Using SHAP Post-Model Explainability as a Model-Agnostic Feature Selection Technique

Joshua Gottlieb[1], Martin Sichali Chunsen[1], and Alexander Yoo[1], advised by Dr. Krishna Bathula[1]
[1]Seidenberg School of CSIS, University of Pace, New York City, NY, USA
jg05394n@pace.edu, cs83339n@pace.edu, ay71173n@pace.edu, kbathula@pace.edu

Feature selection techniques in data science help reduce dimensionality of datasets, remove sparsity, and eliminate noisy patterns, thereby improving generalization of models to future data. Many feature selection techniques fail to optimize for the predictive power of the model, require iterating through the extensive feature space, or apply only to certain model types [1]. SHAP values focus on capturing the predictive power of each feature within the context of the trained model in order to provide explainability. Because SHAP treats the model as a black box, the SHAP technique is model agnostic. In this paper, we employ SHAP values to capture and rank the important features learned by a model and to use these rankings for feature selection. We build upon prior work using SHAP as a feature selection technique [2] by creating the MAX and SUM selection algorithms, which are designed to align with intuitive feature selection goals without the requirement for costly iterative processes. We validate the effectiveness of our SUM and MAX SHAP selection algorithms by testing their performance with five machine learning models using ten diverse datasets, demonstrating that these SHAP selection strategies produce equivalent or superior model performance and greater feature reduction compared to other state-of-the-art techniques.

Algorithm 1
SUM SHAP Pseudocode

Sort the mean absolute SHAP values I_j in descending order: $I_j \geq I_{j+1}$
Calculate the total sum $S = \sum I_j$
Define a proportion constant r in [0, 1].
Initialize a running sum s and an index $j = 0$.
while $s < r \cdot S$ **do**
  Add I_j to s and increment j
end while
return $[f_0, \ldots, f_j]$, the selected sorted features.

Algorithm 2
MAX SHAP Psuedocode

Sort the mean absolute SHAP values I_j in descending order: $I_j \geq I_{j+1}$
Choose $M = \max(I_j) = I_0$
Define a proportion constant ρ in [0, 1].
Initialize an index $j = 0$.
while $I_j \geq \rho \cdot M$ **do**
  Increment j
end while
return $[f_0, \ldots, f_j]$, the selected sorted features.

References

[1] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," Computers & Electrical Engineering, vol. 40, no. 1, pp. 16–28, 2014, 40th-year commemorative issue.

[2] W. E. Marcílio and D. M. Eler, "From explanations to feature selection: Assessing SHAP values as feature selection mechanism," in Proc. 33rd SIBGRAPI Conf. Graph., Patterns Images (SIBGRAPI), Porto de Galinhas, Brazil, 2020, pp. 340–347, doi: 10.1109/SIBGRAPI51738.2020.00053.