REI603M - Assignment 7

Please return this assignment online on Canvas (code and slides). Please submit the files together as a `.zip` file.

Explainability

In this week's assignment, our objective is to explore ways to make models more "explainable." We often rely on models to make predictions and inform our decision-making, but their inner workings can sometimes be a mystery. By making models more explainable, we can gain a better understanding of how they arrive at their predictions and build trust in their results.

It's important to note that explanations might involve understanding the causal relationships within the data. However, we should keep in mind that simply fitting a model to data does not guarantee causality. While models can help us understand what they're doing, they might not provide insight into how the world actually works.

Through this assignment, we'll be investigating various techniques for making models more explainable. By doing so, we hope to gain a better understanding of the models we use and the insights they can provide.

There are several libraries available for this assignment to help make models more explainable. Here are two options that you can use:

- SHAP is a game-theoretic approach that explains the output of any machine learning model. This method helps identify which features are most important for a particular prediction and how each feature contributed to the outcome. If you're interested in studying this approach, you can find the original paper here.
- LIME is another option that provides an interpretable explanation of a model's prediction by learning a simple, interpretable model around it. This method can help you understand how different features influence the final outcome of the model. If you'd like to dive deeper into LIME, you can find the original paper here.
- For image data, you might want to consider saliency maps, grad-cam, LIME and SHAP.
- If you are working with LLMs, you can use this opportunity to incorporate explanations into your project. There are still methods that work for LLMs, but they assume access to the model. If you want, you can also apply the methods above to a new dataset you find online.

It's important to note that other libraries and methods also exist for making models explainable, and you're free to explore those options as well. Ultimately, the goal is to develop a better understanding of how our models make predictions and to build trust in their results.

The Task

Choosing a Dataset and Explainability Method

For this task, it's important to select a dataset that has labeled data. If possible, I encourage you to use the dataset for your final project. Once you've selected your dataset, use a supervised learning approach to train a model on it.

Next, choose an explainability method to study how input features impact your model's performance. For this assignment, you can select from the two options we provided (SHAP and LIME), or explore other explainability methods. Use the method you choose to identify which features are most important for your model's predictions and how each feature contributes to the outcome.

If possible, try to group your input data to gain a more granular understanding of how different features impact model performance. If segmentation is not possible, consider using clustering or UMAP to select representative examples to study. As you analyze your data, take note of any unexpected or counterintuitive findings.

Submission Requirements

For this assignment, you are required to submit the script used for your analysis that includes your original work, as well as slides to present your results. The script can be in the form of a notebook or a collection of `.py` files. Keep in mind that the presentation should be concise and easily understandable within the given time frame. **Your presentation should be 5-7 minutes.**

Peer Evaluation

In this assignment, you are required to make models more "explainable." This rubric is designed to guide the peer evaluation process. A total of 25 points are available, and 20 points are sufficient for a full grade. Exceptional work that goes above and beyond the requirements can earn up to 5 additional points as a bonus.

1. Selection and Preparation of Dataset

- **0 Points:** No dataset selected or provided dataset is inappropriate for the task; lacks labeled data or relevance to explainability methods.
- **1-2 Points:** Dataset is selected but poorly suited for the task; minimal effort in preparation or explanation of its relevance.
- **3-4 Points:** Appropriate dataset selected; adequate preparation and justification for its use in the project.
- **5 Points (Exceptional):** Excellent choice of dataset; thorough preparation and insightful explanation of its importance and relevance to explainability.

2. Application of Explainability Method

- **0 Points:** No application of explainability methods or significant misunderstanding of their purpose.
- **1-2 Points:** Attempted to apply an explainability method but with limited success; lacks depth or proper implementation.

- **3-4 Points:** Successful application of an explainability method; shows understanding of how it works and its relevance to the model.
- **5 Points (Exceptional):** Exceptional application of explainability methods; demonstrates deep understanding and innovative use in the context of the project.

3. Analysis of Model Predictions and Feature Impact

- **0 Points:** No analysis of model predictions or feature impact.
- **1-2 Points:** Basic analysis provided but lacks depth or insight; minimal effort in interpreting the results.
- **3-4 Points:** Solid analysis of model predictions and feature impact; demonstrates good understanding and ability to derive meaningful insights.
- **5 Points (Exceptional):** Comprehensive and insightful analysis; shows exceptional ability to interpret and explain the model's predictions and the impact of features.

4. Presentation and Communication of Findings

- **0 Points:** The presentation is poorly organized or unclear, significantly hindering understanding of the findings.
- **1-2 Points:** The presentation is somewhat clear but lacks polish or coherence; difficult to follow or understand at times.
- **3-4 Points:** Good presentation quality; findings are communicated clearly and effectively with minor issues in organization or clarity.
- **5 Points (Exceptional):** Exceptional presentation; highly effective communication of findings, well-organized, engaging, and clear.

5. Innovation and Depth of Insight

- **0 Points:** No evidence of innovation or depth of insight; fails to move beyond basic analysis or application of methods.
- **1-2 Points:** Limited innovation or insight; shows some attempt at deeper analysis but remains surface level.
- **3-4 Points:** Good level of innovation and insight; demonstrates thoughtful analysis and application of methods to yield new understandings.
- **5 Points (Exceptional):** Exceptional innovation and depth of insight; demonstrates novel approaches and deep, meaningful analysis leading to significant findings.

Please note that the deadline for submitting your assignments is at 23:00 on the 6th of March.

Good luck with the assignment. If you have any questions or concerns, please do not hesitate to reach out through Ed or by sending an email to `hafsteinne@hi.is`.