Exploring the Effectiveness of SHAP over other Explainable AI Methods

Ms. Mayuri Manish Kedar, Ms. Gauri Narendra Mhatre

Guide: Ms. Needhumol Pillai

Keraleeya Samajam's Model College, Khambalpada Road, Thakurli, Dombivli (East), Kanchangaon, Maharashtra

ABSTRACT

Explainable Artificial Intelligence (XAI) has emerged as a critical domainto demystify the opaque decision-making processes of machine learning models, fostering trust and understanding among users. Among various XAI methods, SHAP (SHapley Additive exPlanations) has gained prominence for its theo- retically grounded approach and practical applicability. The paper presents a comprehensive exploration of SHAP's effectiveness compared to other promi- nent XAI methods.Methods such as LIME (Local Interpretable Model-agnostic Explanations), permutation importance, Anchors and partial dependence plots are examined for their respective strengths and limitations. Through a detailed analysis of their principles, strengths, and limitations through reviewing differ- ent research papers based on some important factors of XAI, the paper aims to provide insights into the effectiveness and suitability of these methods.The studyoffers valuable guidance for researchers and practitioners seeking to incorporateXAI into their AI systems.

Keywords: SHAP, XAI, LIME, permutation importance, Anchors and par- tial dependence plots

1 INTRODUCTION

In the world of artificial intelligence (AI), it's often hard to know why AI systems make the decisions they do. Trust issues and ethical concerns may arise due to lack of transparency. That's where Explainable AI (XAI) comes in. XAI helps us understand how AI models work and why they make certain decisions.XAI uses several methods to explain the decisions.One popular XAI method is called SHAP, short for Shapley Additive Explanations is the most popular and widelyused XAI method. It's based on a smart idea from game theory and aims to explain AI decisions by breaking down the importance of each factor. This can give a better understanding of how the AI model reaches its conclusions. But SHAP isn't the only game in town. There are other methods like LIME, which explains local predictions, and techniques that show how important each feature is. Each method has its pros and cons, the paper compares 4 widely used methods with SHAP based on the important factors of XAI and focuses on SHAP to see how it stacks up against other XAI methods.

2 LITERATURE REVIEW

This section presents the study of existing literature focusing on the comparison of SHAP's effectiveness with other methods for explaining AI.

1. Efficient Computation of Shap Explanation Scores for NeuralNetwork Classifiers via Knowledge Compilation

The paper presents an approach to quickly calculate the SHAP explanation scores for neural networks. SHAP is a widely used technique for interpreting machine learning model predictions. However, it can be computationally ex- pensive particularly for large datasets. The study takes on this

problem by effectively approximating SHAP values by utilizing the characteristics of neural networks. The approach greatly minimizes computational overhead by taking advantage of the local linearity of the SHAP explanation function and the lin- earity of neural network layers, while preserving accuracy. As demonstrated in the study, SHAP is highly effective in explainable AI due to efficient computa-tion for neural networks, enhancing accuracy, transparency, scalability, model agnosticism, and interpretability.[1]

2. Shapley Additive Explanations for Knowledge Discovery in Aerodynamic Shape Optimization

The study discusses the application of SHAP in Knowledge Discovery, based on how well it can explain intricate machine learning models. It talks about how SHAP solves the problem of comprehending model predictions by giving a numerical value to each characteristic in a forecast, which indicates how much of a prediction it makes. By breaking down intricate models into more basic elements, SHAP facilitates users understanding of the fundamental dynamics that underlie model predictions. The paper also gives insights of interpretability, flexibility, and computational efficiency of SHAP in various domains, which showcases its potential for knowledge discovery tasks and contributes towards its performance as well. SHAP's effectiveness in enabling knowledge extraction from black-box models suggests a positive performance[7]

3. Understanding Post-hoc Explainers: The Case of Anchors

The research explores the concept of post-hoc explainers, primarily focus- ing on inner workings of an explainable artificial intelligence (XAI) technique: "Anchors" which offers comprehensible justifications for model predictions. The paper highlights their unique features in comparison to other post-hoc explain- ers. In terms of interpretability, faithfulness, and efficiency, the study clarifies the advantages and disadvantages of Anchors through theoretical analysis and actual evaluations. The trade-offs of employing Anchors and other XAI tech- niques are also covered. The research advances knowledge about anchors and how they improve the interpretability of machine learning models.[5]

4. Detection of Risk Factors of Pcos Patients with Local Interpretable Model-Agnostic Explanations (Lime) Method That an Explainable Artificial Intelligence Model

In this study, it is aimed to extract patient-based clarifications of the contri-bution of important features in the decision-making process (estimation) of the Random Forest (RF) model, which is difficult to analysis for PCOS disease risk, with Local Interpretable Model-Agnostic Explanations (LIME). In this study, the Local Interpretable Model-Agnostic Annotations (LIME) strategy was ap- plied to the "Polycystic ovary syndrome" dataset to clarify the Random Forest(RF) model, which is hard to interpret for PCOS risk factors estimation. LIME (Local Interpretable Model-Agnostic Explanations) is a common methodology for making black box Machine Learning (ML) algorithms more interpretable and explainable. LIME often generates an explanation for a single prediction made by any ML model by learning a simpler interpretable model (e.g. linear classifier) around the prediction by randomly perturbing simulated data around the instance and obtaining feature importance through feature selection. LIME and comparative neighborhood algorithms have gained popularity due to their simplicity and straightforwardness. The LIME approach can be used to dis- cover which variables affect each estimation in the model to what extent and in which direction, as well as which variable has a greater impact on the model's outcomes than other factors. This gives a detailed explanation for each ob- servation, allowing any complex classifier to be explained in a straightforward manner. When the observations obtained from the results are interrogated, it can be said that the Follicle (No) L. and Follicle (No) R. variables are the most effective factors on the presence or absence of PCOS. For distinctive esteem ranges of these two factors, the result of PCOS or not varies. Based on this, it can be said that different values of Follicle (No) L. and Follicle (No) R. factors for PCOS status may be effective in deciding the disease.[2]

5. LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS FOR MUSIC CONTENT ANALYSIS

In this work, researcher proposed Sound LIME (LIME), an algorithm that extends the applicability of LIME to Music System Analysis (MCA) systems. They proposed three versions of SLIME and illustrate them with three types of singing voice detection systems to generate temporal and time-frequency ex- planations for the predictions of specific instances. The temporal explanations generated by SLIME are helpful for revealing how the BDT is making decisions based on content that does not contain singing voice despite possessing high classification accuracy for the selected instances. Such issues cast doubt on the generalizability of the model. They also demonstrated that the analysis of time-frequency explanations is helpful to gain trust in the CNN based SVD sys- tem. Researcher compared SLIME based explanations with saliency maps for the neural network model and the results suggest that model-agnostic SLIMEbased explanations agree in many cases with saliency maps.[6]

6. Peeking Inside the Black Box: Visualizing Statistical Learningwith Plots of Individual Conditional Expectation

In this paper, they have developed a suite of tools for visualizing the fitted values generated by an arbitrary supervised learning strategy. There work ex- tends the classical partial dependence plot (PDP), which has honestly become a very popular visualization tool for black-box machine learning result. The lim- ited functional relationship, however, often varies conditionally on the values of the other variables. The PDP offers the average of these relationships and thus individual conditional relationships are subsequently veiled, inconspicuous by the researcher. These individual conditional relationships can now be visu- alized, giving researchers additional insight into how a given black box learning algorithm makes use of covariance to generate predictions. Thus, a normal bend, such as the PDP, can complicate the complexity of the modeled relation- ship. Accordingly, ICE plots refine the partial dependence plot by graphing the useful relationship between the predicted response and the feature for indi- vidual observations. Specifically, ICE plots highlight the variation in the fitted values across the range of a covariance, suggesting where and to what extent heterogeneity might exist. In addition to giving a plotting suite for exploratory examination, researcher include a visual test for additive structure in the data generating model. Through assumed examples and real data sets, we demon- strate how ICE plots can shed light on evaluated models in ways PDPs cannot. By creating additive models from an invalid conveyance and presenting the real ICE plot into the lineup, interaction effects can be distinguished from noise, providing a test at a known level of significance. Future work will expand the testing methodology to other null hypotheses of interest.[3]

7. A Permutation Importance-Based Feature Selection Method forShort-Term Electricity Load Forecasting Using Random Forest

In this paper, a novel random forest (RF)-based feature selection method for STLF is proposed. Subsequently, the original feature set was utilized to train an RF as the original model. The optimal forecasting feature subset is selected only by the improved SBS method with simple principle and high efficiency. After the training procedure, the prediction error of the original model on the test set was recorded and the permutation importance (PI) value of each fea- ture was obtained. As it were two parameters of RF required to be adjusted, and the parameter selection method is clear. Considering this advantage, the proposed approach maintains a strategic distance from the influence of unrea- sonable model parameters on the feature selection outcomes. The conventional SBS method is optimized to diminish the number of iterations. Therefore, the efficiency of the search strategy is dramatically improved. The experimental re- sults based on real load data confirm the effectiveness of the proposed RF-based feature selection method for STLF. In addition, the optimized RF has superior generalization capability than SVR and ANN.[4]

8. Generative Local Interpretable Model-Agnostic Explanations

The linked paper introduces a method called Generative Local Interpretable Model-Agnostic Explanations (GenLIME), designed to provide local and inter- pretable explanations for machine learning models. GenLIME aims to address the limitations of existing explanation methods by generating synthetic samples around specific instances and training interpretable models on these samples. By leveraging generative models, GenLIME creates interpretable explanations that capture the local behavior of the underlying model. The paper discussesthe implementation and evaluation of GenLIME on various datasets, demon- strating its effectiveness in providing accurate and interpretable explanations for model predictions. Overall, GenLIME presents a promising approach for enhancing the interpretability of machine learning models, particularly in local contexts where precise explanations are needed.(plagarize)

3 ANALYSIS OF THE STUDY

1. **SHAP (SHapley Additive exPlanations):**
SHAP (SHapley Additive exPlanations) is a method which explains the outputof machine learning models. SHAP operates by analyzing a dataset containinginput features and corresponding predictions made by a machine learning model.
Principles:

1. **Local Accuracy:**

SHAP ensures that its explanations are locally accurate, meaning theyfaithfully represent how the model behaves in the vicinity of a specific instance. The explanations should reflect the model's decision-making process for that particular instance.

2. **Consistency:**

SHAP adheres to the principle of consistency, which means that similarinstances should receive similar explanations. Consistency ensures that the explanations provided by SHAP are stable and reliable across different runs of the model or variations in the dataset.

3. **Fairness:**

SHAP aims to provide fair explanations by ensuring that each feature'scontribution to the model's prediction is accurately represented, regardless of its nature or context. Fairness principles ensure that explanations do not exhibit biases based on sensitive attributes such asgender, race, or ethnicity.

4. **Robustness:**

SHAP seeks to provide robust explanations that are consistent acrossdifferent subsets of the data or variations in the model architecture. Robust explanations ensure stability and reliability in model interpretation, even in the presence of noisy or uncertain data.

5. **Interactivity:**

SHAP emphasizes interactivity by offering interactive visualization toolsthat enable users to explore and manipulate the explanations based on their specific needs and preferences. Interactive features enhance user engagement and understanding of the model's behavior.

2. **LIME (Local Interpretable Model-agnostic Explanations):**

LIME generates explanations for individual predictions of machine learning models through a systematic process. It begins by creating perturbed instances around the instance of interest, ensuring proximity to the original instance by random sampling from the feature space.
Principles:

1. **Local Interpretability:**

LIME focuses on explaining individual predictions of machine learning models, providing insights into the

model's decision-making process at a local level.

2. **Model Agnosticism:**

LIME is model-agnostic, meaning it can be applied to any machine learn- ing model regardless of its architecture or training algorithm.

3. **Human Interpretability:**

LIME presents feature importance rankings in an understandable and in- terpretable manner, allowing users to gain insights into the reasons behindspecific predictions.

3. **Permutation importance:**

Permutation importance is a method used in explainable AI (XAI) to under- stand the importance of input features in a predictive model. It is a method used for explaining black-box models. Permutation Importance is a good method which can be used for model interpretability. But completely depending on permutation importance can be risky, especially without the knowledge of the feature set.

Principles:

1. **Interpretability:**

Permutation importance provides a straightforward method for under- standing feature importance, allowing users to interpret which features have the most significant impact on model predictions.

2. **Model Agnosticism:**

Permutation importance is model-agnostic, meaning it can be applied to any machine learning model without needing to understand its internal workings, promoting flexibility and compatibility across different algo- rithms.

3. **Accuracy:**

By evaluating the impact of each feature on model performance, permu- tation importance helps prioritize features that contribute the most to accurate predictions, enhancing overall model performance.

4. **Robustness:**

Permutation importance is robust to noise and outliers in the data, as it evaluates feature importance based on changes in model performance rather than individual data points, enhancing the reliability and stabilityof the results.

5. **Transparency:** Permutation importance promotes transparency by re- vealing which features are most influential in driving model predictions, enabling stakeholders to understand the reasoning behind model decisionsand build trust in the model's behavior.

4. **Partial dependence plots:**

Partial dependence plots (PDPs) understands the relationship between one or more input features and the prediction outcome of a machine learning model. It understands how individual or combined features influence the predictions of a machine learning model.

Principles

1. **Visualization:**

They visually represent the relationship between input features and the model's predictions, with the y-axis depicting the predicted outcome and the x-axis showing the range of values for the selected features.

2. **Model Agnosticism:**

Partial dependence plots are agnostic to the specifics of the underlying model, allowing them to be applied universally across different types of machine learning algorithms.

3. **Interpretation:**

They provide insights into the impact of individual features on predictions, helping to understand whether relationships are linear, non-linear, or if there are any threshold effects.

5. **Anchors:**

Anchors are simplified rules or conditions that provide understandable ex-planations for individual predictions made by machine learning models.

Principles:

1. **Local Interpretability:**

They offer local interpretability by focusing on individual predictions rather than the entire model.

2. **High Certainty:**

Anchors aim to provide explanations with high certainty, meaning they identify conditions under which the model is highly confident about its prediction.

3. **User-friendly:**

These explanations are designed to be user-friendly and understandable by non-experts, aiding in model transparency and trustworthiness.

4. **Model-Agnostic:**

Anchors can be applied across different types of machine learning models, making them model-agnostic.

4 CONCLUSION

The exploration of SHAP's effectiveness over other Explainable AI methods presents a understanding of interpretability in machine learning models. SHAP stands out for its ability to provide highly interpretable explanations. Its model- agnostic nature ensures applicability across various machine learning models, enhancing its versatility. Delivering high-performance, accurate and transparent insights into model decisions, SHAP stands out as a robust choice. While LIME and Anchors may offer relative advantages in ease of implementation, SHAP's overall adaptability, scalability, and transparency surpass those of other methods evaluated. Consequently, researchers and practitioners can confidently rely on SHAP as a powerful tool for elucidating and comprehending the inner workings of complex machine learning models.

References

[1] Leopoldo Bertossi and Jorge E Leόn. Efficient computation of shap ex- planation scores for neural network classifiers via knowledge compilation. In *European Conference on Logics in Artificial Intelligence*, pages 49–64. Springer, 2023.

[2] İpek Balıkçı Çiçek, Zeynep Küçükakçalı, and Fatma Hilal Yağın. Detection of risk factors of pcos patients with local interpretable model-agnostic ex- planations (lime) method that an explainable artificial intelligence model. *The Journal of Cognitive Systems*, 6(2):59–63, 2021.

[3] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *journal of Computational and Graphical Statistics*, 24(1):44–65, 2015.

[4] Nantian Huang, Guobo Lu, and Dianguo Xu. A permutation importance-based feature selection method for short-term electricity load forecasting using random forest. *Energies*, 9(10):767, 2016.

[5] Gianluigi Lopardo, Frédéric Precioso, and Damien Garreau. Understanding post-hoc explainers: The case of anchors. *arXiv preprint arXiv:2303.08806*, 2023.

[6] Saumitra Mishra, Bob L Sturm, and Simon Dixon. Local interpretable model-agnostic explanations for

music content analysis. In *ISMIR*, vol- ume 53, pages 537–543, 2017.

[7] Satoshi Takanashi, Shinsuke Nishimura, Kaoruko Eto, and Keita Hatanaka. Shapley additive explanations for knowledge discovery in aerodynamic shape optimization. In *AIAA SCITECH 2023 Forum*, page 0904, 2023.