k_{th} feature compared with y_0. The Shapely value $f(x_{ik})$ indicates the expected marginal contribution of feature k across all orderings of input variables:

$$f(x_k) = \sum_{S \subseteq \{1,\ldots,n\} \setminus \{k\}} \frac{|S|!\,(n - |S| - 1)!}{n!} \left(val(S \cup \{k\}) - val(S)\right)$$

where n is the total number of features. S is a subset of the features used in the model except for k. $|S|$ is the number of features in subset S. $val(S \cup \{k\})$ and $val(S)$ are the predictions with and without variable k. $val(S \cup \{k\})$ and $val(S)$ are usually calculated as conditional expectations from the training dataset as most ML models can't handle missing features. We adopted the SHAP algorithm designed for tree-based models to calculate the contributions of local covariates. It should be noted that there are other SHAP approximation methods (e.g., Kernel SHAP and Deep SHAP). We prefer the tree-based method because it calculates accurate SHAP value in polynomial time instead of exponential time when the feature numbers increase.

Ref:

1. Lundberg, S.M., Erion, G.G., Lee, S.-I., 2018. Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888.

2. Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems 30.