Decoding AI Complexity: SHAP Textual Explanations via LLM for Improved Model Transparency

Chung-Chian Hsu, I-Zhen Wu, Shih-Mao Liu
Department of Information Management
National Yunlin University of Science and Technology, Yunlin, Taiwan
{hsucc, m11123005, m11223012}@yuntech.edu.tw

***Abstract*—With the continuous advancement of artificial intelligence (AI), particularly in widespread domains such as healthcare and environmental applications, there is an increasing demand for model interpretability. Understanding the decision-making process of models contributes to building trust in them. Hence, the development of Explainable AI (XAI) has become crucial. This study proposes an approach to generate text via a large language model (LLM) for interpretation to enhance the interpretability of SHAP (Shapley Additive exPlanations) plots. The goal is to make the interpretability of model decisions accessible even to non-IT experts through textual explanations.**

Keywords—Explainable AI, Generative AI, SHAP, LLM

I. Introduction

As AI technology advances, Explainable AI (XAI) has become crucial to address confusion around complex machine learning models. Traditional black-box models are shifting towards more interpretable ones, enhancing reliability [1]. In sensitive domains like healthcare and environment, the need for explainability is growing, given the profound implications of model decisions. Understanding model decision-making is crucial to building trust and ensuring persuasive predictions. However, without specific explanations, visualization tools in XAI methods, such as SHAP plots as shown in Fig. 1 [2], may be challenging for individuals outside the relevant fields to understand.

To make model explanations accessible to a wider audience, this study aims to automatically transform these graphical representations into easily understandable textual explanations. This transformation not only contributes to improving the interpretability of models but also makes model applications more transparent and trustworthy, further meeting societal demands for model explainability.

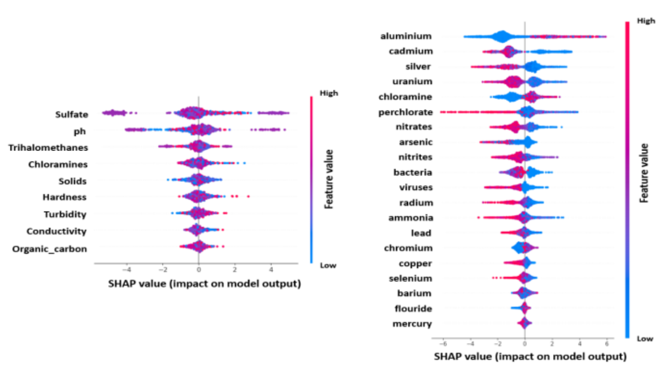

Fig. 1. Two SHAP Global Plots, which are not easily understandable.

By presenting the complex decision-making processes of models in clear and simple text, this study seeks to bridge the gap between AI technology and society. The goal is to enable a broader audience, including non-experts, to easily comprehend and trust these technologies, facilitating the widespread application of artificial intelligence across various domains.

II. Method

The proposed approach is shown in Fig. 2. First, numerical values are obtained from the SHAP global plot via designated formulas. Subsequently, the Pearson correlation coefficient between feature values and SHAP values is calculated. Next, all relevant information is transformed into table format. Finally, these tables are input into the LLM model to generate textual explanation automatically.

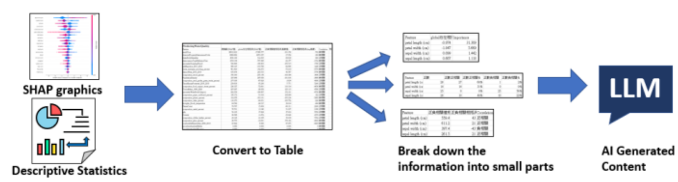

Fig. 2. Process of generating text to explain SHAP diagram for non-IT experts

A. Applying SHapley to the data

Explainable artificial intelligence enhances our understanding of model predictions. In this regard, SHAP demonstrates unique advantages. It not only provides explanations for the global model but also offers local explanations for individual predictions. The formula for local explanations is given by (1), where x denotes a data point, i represents a feature, and $\phi_i(x)$ signifies the feature contribution for a specific data point.

In global explanation, the focus lies on explaining the contribution of features in the model as well as their positive or negative correlations. The calculation formula for feature contribution is illustrated in ( ), where the overall contribution of a feature is determined by summing up the absolute values of its contributions across all data points [3]. A higher value of Φ_i indicates a higher level of contribution, i.e., more important.

$$\phi_i(x) = \sum_{S\subseteq N\setminus\{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!}\left[f(x_{S\cup\{i\}}) - f(x_S)\right] \quad (1)$$

$$\Phi_i = \sum_x^X |\phi_{i,x}| \quad (2)$$

B. Correlation coefficient between feature values and SHAP values

The other information which a SHAP diagram reveals is how a feature value impacts model output. Red/blue color indicates a high/low feature value. A far right/left position along the X axis indicates larger positive/negative impact on the model output. The Pearson correlation coefficient is widely trusted in statistics and machine learning, so we propose to use Pearson correlation to quantify the relationship between feature values and the impacts, i.e., the SHAP values. The Pearson formula is as shown in (3), where X_i, Y_i, $\hat{X}$ and $\hat{Y}$ represent the SHAP value, the feature value and the averages, respectively.

$$\gamma = \frac{\sum(X_i-\hat{X})(Y_i-\hat{Y})}{\sqrt{\sum(X_i-\hat{X})^2\sum(Y_i-\hat{Y})^2}} \tag{3}$$

C. Generating explanation via LLM Model with Partial Tables

Considering the utilization of the LLM model in this research, designed primarily for text processing, we transform SHAP diagrams and other pertinent information into tabular formats for presentation.

To aid in the understanding of information in each category, this study subdivides the data into partial tables based on specific topics and inputs them individually into the LLM model for interpretations. The proposed procedure enables the LLM model to accurately comprehend and respond to crucial features present in the SHAP diagram via extracted descriptive statistics.

III. Experimental Results

In the field of environmental applications, two water quality datasets[1,2] were used for experimental comparison. XGBoost was employed as the predictive model. In the two SHAP global plots in Fig. 1, the accuracy of the left plot was 0.765, while that of the right plot was 0.965. For this study, the ChatGPT API was utilized to build the prototype system for the LLM model.

In Table 2, the Response was automatically generated by ChatGPT as the explanation of the SHAP feature importance rankings regarding the right plot in Fig. 1. The experimental result demonstrates that ChatGPT can accurately respond to the importance rankings of this dataset. In the Table, "X" represents the extracted statistics in CSV file with columns indicating "Feature" and "Importance" from the right plot of Fig. 1.

TABLE 2. Explanation of feature importance w.r.t. the right plot of Fig. 1

Prompt
About 100 words of explanation and analysis: X
Response by ChatGPT
This set of data represents the importance or impact of different features in a certain context, with the highest importance assigned to "aluminum" at 4523.131 and the lowest to "mercury" at 265.849. The values indicate the relative significance of each feature in the given scenario. For instance, "aluminum" seems to be the most influential feature based on the provided data. On the other hand, "mercury" holds the lowest importance among all features listed. This analysis can be used to prioritize resources or interventions based on the significance of each feature in the dataset.

In Tables 3 and 4, ChatGPT elucidates positive/negative correlation of the features and their strengths of Fig. 1. "Y" represents calculated statistics with columns "Feature" and "Pearson correlation between feature value and SHAP value".

TABLE 3. Explanation of correlation of the left in Fig. 1 (Less Discernible)

Prompt
Please analyze and explain positive/negative correlations and correlation strengths within 100 words : Y
Response by ChatGPT
Positive correlations indicate that as one variable increases, the other variable also tends to increase. In this dataset, Turbidity (0.237) and Solids (0.196) have a moderate positive correlation with the feature value and Shap value. Hardness (0.056) and pH (0.054) have weak positive correlations. Negative correlations show that as one variable increases, the other tends to decrease. Chloramines (-0.11) and Conductivity (-0.09) exhibit moderate negative correlations, while Trihalomethanes (-0.081) and Organic_carbon (-0.083) have weak negative correlations with the feature value and Shap value.

TABLE 4. Explanation of correlation of the right in Fig. 1 (More Discernible)

Prompt
Please analyze and explain positive/negative correlations and correlation strengths within 100 words : Y
Response by ChatGPT
Positive correlations indicate that as the value of one feature increases, the value of another feature also tends to increase. In this dataset, "aluminum" has a strong positive correlation of 0.7561. Negative correlations mean that as one feature increases, the other tends to decrease. For example, "lead" has a negative correlation of -0.627. Strong negative correlations are seen in features like "radium" (-0.8751) and "nitrates" (-0.858). The strengths of correlations range from weak (e.g., "barium" at -0.2018) to strong (e.g., "radium" at -0.8751).

According to the experimental results, when observing the two SHAP global plots in Fig. 1, it's more challenging to discern positive/negative correlations in the left plot. The computed correlation coefficients tend towards 0, which is also reflected in the explanation of Table 3, where ChatGPT's responses indicate weaker relationships. In contrast, in the right plot, positive/negative correlations are clearly observable. Many feature correlation coefficients approach 1 or -1, indicating stronger correlations, consistent with the responses in Table 4.

IV. Conclusion

We proposed an approach to automatically generate textual explanations via LLM w.r.t. the SHAP diagram outputted by the XAI technique SHAP. Experimental results on two datasets demonstrated the proposed approach is feasible and promising. The textual explanations contributed to enhancing the overall interpretability of the AI models, especially for non-IT experts.

REFERENCES

[1] Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Ser, J. D., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., Jiang, R., Khosravi, H., Lecue, F., Malgieri, G., Páez, A., Samek, W., Schneider, J., Speith, T., & Stumpf, S. (2024). Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, *106*, 102301.

[2] Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. arXiv:1705.07874. Retrieved May 01, 2017.

[3] van Zyl, C., Ye, X., & Naidoo, R. (2024). Harnessing eXplainable artificial intelligence for feature selection in time series energy forecasting: A comparative analysis of Grad-CAM and SHAP. *Applied Energy*, *353*, 122079.

[1] https://www.kaggle.com/datasets/adityakadiwal/water-potability

[2] https://www.kaggle.com/datasets/mssmartypants/water-quality