SHAP

SHapley Additive exPlanations

- SHAP is a mathematical concept from the field of economics and Game Theory which was developed in the 50's. In it's modern use, it is used to explain how each feature of an ML Model contributed to the prediction. 1
- SHAP and LIME are methods to understand feature importance within a prediction of a ML Model. They expected to give similar results. [2]
- SHAP Explainer (model, X) function outputs .values, .base_values and .data. Values reflects shap values, .data reflects the training set (X), and .base_value reflects the model output without any features (probability of getting a specific label in the train set, regardless of having a model). Then, SHAP values are analyzed in relation to the base_value, where big deviation

(negative/positive) from it means big importance for the classification. [2]

- Positive SHAP value of a feature indicate it increases prediction value, while negative SHAP indicate it decreases prediction value. SHAP = 0 has no impact on prediction outcome. Therefore, absolute SHAP values reflect the importance of each feature for the prediction. [1,3]
- Feature variance have a big impact on SHAP values, where smaller variance will tend to present lower SHAP's. [4]
- Variance in SHAP values of a particular feature in a beesswarm plot could suggest that the feature has high dependence on other features that the model were trained on. [1]
- The SHAP interaction scores gives the paired interaction between each feature with all other features in the train set. Positive SHAP interaction score would mean that the variables share the same SHAP trend, and negative value will indicate they have opposite SHAP trends. The interaction

score can vary between different feature values (different samples).

- Notice that SHAP values that are dependent could give a certain feature a **negative** interaction score with another feature, although they both has **high positive SHAP** value.

1. SHAP Excercise AI HUB
2. https://svitla.com/blog/interpreting-machine-learning-models-lime-and-shap#Why%20interpreting%20models%20is%20important.
3. ChatGPT
4. https://www.data-cowboys.com/blog/shap-values-and-feature-variance