Improving Explainable AI in Machine Learning Models Using SHAP

1 Introduction

~~As **artificial intelligence (AI)** advances, smart machines are increasingly integrated into our daily lives.~~ Recommendation systems, text analytics, autonomous vehicles, medical diagnostics, and other AI-driven technologies are shaping critical aspects of society as we place our trust in these tools. However, as these systems grow more complex, understanding a model's decision-making process is essential for debugging errors, detecting bias, ensuring transparency, and complying with regulations.

Explainable AI (XAI) aims to make AI decision-making more transparent by providing methods to interpret and explain model outputs. Current literature reveals two distinct XAI approaches: **machine learning (ML)** and **user experience (UX)**.

The ML approach utilizes various tools to reveal how a model arrives at its conclusions. The field aims to improve model performance, identify and mitigate bias, debug errors, and build industry trust in AI systems.

The UX approach focuses on how users perceive and interact with AI explanations, drawing from psychology and **human-computer interaction (HCI)**. This approach evaluates whether a model's output is interpretable (users can understand it) and explainable (users can predict how changes in input affect output). The primary aim of this approach is to foster user trust and reduce perceived risk.

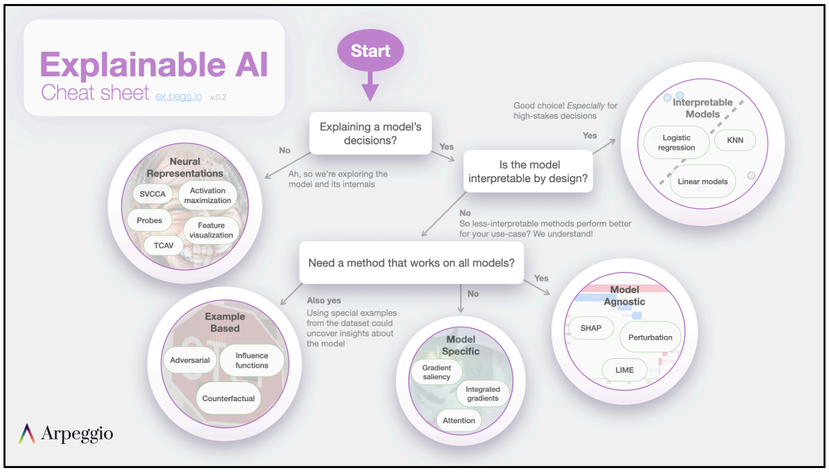
Explainable AI
Cheat sheet ex.pegg.io v.0.2
Start
Explaining a model's decisions?
No
Ah, so we're exploring the model and its internals
Yes
Is the model interpretable by design?
Good choice! Especially for high-stakes decisions
Yes
Interpretable Models
Logistic regression
KNN
Linear models
No
So less-interpretable methods perform better for your use-case? We understand!
Need a method that works on all models?
Neural Representations
SVCCA
Activation maximization
Probes
Feature visualization
TCAV
Also yes
Using special examples from the dataset could uncover insights about the model
Example Based
Adversarial
Influence functions
Counterfactual
No
Model Specific
Gradient saliency
Integrated gradients
Attention
Yes
Model Agnostic
SHAP
Perturbation
LIME
Arpeggio

Figure 1.[1] *A flowchart depicting Explainable AI (XAI) methods based on model interpretability and other characteristics. SHAP is a model-agnostic tool that explains AI decisions and can be applied to less interpretable models.*

[1] Alammar, Jay. 2021. "Explainable AI Guide." Pegg.io. 2021. https://ex.pegg.io/.

Various machine learning techniques in the field of XAI can be categorized based on model features and characteristics (Figure 1). SHAP (SHapley Additive exPlanations) is a Python, open-source, XAI tool. SHAP is model-agnostic, meaning it can be applied to any machine learning model, regardless of its architecture or underlying algorithm. This study explores SHAP's role in improving machine learning interpretability, highlighting its benefits and limitations.

~~*Outline*~~

~~The remainder of this paper is structured as follows. *Section 2* introduces SHAP as an explainability method derived from Shapley values in coalitional game theory. *Section 3* discusses the desirable mathematical properties of SHAP, such as efficiency, symmetry, dummy, and linearity, and highlights SHAP's ability to provide both local and global interpretability, making it a versatile tool in XAI. *Section 4* examines the challenges associated with SHAP, including computational complexity, difficulties with high-dimensional data, sensitivity to correlated features, and the risk of misinterpretation. *Section 5* discusses future research, and outlines areas for further investigation, including SHAP's comparison with other feature attribution methods and its applicability to non-numeric data. The article concludes in *Section 6*.~~

2 SHAP Overview

SHAP is an explainability method based on Shapley values in coalitional game theory.[2] Coalitional game theory, also known as cooperative game theory, is a model that describes how groups of players, or coalitions, work together.

Let's look at an analogy:

> Imagine you're at a potluck dinner where each guest brings a dish. The overall meal enjoyment depends on the combination of dishes. Some dishes, like a well-seasoned main course, might have a greater impact on the meal's success, while others, like a simple side dish, contribute less. How do we determine how much each guest contributed to the overall meal enjoyment?

In this analogy, each guest represents a *feature* (variable), and their dish is the *contribution.* The overall meal satisfaction is the *prediction.* Shapley values reveal how much each guest contributed to meal enjoyment by evaluating different combinations of dishes and the resulting meal enjoyment. In other words, Shapley values quantify how each feature contributes to the overall prediction, and how that prediction changes when joined to every possible combination of other features.[3]

[2] Lundberg, Scott. 2018. "Shap." GitHub. June 28, 2018. https://github.com/shap/shap.

[3] Pawlicki, Marek, Aleksandra Pawlicka, Federica Uccello, Sebastian Szelest, Salvatore D'Antonio, Rafał Kozik, and Michał Choraś. 2024. "Evaluating the Necessity of the Multiple Metrics for Assessing Explainable AI: A Critical Examination." *Neurocomputing* 602 (October): 128282–82. https://doi.org/10.1016/j.neucom.2024.128282.

Feature contribution refers to how much a specific feature contributes to a single prediction whereas *feature importance* is a global measure that summarizes the overall impact of a feature across many predictions.

By considering all possible combinations of features, Shapley values provide a holistic view of how each feature influences the model's prediction.

3 Mathematical Properties and Strengths

SHAP as an XAI model has numerous strengths. In addition to providing a comprehensive view of feature importance, Shapley values are considered the "definition of a fair weight"[4] due to their mathematically desirable properties[5]:

1. *Efficiency.* The sum of the Shapley values of all features equals the value of the prediction made with all the features, ensuring that the prediction is fairly distributed among them.
2. *Symmetry.* If two feature values contribute equally to all possible coalitions, their Shapley value is the same.
3. *Dummy.* A feature that does not change the predicted value, regardless of which features are included, have a Shapley value of 0.
4. *Linearity.* If two models are combined, the prediction should be distributed according to each model's contributions.

These properties ensure that SHAP provides a fair, consistent, and efficient method for explaining model predictions.

As a model-agnostic explainability method, SHAP can be applied to any machine learning model, including enigmatic black-box models like deep neural networks, to provide insights into the model's inner workings despite their complexity.[4]

Another key advantage of SHAP is the method's ability to provide both local explanations and global insights. Local explanations reveal how the model made individual decisions, such as whether or not to approve a loan. Global insights average Shapley values across various individual decisions to reveal the model's overall behavior and key features that drive predictions. The information gained from global analysis can be used to fine-tune the model and address potential biases.[4]

[4] Aboze, Brain John. 2023. "A Comprehensive Guide into SHAP Values." Deepchecks. May 16, 2023. http://deepchecks.com/a-comprehensive-guide-into-shap-shapley-additive-explanations-values/.

[5] Fadel, Soufiane, and Statistics Canada. 2022. "Explainable Machine Learning, Game Theory, and Shapley Values: A Technical Review." Www.statcan.gc.ca. February 28, 2022. https://www.statcan.gc.ca/en/data-science/network/explainable-learning.

4 Limitations

The main limitation of SHAP is the exponential growth rate of computational complexity with each feature added to a model.[6] Approximation methods, such as Monte Carlo sampling, KernelSHAP, TreeSHAP, and DeepSHAP attempt to accurately estimate Shapley values using fewer computational resources.

While Shapley values provide holistic insight into a model's decision making process, Shapley values, and Shapley approximation methods, are prone to the following limitations:

1. ***Challenges with high-dimensional data.*** A high-dimensional dataset is one with many features. SHAP can become computationally infeasible with high-dimensional data, limiting its ability to provide accurate and timely explanations.[4]

2. ***Interpretability vs. Complexity (Accuracy) trade-off.*** SHAP offers local interpretability at the cost of global accuracy. It can approximate each feature's contribution, but may not accurately reflect the model's global, intricate relationships.[4]

3. ***Sensitivity to correlated features.*** SHAP assumes that features are independent, but real-world datasets often contain correlated variables. As a result, SHAP cannot be reliably used for causal inference.[7]

4. *Subject to human error*. Accurate interpretation of SHAP values requires domain expertise. Without it, there is a risk of misinterpretation, confirmation bias, and misleading conclusions.[7]

A 2023 paper, *The Inadequacy of Shapley Values for Explainability*, argues that Shapley values can provide misleading results.[8] Further analysis and cross-validation of these claims will be explored in future work.

5 Future Research

In future iterations of my research, I will aim to answer the following questions:

1. Can SHAP be applied to non-numeric input and output, such as prose?
2. How does SHAP compare with other feature-based XAI methods?
3. How do SHAP approximation methods, such as Monte Carlo sampling, KernelSHAP, TreeSHAP, and DeepSHAP, compare regarding functionality and performance?

[6] Aas, Kjersti, Martin Jullum, and Anders Løland. 2021. "Explaining Individual Predictions When Features Are Dependent: More Accurate Approximations to Shapley Values." *Artificial Intelligence*, March, 103502. https://doi.org/10.1016/j.artint.2021.103502.

[7] A Data Odyssey. 2023. "4 Significant Limitations of SHAP." YouTube. April 10, 2023. https://www.youtube.com/watch?v=zIbQgYxRBUc.

[8] Huang, Xuanxiang, and Joao Marques-Silva. 2023. "The Inadequacy of Shapley Values for Explainability." February 16, 2023. http://arxiv.org/pdf/2302.08160.

4. What data visualization tools exist in the SHAP library?
5. Given discrepancies in the adequacy of Shapley values for explainability, how does this impact future research? Is it possible to reconcile these inconsistencies?

6 Conclusion

As AI becomes more integrated into society, transparency in model decision-making becomes imperative. XAI tools provide insights into how AI models generate predictions. Among XAI techniques, SHAP is a mathematically grounded, model-agnostic approach for determining feature contribution and importance, offering both local and global interpretability.

This study explored SHAP’s roots in coalitional game theory, its desirable mathematical properties, and its local and global explanation abilities. However, the method is not without limitations. Computational complexity, challenges with high-dimensional data, sensitivity to correlated features, and the risk of misinterpretation should be examined when applying SHAP to real-world problems.

Future research will focus on exploring SHAP’s applicability to non-numeric data, comparing SHAP with other feature-based XAI techniques, exploring SHAP approximation methods, and examining visualization tools. Additionally, ongoing discussions about the adequacy of Shapley values warrant deeper exploration.

Bibliography

A Data Odyssey. 2023. “4 Significant Limitations of SHAP.” YouTube. April 10, 2023. https://www.youtube.com/watch?v=zIbQgYxRBUc.

Aas, Kjersti, Martin Jullum, and Anders Løland. 2021. “Explaining Individual Predictions When Features Are Dependent: More Accurate Approximations to Shapley Values.” *Artificial Intelligence*, March, 103502. https://doi.org/10.1016/j.artint.2021.103502.

Aboze, Brain John. 2023. “A Comprehensive Guide into SHAP Values.” Deepchecks. May 16, 2023. http://deepchecks.com/a-comprehensive-guide-into-shap-shapley-additive-explanations-values/.

Alammar, Jay. 2021. “Explainable AI Guide.” Pegg.io. 2021. https://ex.pegg.io/.

Fadel, Soufiane, and Statistics Canada. 2022. “Explainable Machine Learning, Game Theory, and Shapley Values: A Technical Review.” Www.statcan.gc.ca. February 28, 2022. https://www.statcan.gc.ca/en/data-science/network/explainable-learning.

Huang, Xuanxiang, and Joao Marques-Silva. 2023. “The Inadequacy of Shapley Values for Explainability.” February 16, 2023. http://arxiv.org/pdf/2302.08160.

Lundberg, Scott. 2018. “Shap.” GitHub. June 28, 2018. https://github.com/shap/shap?tab=readme-ov-file#citations.

Pawlicki, Marek, Aleksandra Pawlicka, Federica Uccello, Sebastian Szelest, Salvatore D’Antonio, Rafał Kozik, and Michał Choraś. 2024. “Evaluating the Necessity of the Multiple Metrics for Assessing Explainable AI: A Critical Examination.” *Neurocomputing* 602 (October): 128282–82. https://doi.org/10.1016/j.neucom.2024.128282.