Exploring the Relationship Between Feature Attribution Methods and Model Performance

Paper submitted to AI4ED-AAAI-2024 — Track: Responsible AI for Education (Day 2)

Dear Editor,

We thank all the AI4ED reviewers for their detailed feedback on the version of the paper submitted to the workshop. We have revised the paper based on their comments and suggestions. Below, we detail how the reviewers' comments have been addressed.

Sincerely,

The Authors.

Response to Reviewer Comments

Reviewer 1

This paper looks at the agreement between different explanation or feature attribution methods, in the context of models (simple multilayer perceptrons) for binary prediction of student performance. Using on the agreement methods in Krishna et al. (2022)—sign agreement, rank agreement, feature agreement (i.e. top-k membership overlap) and signed rank agreement—explanation methods are compared and evaluated for consistency. This paper finds that increased agreement correlates with increased model performance: more accurate models have more consistent explanations across explanation methods. The datasets and models here are pretty small and I'm not sure how relevant these results would be in other settings, but it's an interesting result and this seems like a worthwhile paper.

We thank the reviewer for his/her feedback. We have considered and addressed all of his/her issues and comments.

R1.1. rework the figures to be a little larger (trim whitespace, reduce padding, etc) as they are hard to read at this scale, especially Figures 1 and 5. Even zooming in they're blurry. You can export the image as a PDF at high res and embed with LaTeX to avoid loss of resolution.

We have reworked all the figures to ensure they are more readable.

R1.2. add the datasets to the captions of Figures 3 and 4.

The datasets have been added to the captions of the figures.

R1.3. adding more information about the source of the Intro to Programming dataset (understandable if it was excluded for anonymization).

More details about the Intro to Programming dataset were added in section 2.2 and in Appendix C.

Reviewer 2

Quality: The paper is of great quality, well written and the narrative follows a good flow of ideas. It offers a robust background about the problem to address, limitations of ongoing approaches, and the need for a new approach; which is at the core of the paper. Specifically, the problem of disparities in determining the importance of features/factors' in a predictive model, known as a disagreement problem. The research question is well articulated and properly introduced. Good references to support the main arguments of the paper.

Clarity: The introduction clearly situates the problem the authors are addressing; namely, the lack of explainability in deep learning models used in education, because they can impact decisions made about students's success or devising sound intervention strategies. Initially, the authors indicate using feature importance to elucidate what factors influenced the decision of a model.

Good use of the Appendix section to elaborate on different explanation methods.

Originality and significance of this work: The authors address the need to better understand what influences model's predictions in the context of explainability in AI-driven systems in education. They accomplish this by comparing nine methods in two aspects: 1) generating explanations, and 2) prediction performance.

Pros: Overall a great methodology for comparing different methods used to explain a neural network-based classifier's predictions, which has high relevance in this field. I look forward to seeing future work of this methodology applied to more datasets.

We thank the reviewer for recognizing the importance of our study and for valuable suggestions. We have considered the reviewer's criticisms and we believe we have addressed them properly.

R2.1. On the problem formulation, it seems that the "n" is the index that denotes the number of features in a model, if so please state that to avoid confusing with n instances or observations in the dataset.

The index was modified and explanatory text was added to avoid confusion between the number of features in the model and the number of instances in the dataset.

R2.2. Perhaps explain what is the meaning of positive and negative sign in the feature importance (my sense is the two classes to predict).

The positive and negative signs indicate the effect of a feature on the model's prediction. An explanation of this was included in Section 2.1.

R2.3. In Appendix A, it is stated that two metrics that rely on domain knowledge were left out. Were those methods included, what would be the impact on the study?

These metrics are useful when an expert already knows the most important features and their order of importance for prediction. However, using these metrics can add subjectivity to the study as it is

difficult for an expert to determine with certainty which features in the dataset and their ranking should be considered most important.

R2.4. Minor suggestion: Move the reference to Appendix B to the sentence: "After the data preprocessing step.." Something like: "After the data preprocessing step (see Appendix B for more details on preprocessing)..."

We have updated the text.

R2.5. In section 2.1 add an Appendix to briefly explain the nine feature attribution techniques.

Added (see Appendix B).

R2.6. Cons: Not necessarily a major con, but in section 2.2.2, the authors use the average of the disagreement methods, how about including also the standard deviation, which will also offer some insights about the spread of the scores.

We thank the reviewer for the suggestion. We have addressed the issue by providing information about the spread of scores for agreement metrics in Appendix D.