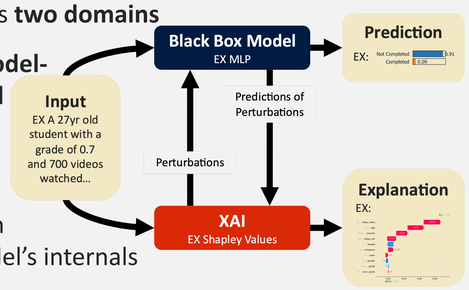
Explainable AI: Breaking Down the Black Box

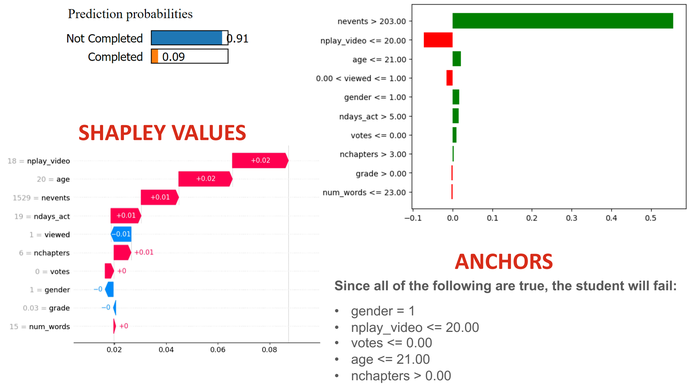
Adrian Boskovic, Sam Johnson-Lacoss, Chris Melville, Josh Moore, Thomas Pree, **Lev Shuster**

What is XAI?

Explainable Artificial Intelligence (XAI) techniques provide insight into ML models' **assumptions** and **priorities,** identifying blind spots, biases, and possible improvements, influencing model trust among users. To investigate, we...

- Compared three XAI techniques' **theoretical bases** and **usability** across **two domains**
- Focused on **model-agnostic and local** XAI: techniques which explain a single prediction, without relying on access to the model's internals





XAI Techniques

LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS (LIME)

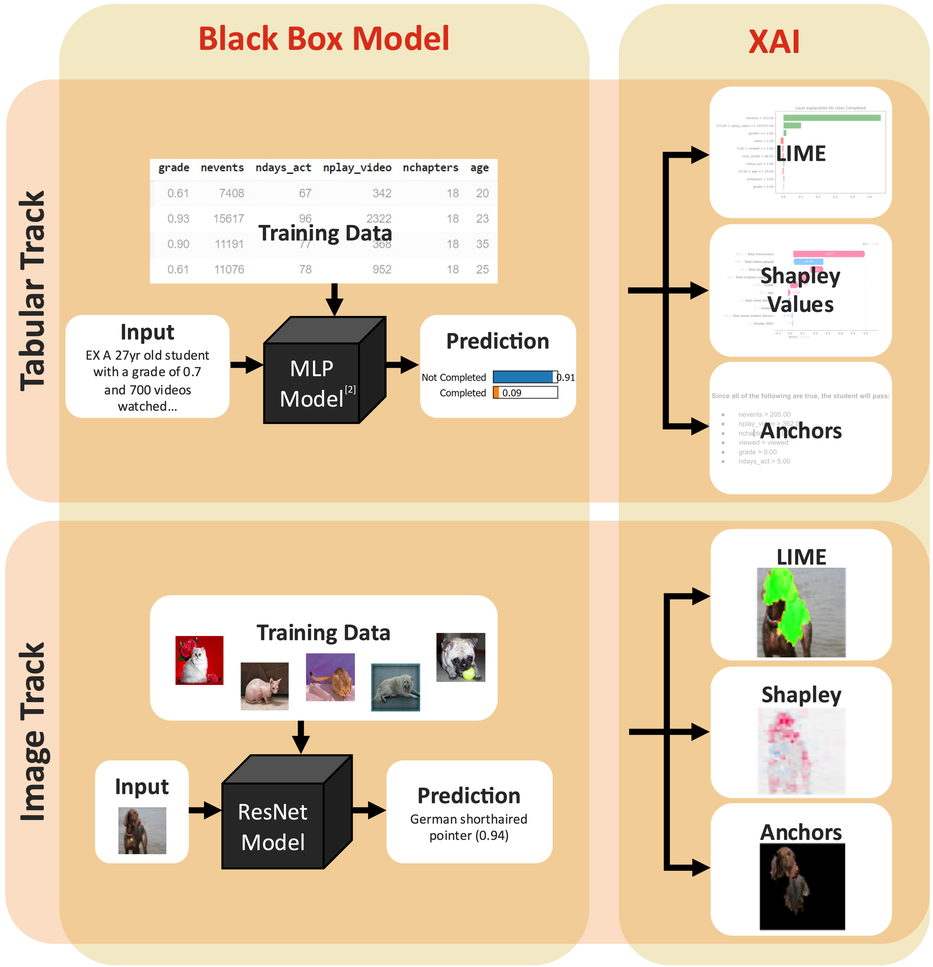
LIME generates a simple (therefore explainable) **surrogate model** to **approximate** the Black Box model across a **small vector-space**.

SHAPLEY VALUES

With a strong mathematical basis in **cooperative game theory**, Shapley calculates each feature/player's **relative contribution** (positive or negative) to the prediction.

ANCHORS

Using a similar technique to LIME, Anchors outputs a set of **rules** which anchors the input of interest. Changing the input's features **without modifying the anchors** would not effect the black box model's prediction.



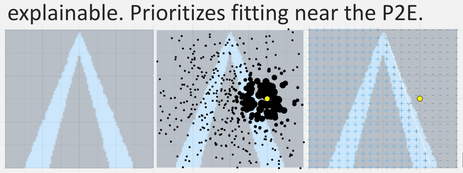
LIME Continued

HOW LIME WORKS

1. Given a black box model and a prediction to explain (P2E), LIME **finds** a simple model that mirrors the black box near the P2E.

a. Generates Training Dataset by perturbing the P2E's features along a gaussian distribution [1].

b. Builds Local Surrogate Model using the training dataset. The model is most commonly linear regression, but can be any model simple enough to be inherently explainable. Prioritizes fitting near the P2E.



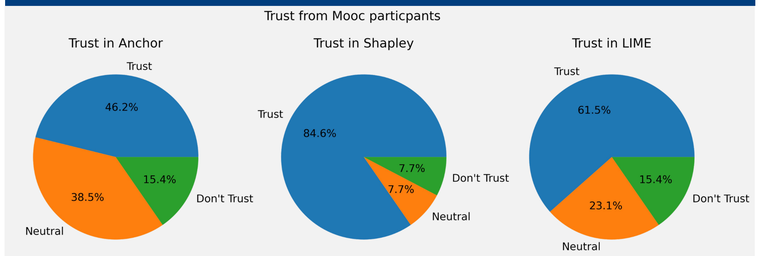
[1]

2. Given a linear regression model and a P2E, LIME **finds** the relative importance of the P2E's feature values, and where they fall relative to the decision boundary

TABULAR USER STUDY CONCLUSIONS

- Consistently ranked behind Shapley but ahead of Anchors, LIME was the **preferred choice by non-technical** participants.
- While **visually busy**, LIME was the only technique that provided both **rules** and **weights**, giving a solid starting point no-matter what participants where looking for.

User Studies: Comparing Techniques



*"I think these will overall **increase trust** in ML."*

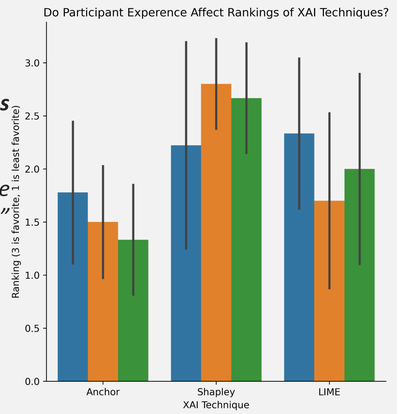
*"I appreciate seeing [the] amount of mattering **[weighting] of Shapley/LIME**, as compared to **binaries in anchoring**."*

*"I **trust the explanations** to do their thing, but they make me really **distrust the model's** processes."*

*"This makes me **trust the model less** because of the sexism."*

*"I **trust none** of these because it [the model] needs to weigh grade more."*

*"I don't think you can be fully accurate then, since they're not **consistent** [across explanations]."*



Key Takeaways

- XAI techniques can be applied in **under an hour**, using off the shelf packages, to deliver valuable insight.
- LIME, Shapley, and Anchor's conclusions are **limited** by their focus on **local explanations**, as well as **inconsistency** between techniques.
- XAI makes ML models **more understandable**. However, without an understand of how the technique works, overgeneralizations are the norm.
- After ranking highest in nearly every metric, **Shapley is an easy decision** when picking a starting XAI technique.

[1] Molnar, C. (2023, August 21). Interpretable Machine Learning. Retrieved January 15, 2024, from christophm.github.io/interpretable-ml-book/ [2] Muthukumar, V. (2019). MOOCs-Dropout-Prediction. Retrieved January 19, 2024, from github.com/vickymhs/MOOCs-Dropout-Prediction. Scan the QR-code for additional references.

Thanks to professor Rafferty, user study participants, and the members of our comps group.