Explainable AI: Breaking Down the Black Box

An exploration of three model-agnostic techniques to "explain" classifications given by machine learning models

Carleton College Computer Science Comps

29 February 2024

Thomas Pree,

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

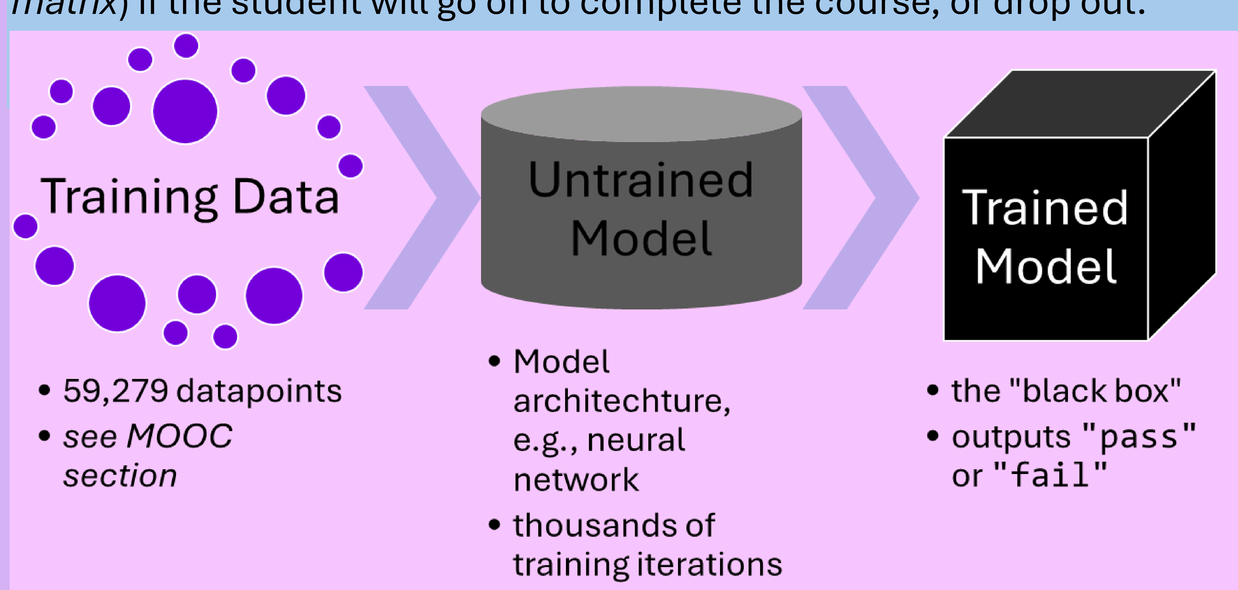
Advised by Professor Anna Rafferty

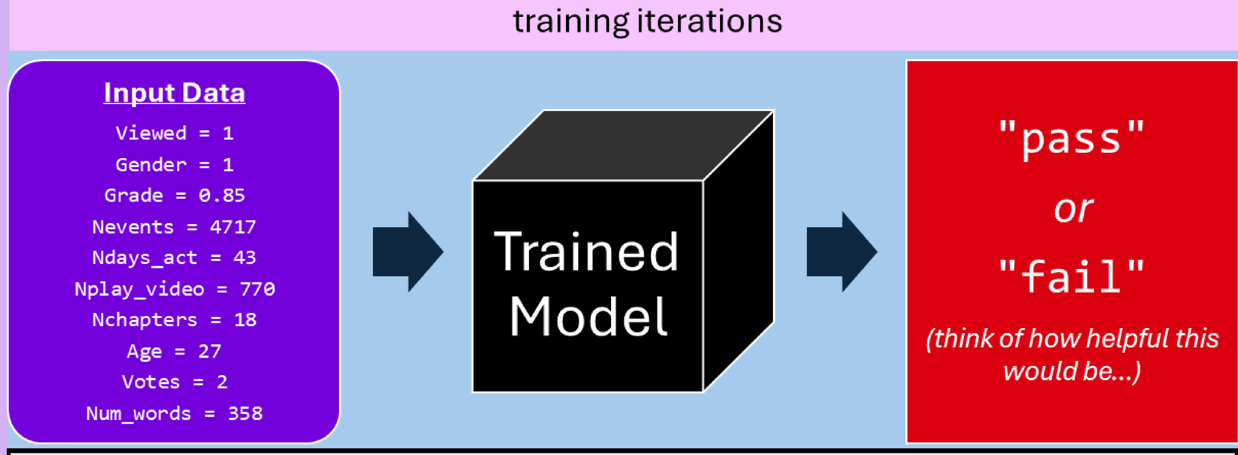
Introduction

Artificial intelligence is known for making accurate predictions, but providing little in the way of reasoning for its "thinking". *XAI: Breaking Down the Black Box* is an exploration of three common techniques for model-agnostic prediction explanation: LIME, Shapley, and Anchors. These techniques generally work by testing perturbations of the sample input through the classifier, and assessing which features have the most significance to the model's output. Models were trained in two domains: images (cats vs. dogs), and tabular data (MOOC dropout prediction), and each technique was implemented on each domain. Students and recent graduates were surveyed to determine a general sense of understandability and trust the explanations provided with respect to the underlying models. A website to interact with the models and host the full writeup will also be available.

MOOC

MOOCs (Massively Open Online Courses) are typically university courses easily available in an online, asynchronous format. This data comes from a MOOC course, introductory programming, from several years ago. About 5 weeks into the course, the training data was collected, which contains information like how many chapters of the textbook the student opened, their current grade, etc. This data was used to train a neural network that predicts (with very high accuracy — see *confusion matrix*) if the student will go on to complete the course, or drop out.



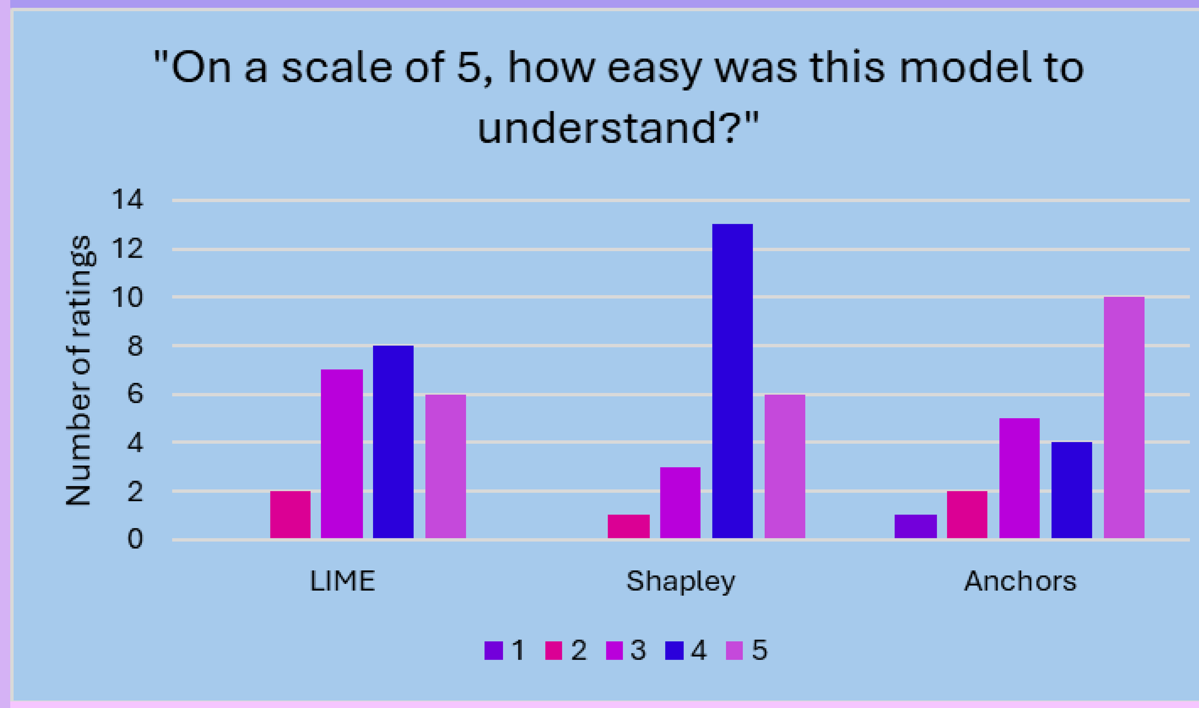


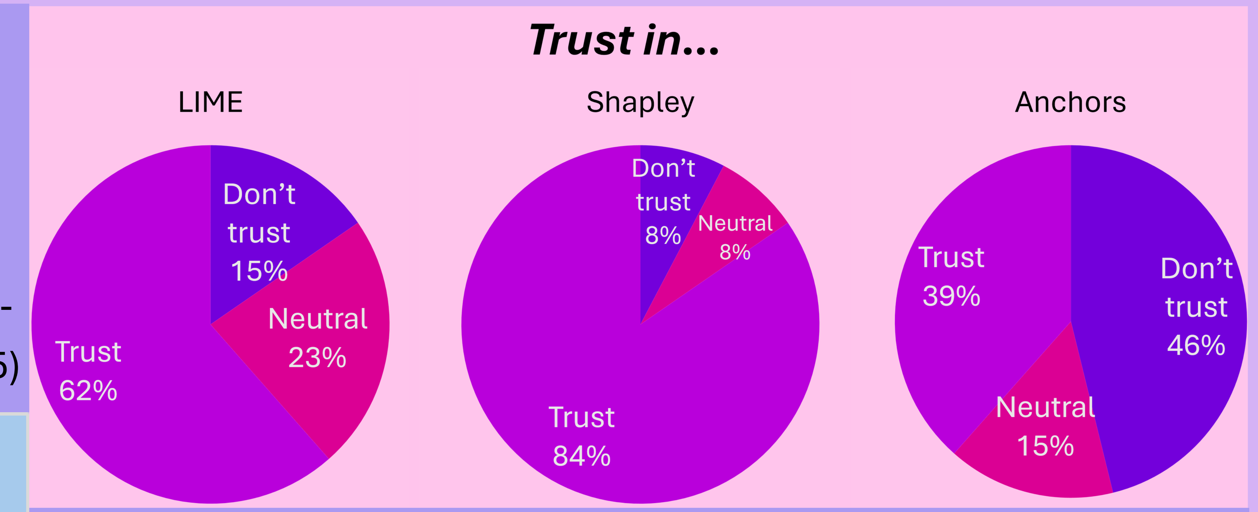
MOOC Model Confusion Matrix

outcome \ prediction	Not Completed	Completed
Not Completed	11,201 *(true -)*	17 *(false +)*
Completed	31 *(false -)*	342 *(true +)*

User Study

Current undergraduates and recent graduates were surveyed to gauge general reactions to the different XAI techniques presented here. After being introduced to the underlying model, respondents were shown a sample input, and an explanation, and answered a series of subjective, qualitative questions. (Overall *n*=31, MOOC *n* = 15)



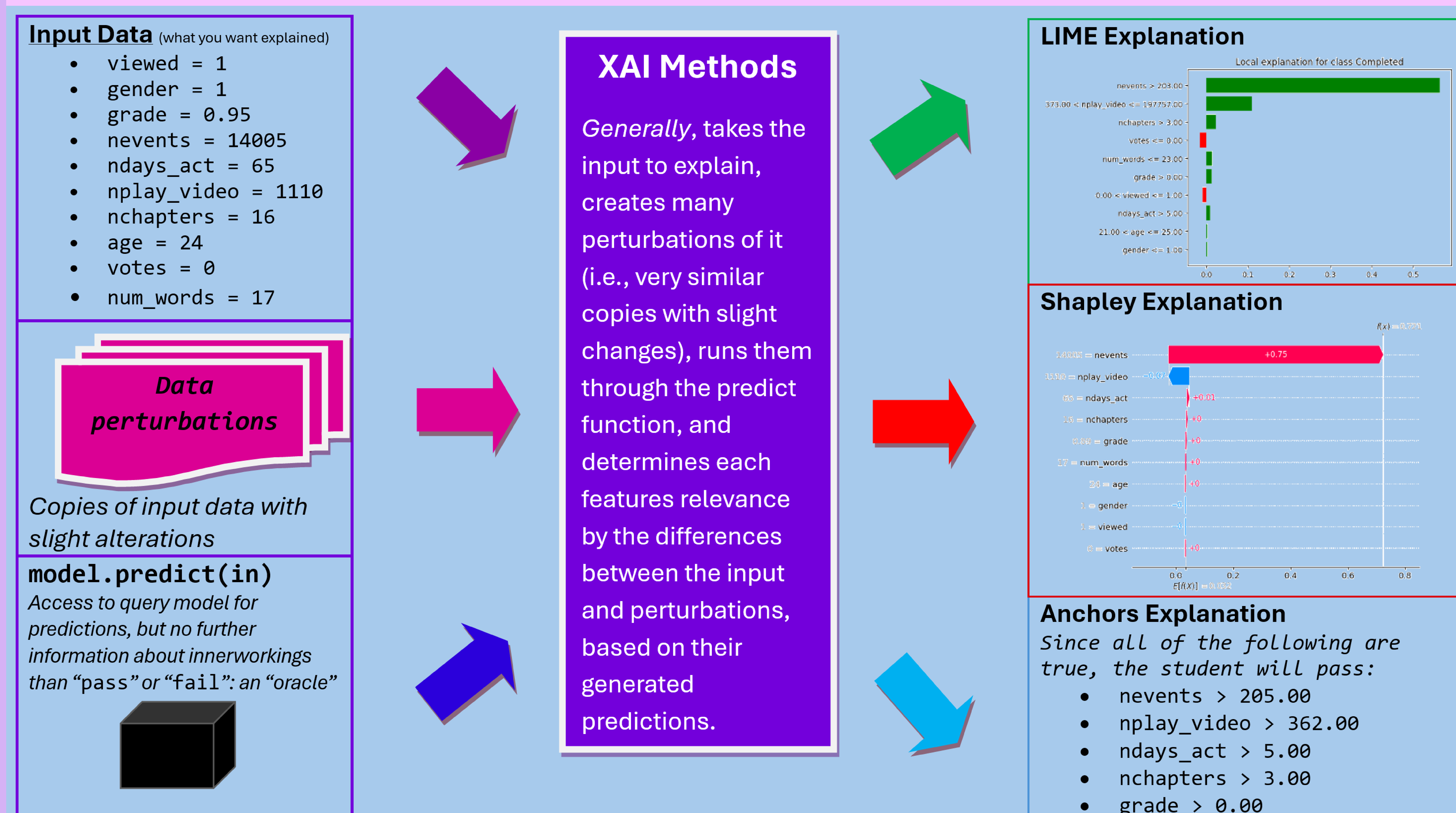


In general, Anchors preformed poorly in building trust in the underlying model in the respondents. When compared to the graphical representations presented by LIME and Shapley, the simple rules that did not weigh each feature were found to be less informative. However, due to its simple nature and clearly defined scope, it scored more 5 out of 5s for ease of understanding than LIME or Shaley. Anchor was most preferred among respondents who had little to no background in machine learning or STEM. Misinterpretation of all methods was common across respondents, demonstrating a need to understand the techniques and their scope before being implemented.

"I think it makes it more clear that these models follow rules, even if we don't know what the rules specifically are. It really breaks down the idea of the black box model."

"Anchor made me really trust the model. When I read the rules for anchor, I felt that the rules were so rigid that I could be most comfortable predicting."

"Anchor gives too hard rules with too little nuance."



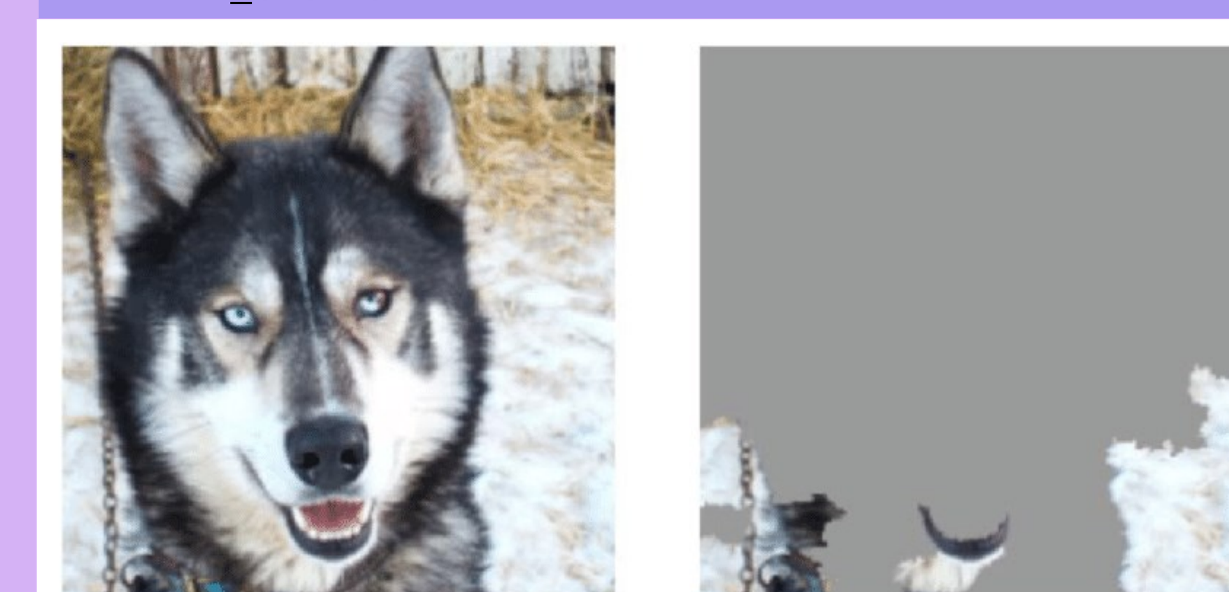
preet@carleton.edu ● https://github.com/cosmcbun/Explainable-Ai-Comps-2024

Anchors

Anchors differentiates itself from LIME and Shapley in two ways: first, visually, it presents information in a fundamentally different format, and second, it clearly defines its coverage. Instead of a chart that shows the significance of every feature as its explanation, Anchors provides a more succinct list of a handful of rules that *must be satisfied* to make the prediction. In doing this, Anchors clearly defines its scope (i.e., where the explanation it gives is applicable) — this is an issue for Shapley and LIME, because it is unclear to the reader how much can be extrapolated from the given explanation.

Conclusions

- There can be discrepancies across explanations
  - They take fundamentally different approaches to finding an explanation
- Troubled by confirmation bias
  - Respondents were more likely to find an explanation that reflected their beliefs useful
- *Explainable* AI is makes only local explanations
  - *Interpretable* AI makes global explanations
- These techniques are remarkably easy to implement with "off the shelf" packages
- Great tool for identifying bias, oddities in model, e.g.,
  - Found age/gender bias
  - Realized its heavy reliance on nplay_videos, n_events



Acknowledgements & Citations

I would like to thank our advisor, Professor Anna Rafferty, as well as the group members, survey respondents, and friends for their guidance, hard work, time, and support respectively.

Vignesh Muthukumar and Bhalaji Natarajan, "MOOCVERSITY - Deep Learning Based Dropout Prediction in MOOCs over Weeks," *Journal of Soft Computing Paradigm* 2, no. 3 (June 27, 2020): 140–52, https://doi.org/10.36548/jscp.2020.3.001.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, "Anchors: High-Precision Model-Agnostic Explanations," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

Philippe Besse et al., "Can Everyday AI Be Ethical. Fairness of Machine Learning Algorithms," *arXiv.Org*, 2018, https://doi.org/10.48550/arxiv.1810.01729.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," *arXiv.Org*, 2016, https://doi.org/10.48550/arxiv.1602.04938.