Exploring the Relationship Between Feature Attribution Methods and Model Performance

Priscylla Silva [1,3] Claudio Silva [2] Luis Gustavo Nonato [1]

[1]Universidade de São Paulo [2]New York University [3]Instituto Federal de Alagas

Abstract

Machine learning and deep learning models are pivotal in educational contexts, particularly in predicting student success. Despite their widespread application, a significant gap persists in comprehending the factors influencing these models' predictions, especially in explainability within education. This work addresses this gap by employing nine distinct explanation methods and conducting a comprehensive analysis to explore the correlation between the agreement among these methods in generating explanations and the predictive model's performance. Applying Spearman's correlation, our findings reveal a very strong correlation between the model's performance and the agreement level observed among the explanation methods.

Introduction

- Predicting student success is challenging, and many models used for this purpose are difficult to interpret because of their black-box nature;
- It is insufficient to identify a potential student failure only; it is necessary to identify the factors analyzed by the model to generate the predictions;
- By employing feature attribution explanation methods, practitioners can gain insights into the importance of each input feature in influencing model predictions.

The Disagreement Problem

A local attribution method for the model f is a mapping $g : (f, X) \rightarrow E$ that, based on f, takes instances from X to the explanation space E, where $g(x) = (e_1, \ldots, e_K)$ is a point in E, K denotes the number of features, and e_i are the importance of each feature as to f. Consider two distinct local attribution methods, g_1 and g_2. For a given instance x, let $g_1(x)$ and $g_2(x)$ be the explanations generated by g_1 and g_2 in x, respectively. The disagreement problem occurs when $g_1(x) \neq g_2(x)$.

Research gaps

- Significant efforts have been made to use explanation methods to understand how a model predicts student success. However, there is a considerable gap in the literature when it comes to explaining the results in the field of education;
- The *disagreement problem* remains unresolved in the existing literature.

Research Goal

Is there a correlation between the model's performance and the disagreement level observed among explanation methods?

Study methodology

- In our experiment, we used two datasets to train binary classification models, predicting student success (success or failure) in one course;
- We trained Neural Network models for each dataset. Throughout the training, we systematically saved the models from intermediate epochs, creating a series of snapshots that captured the evolving state of the neural network;
- Using the test data, we computed the AUC metric for each model from the intermediate epochs;
- We employed nine state-of-the-art feature attribution techniques to explain the predictions made by the models: KernelShap, Guided Backprop, Input X Gradient, Occlusion, Smooth Gradient, Vanilla Gradients, LIME, Integrated Gradients, and DeepLift;
- We employed established (dis)agreement metrics (FA, SA, RA, and SRA) to quantify the (dis)agreement level between the explanation methods.

$$FA(g_1(x), g_2(x), k) = \frac{|\, top_k(g_1(x)) \cap top_k(g_2(x))\,|}{k} \tag{1}$$

$$SA(g_1(x), g_2(x), k) = \frac{|\bigcup_{s \in S}\{s \,|\, s \in top_k(g_1(x)) \wedge s \in top_k(g_2(x)) \wedge sign_k(g_1(x)) = sign_k(g_2(x))\}\,|}{k} \tag{2}$$

$$RA(g_1(x), g_2(x), k) = \frac{|\bigcup_{s \in S}\{s \,|\, s \in top_k(g_1(x)) \wedge s \in top_k(g_2(x)) \wedge rank_k(g_1(x)) = rank_k(g_2(x))\}\,|}{k} \tag{3}$$

$$SRA(g_1(x), g_2(x), k) = \frac{|\bigcup_{s \in S}\{s \,|\, s \in top_k(g_1(x)) \wedge s \in top_k(g_2(x)) \wedge sign_k(g_1(x) = sing_k(g_2(x) \wedge rank_k(g_1(x) = rank_k(g_2(x)\}\,|}{k} \tag{4}$$

Figure 1. Disagreement Metrics by Krishna et al. (2022).

- We compute Spearman's rank correlation to explore the relationship between model performance, as measured by the AUC metric, and the (dis)agreement level.

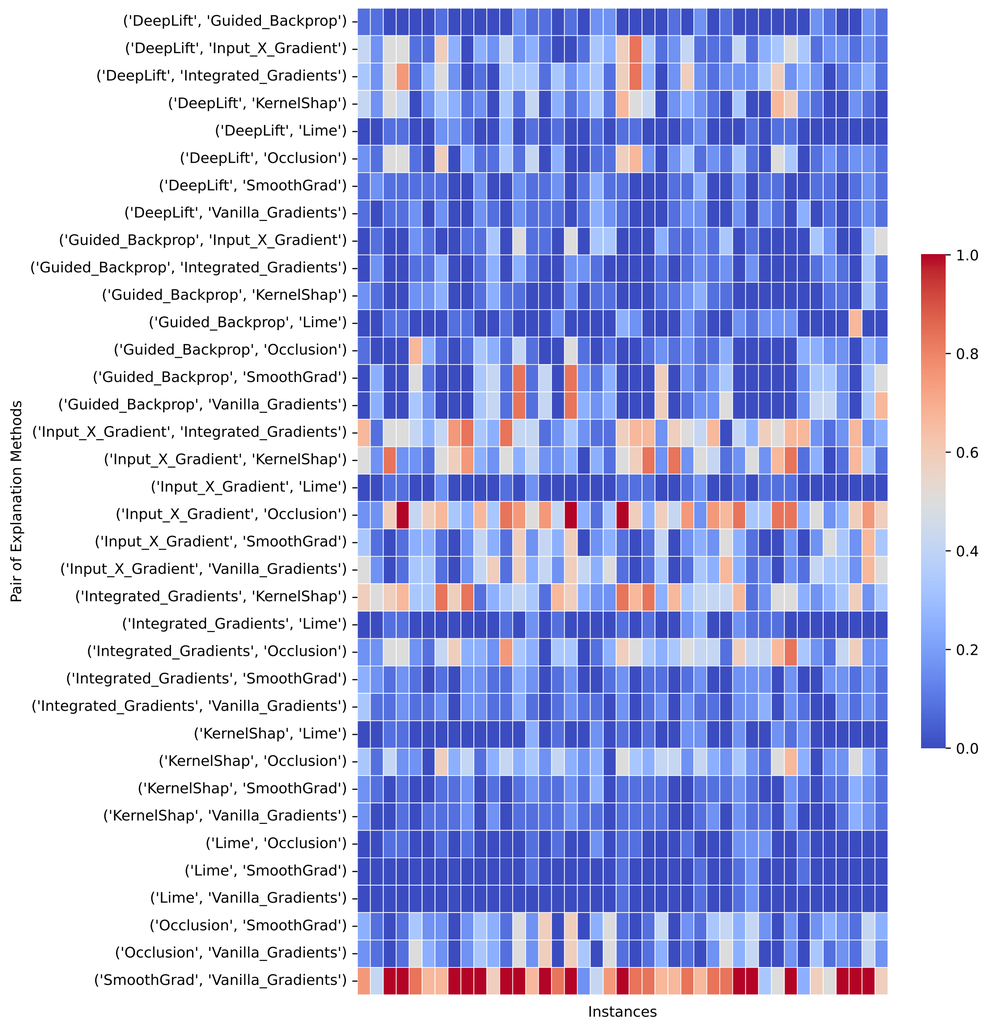

Figure 2. Heatmap illustrating the (dis)agreement levels between explanation methods.

Results and discussion

In both datasets analyzed in the student success prediction task, we were able to observe that there is a strong correlation between the model's performance, measured using AUC, and the (dis)agreement level between the methods, measured using the FA, SA, RA, and SRA metrics. The strong correlation we identified implies that the agreement among explanation methods becomes more evident as the model's performance improves. A higher-performing model tends to yield explanations that exhibit more substantial consensus across various explanation techniques. This finding underscores the intrinsic connection between model quality and the interpretability of its predictions.

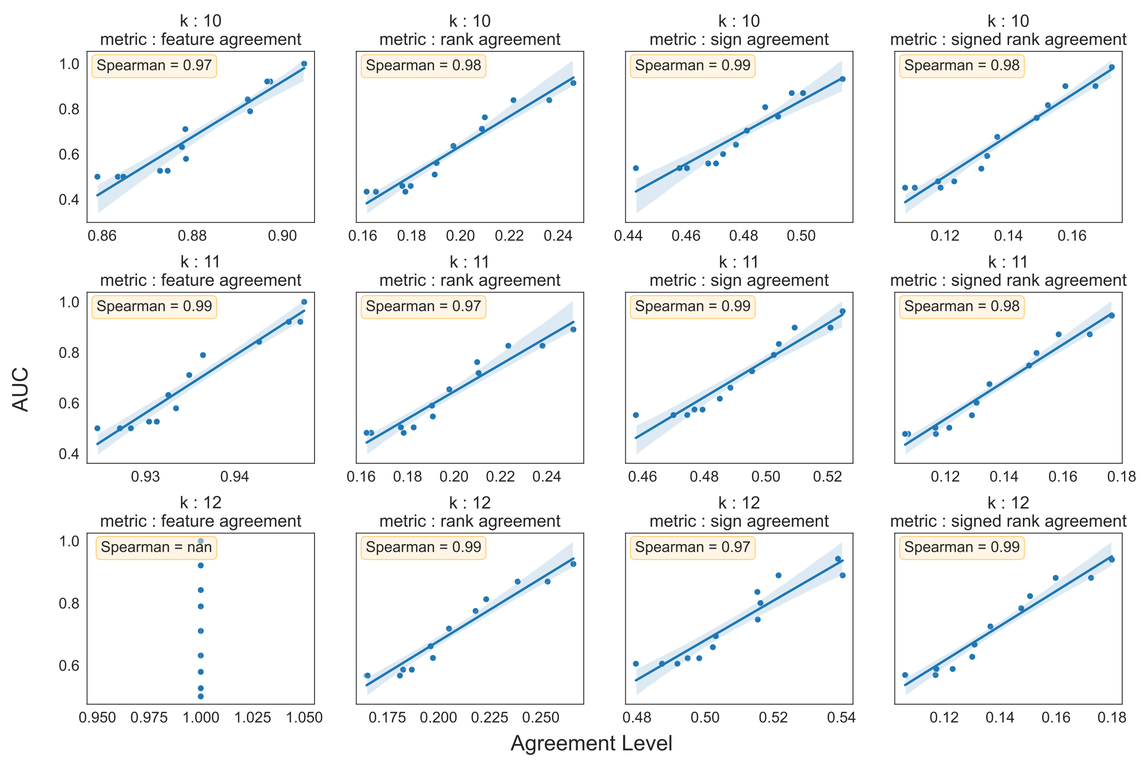

Figure 3. Correlation between Model Performance (AUC) and (Dis)agreement Metrics.

Practical implications

Practitioners should thoughtfully consider the model's performance before employing any explanation method. Our results show that models with an **AUC greater than or equal to 0.8** consistently exhibit the highest levels of agreement among explanation methods.

Author[1] Website