Reliance in human-AI decision-making

Second Checkpoint

Stanisław Giziński, Michał Tyrolski, Emilia Wiśnios

Our study aimed to expand upon previous research by using a dataset that may be more intuitive for random people to understand and by adding a personality test and metadata such as gender, age, and education level. Our dataset contained information about individuals, including their age, workclass, number of years of education, marital status, occupation, relationship to family, race, and sex. After cleaning the dataset, there were 30162 records left.

The selected model for our study was Explainable Boosting Machines (EBMs), which are a type of machine learning model that uses boosting algorithms to make predictions and provide explanations for their predictions. In our study, we trained the EBM model on the income dataset and obtained both local and global explanations. We also selected examples for the user study and designed the study itself.

Unfortunately we can't disclose the details of our methodology, selected examples and information about global and local explanations, because a great number of people from our group will take part in our study.

Regarding the design of the study we want to divide people into 4 groups. Each of them will have the same examples but with different information:

- Control group (only features from the dataset), no additional information,
- Group with model prediction without any explanations,
- Group with model prediction and local explanations,
- Group with model prediction and global explanations.

Additionally we will ask about metadata (age, sex, education level) and 10-Item Personality Inventory. The study will be conducted in English. At the beginning of the test a quick guide will be provided. Each feature will be explained (but not in the XAI understanding) before the study. The final number of chosen examples will be determined after the testing phase (we need to determine how long it is to complete the study).

Our next steps will involve conducting the study using the aforementioned schema, analyzing the results, and preparing a final report. The study will be implemented in the internal tool of the Warsaw University of Technology and will be conducted on students from various faculties (MIM, Political Sciences, Physics, Psychology).

Reference: Johannes Jakubik, Jakob Schöffer, Vincent Hoge, Michael Vössing, and Niklas Kühl. An Empirical Evaluation of Predicted Outcomes as Explanations in Human-AI Decision-Making, 2022. 6.