Understanding Interactions in Statistical Models: Unravelling Complex Relationships with SHAP

Guillermo Dominguez[1], Minh Tuan Nguyen[1], Joana Levtcheva[1] and Mays Azeez[1]

[1] Imperial College London

Keywords: *Interaction Effects,* Statistical Models, *Shapley Values,* Model Interpretability, *SHAP Python Package.*

The rapid adoption of machine learning in various fields, from tech companies to clinical trials, has ignited ethical debates, primarily around model interpretability. A crucial aspect often overlooked is the role of interactions among predictors, i.e., the mutual dependence of variables in shaping the response variable. Ignoring these interactions can lead to misinterpretation of data, with potentially severe consequences, particularly in sensitive domains such as healthcare. This study aims to define and classify interactions, and provide an overview of methods to detect and interpret them, focusing on Shapley values and SHAP. Linear models without interaction terms assume that the effect of each predictor on the response variable is independent, an assumption often too simplistic to capture complex phenomena. Introducing interaction terms allows capturing joint effects of multiple predictors on the response variable, leading to a more nuanced model. Interactions can be qualitatively classified as positive or negative, or based on variable types into continuous-continuous, categorical-categorical, and categorical-continuous. Shapley values, rooted in game theory, can measure the contribution of each predictor to the model outcome considering all possible feature combinations. This approach further extends to the Shapley interaction index, offering a measure of combined feature effects beyond individual contributions. We illustrate these concepts with a clinical study on the interaction effects of dose and a genetic variant on treatment response in rheumatoid arthritis, employing the SHAP Python package for analysis and interpretation. This work underscores the critical role of understanding interactions in statistical models to unravel complex relationships, enhancing model interpretability, and thus, decision-making.