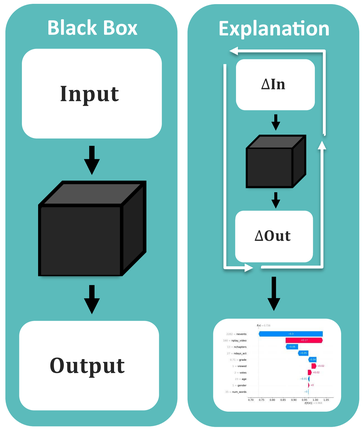
Explainable AI: Breaking Down the Black Box

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

What is XAI?



The field of ML has long been marked by a need for explainable artificial intelligence. This act of explanation intends to influence trust and build understanding in a model and illuminate potential biases in its data or training [1].

Why do we care?

Machine learning is often applied in high-risk areas such as forensics and increasingly in medicine [2]. Moreover, the EU and US have both drafted bills citing a *Right to Explanation* for every model. As such, methods for XAI techniques are in very high demand.

Our XAI Techniques

Shapley values

Shapley values leverage **coalitional game theory** to find how each feature factors into a prediction. In his original work, Shapley shows the calculation's robustness through the following four axioms [4]:

- **Efficiency:** All Shapley values must sum to the difference between prediction on the input and the average prediction.
- **Symmetry:** Features i and j have the same contribution to the prediction if they contribute identically to all coalitions.
- **Nullity:** If feature i changes nothing in the prediction, its contribution is 0.
- **Additivity:** For a prediction with multiple components $p + p'$, the Shapley values are calculated as ϕ + ϕ'.

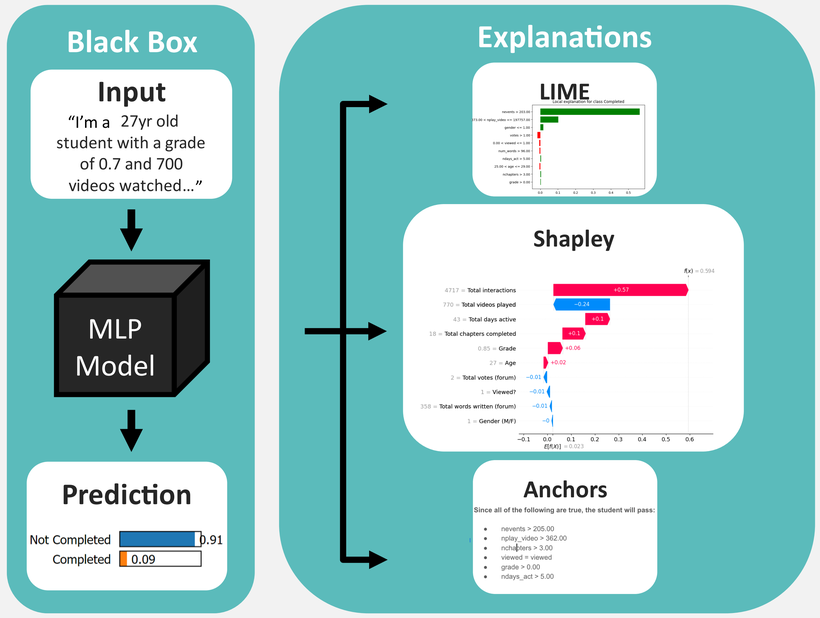
LIME

LIME uses a **local surrogate model** to explain how a model came to its decision. The method trains a simpler model on the outputs of the black box, which then allows the user to analyze the surrogate.

Anchors

This technique perturbs the data in a similar way to LIME, searching for the part of the input that **anchors the prediction,** where the model's output is the same so long as these anchors are all present.

Methodology



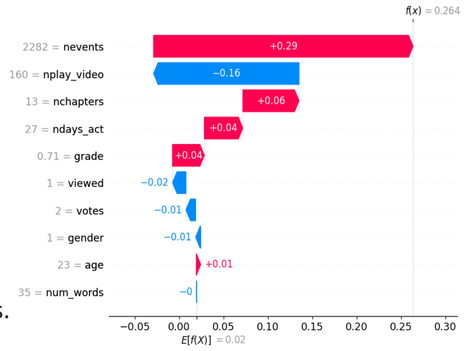
Shapley values — Continued

Mathematical Basis

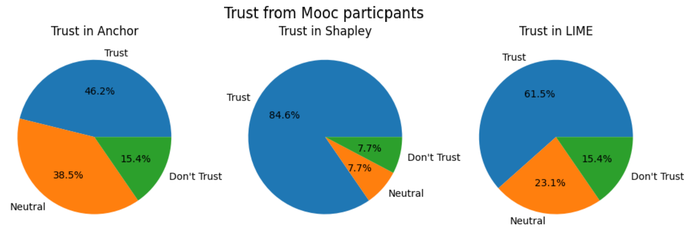
Let the act of prediction be a game, where the feature values are the players, and let the model output be the game's payout. To calculate each player's contribution to the result, have each feature begin to participate in the game in a random order [3]. **A feature's Shapley value is the difference between payouts when a feature either *is* and *isn't* playing.**

Interpreting Shapley values

Each Shapley value tells us how the input caused this prediction to deviate from the norm. Although this gives us some good insights into the model's processes, we need to keep in mind this method's specific definition of fairness.



User Study



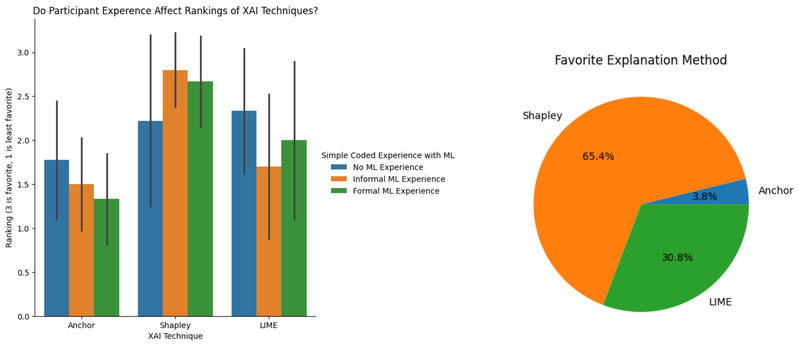
"It's doing a good job of explaining what happened, but not how it comes to conclusions."

"We don't know how much better it gets when they're over the threshold."

"Big arrow go right. I like. Sum of big arrows is prediction. I like.

"Big fan of the Shapples, as they say. Makes me feel warm and fuzzy."

"I don't get the math, but I intuitively get it."



Key Takeaways

- Each of these techniques are **extremely quick and easy to implement**, and they provide **valuable insights** to those of a wide range of technical backgrounds.
- Even with such different approaches, all of these techniques can identify biases in the model in an accessible manner.
- **Misinterpretation is common** without a deeper understanding of the applied technique, and **not all of these work globally**.

[1] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why should I trust you?"* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/2939672.2939778
[2] Sim, J., Fong, Q., Huang, W., & Tan, C. (2021). *Machine learning in medicine: What clinicians should know.* Singapore Medical Journal. doi.org/10.11622/smedj.2021054
[3] Shapley, L. S. (2023). *A Value for n-person Games*. The RAND Corporation.
[4] Molnar, C. (2023). Shapley Values. In Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Independently Published.

I'd like to thank Professor Anna Rafferty, the group, and our survey's participants for their incredible support.