### Shapley values

SHAP (SHapley Additive exPlanations) is an interpretability tool used for "open the black box". It is released by Lundberg and Lee (2016) to explain the prediction of each instance due computing the contribution of each feature (for this thesis it is input variable) to the prediction. SHAP is constructed by the concept of the Shapley values, a method from coalitional game theory that is built to fairly distribute a "pay-out" from a multi-player game (Woloszko, 2020; Molnar, 2020). In this matter, the "game" is a single instance prediction task, the "players" are the predictors, and the "pay-out" equals actual this instance prediction minus the average prediction for all instances (Molnar, 2020). The Shapley value is the average marginal contribution of an estimator value to the prediction across all possible coalitions, which is designed as a number of predictors taking the value that is noticed rather than their average or any arbitrary value. The model prediction is added up by Shapley values. In the mathematical term, for simple linear model prediction described by:

|  |  |  |
| --- | --- | --- |
|  |  | *(3.4)* |

Where is the instance which is desired to compute the contributions. is a feature value, with j = 1,…,p. The is the weight corresponding to predictor j. The contribution of the j-th feature on the prediction is:

|  |  |  |
| --- | --- | --- |
|  |  | *(3.5)* |

Where represents the mean impact estimate for feature j. The contribution equals the feature impact minus the average impact. The following is the result when taking sum all the feature contributions for one instance:

|  |  |  |
| --- | --- | --- |
|  |  | *(3.6)* |

|  |  |  |
| --- | --- | --- |
|  |  | *(3.7)* |

As Woloszko (2020) said, Shapley value is the only ascription technique that associate the following properties: efficiency (Shapley values sum to the prediction minus its average), symmetry (variables which have the Shapley value should have equal contribution to all coalition), dummy (a variable value with no impact on the prediction whatever the coalition has Shapley value equal to zero) and additivity. The considerable value of this method is the tendency to provide both local interpretability (explaining given predictions) and global interpretability (the general functioning of a model).