Explainable AI: Breaking Down the Black Box

Adrian Boskovic, Sam Johnson-Lacoss, Chris Melville, Josh Moore, Thomas Pree, Lev Shuster

Carleton College Computer Science | Winter 2024 | Advised by Anna Rafferty

Introduction: What is Explainable AI?

Machine learning models are complex, and their predictions are not always easy to understand. How can we, as users, determine a model's "thought process"?

Explainable artificial intelligence (or XAI) uses post-hoc model-agnostic techniques to improve a user's trust in the model's prediction.

- Visualizes an explanation for the model's prediction after it is made.
- Treats the model as a **"black box"** (i.e. we only know its output).

A famous example of XAI in action:

An XAI technique can help uncover issues with a model by showing *why* a mistake was made (like a model learning to recognize snow in the background).

Our Explainable AI Techniques

Our goal was to test three XAI techniques on two ML models: ResNet (images) and MOOC (tabular data). Here are the techniques.

LIME (Local Interpretive Model-Agnostic Explanations):

- Written by Ribeiro et al, this method trains a local surrogate model using linear regression between the input and black box prediction.

Anchors:

- By the same authors as LIME, this method "anchors" a precise data point of interest (by perturbing it) to see how black box results change.

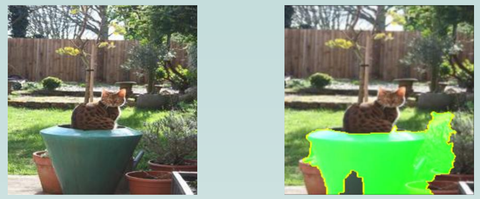
Shapley values:

- Mathematical approach written by Lloyd Shapley. It uses cooperative game theory, treating features as "players" that each contribute a "Shapley value" to the avg. prediction.

LIME: Explaining a Residual Neural Network

ResNet is a neural network trained to predict 1000 possible classes from the ImageNet database, including images of cats and dogs. Giving LIME the predict function and the image (in NumPy array form) produces an explanation.

A cat predicted to be a pot (0.98) by ResNet:

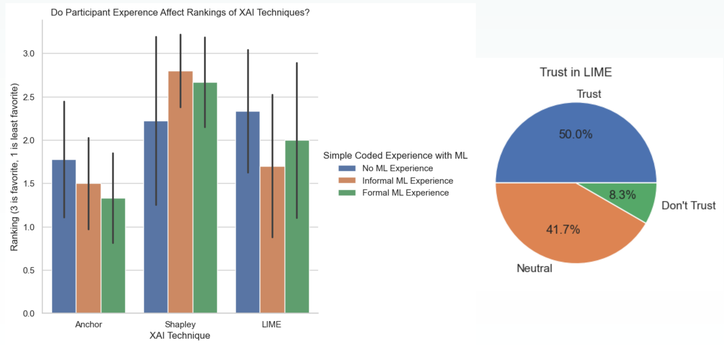


According to LIME, ResNet predicted "pot" with high certainty because there was a pot in the image (highlighted in green). A clear explanation like this could improve trust in ResNet, assuming the user trusts LIME.

User Study

A dog predicted to be a polar bear (0.56), explained by each technique:

For our survey, we collected data by showing users like the one you see above, with an input, the predicted class, and explanations from each technique. For each image we asked if users trusted the model's prediction, and how each explanation improved their understanding in the model's thought process.



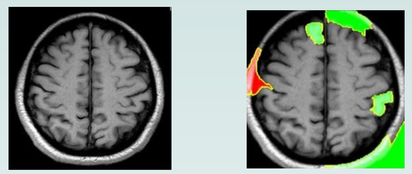
Results

Some key takeaways:

- Shapley was the favorite technique among participants, especially those with ML background. Those without ML background preferred LIME.
- Lack of consistency between XAI techniques undermined user understanding of the underlying model (there was no way to know which was correct)
- Trust in the underlying model decreased if incorrect predictions weren't given a clear explanation.

Tumors Example

One medical application of ML is training models to recognize brain tumors in MRIs. We trained a similar model (based on ResNet) and used XAI to see how it worked.



Our model classified this image of a healthy brain as one with a tumor. LIME helped us visualize our model's error: it wasn't looking for a tumor or even at the brain. Thus, our model would not be useful in a medical setting.

Conclusions

- Every technique works quickly and is easy to understand, even for those without an ML background. The insight can be extremely helpful!
- XAI has limitations in what it can explain (doesn't say *why* model is focusing on a certain area), and it is easy to misinterpret results.

Shoutout: *A special thanks to our advisor Prof. Anna Rafferty, as well as all our survey participants for making this project possible!*

Citations

[1] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier* (arXiv:1602.04938). arXiv. doi.org/10.48550/arXiv.1602.04938

[2] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition (Version 1). arXiv. doi.org/10.48550/ARXIV.1512.03385

[3] Sarta, J. (2021). Brain Tumor Classification (MRI). www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri/data?select=Testing