HELPING NON-EXPERT STAKEHOLDERS WITH STRATEGIES THROUGH EXPLANATION OF THE DECISION-MAKING ALGORITHM

Hari Krishna Chilakala Assistant Professor, RISE Krishna Sai Gandhi Group of Institutions, Ongole, ch.harikrishna2003@gmail.com
Lavanya Baviri Associate Professor, RISE Krishna Sai Gandhi Group of Institutions, Ongole, laavanyajayakrishna@gmail.com

ABSTRACT

Increasingly, algorithms are used to make important decisions across society. However, these algorithms are usually poorly understood, which can reduce transparency and evoke negative emotions. In this research, we seek to learn design principles for explanation interfaces that communicate how decision-making algorithms work, in order to help organizations explain their decisions to stakeholders, or to support users' "right to explanation". We conducted an online experiment where 199 participants used different explanation interfaces to understand an algorithm for making university admissions decisions. We measured users' objective and self-reported understanding of the algorithm. Our results show that both interactive explanations and "whitebox" explanations (i.e. that show the inner workings of an algorithm) can improve users' comprehension. Although the interactive approach is more effective at improving comprehension, it comes with a trade-off of taking more time. Surprisingly, we also find that users' trust in algorithmic decisions is not affected by the explanation interface or their level of comprehension of the algorithm.

CCS CONCEPTS • Human-centered computing → Interactive systems and tools; Empirical studies in HCI;

KEYWORDS Algorithmic Decision-making, Explanation Interfaces

INTRODUCTION

Automated and artificially intelligent algorithmic systems are helping humans make important decisions in a wide variety of domains. To name a few examples, recidivism risk assessment algorithms such as COMPAS have been used to help judges decide whether defendants should be detained or released while awaiting trial. Allegheny County in Pennsylvania has been using an algorithm based on Predictive Risk Modeling (PRM) to help screen referral calls on child maltreatment . And according to an article in The Wall Street Journal, the proportion of large companies using Applicant Tracking Systems to automatically filter and rank applicants is in the "high 90%" range . Researchers, government bodies, and the media have argued that data users should have the "right to explanation" of all decisions made or supported by automated or artificially intelligent algorithms. The approval in 2016 of the European Union General Data Protection Regulation (GDPR) mandates that data subjects receive meaningful information about the logic involved in automated decision-making systems. However, it is challenging for people who are not algorithm experts to understand algorithmic decision-making systems. Due to this literacy gap, recipients of the algorithm's output have difficulty understanding how or why the inputs lead to a particular outcome . The recent surge of interest in explainable artificial intelligence (XAI) has lead to great progress on transforming complex models (such as neural networks) into simple ones (such as linear models or decision trees) through approximation of the entire model or local approximation . Despite its mathematical rigor, there are recent critiques that this line of research is based on the intuition of researchers, rather than on a deep understanding of actual users. There is limited empirical evidence on whether these “intelligible models” and explanation interfaces are actually understandable, usable, or practical in real world situations . On the other hand, HCI researchers have conducted surveys, done interviews, and analyzed public tweets to understand how real-world users perceive and adapt to algorithmic systems.

The goal of this paper is to bridge these different research areas through conducting human-centered design and empirical comparisons of parallel interface prototypes to explore the effectiveness and trade-offs of different strategies to help non-expert stakeholders understand algorithmic decision making. We understand that there might not be a universally effective strategy for algorithmic decision making. Therefore, we focus on whether there are more effective strategies in the context of profiling, defined as the processing of personal data to evaluate certain aspects relating to a natural person 1 . In profiling tasks, the actual evaluation outcomes (e.g. the risk of offenders, or the suitability of applications to an organization or a university) are difficult to observe. We examined two sets of strategies for designing interfaces to explain algorithmic decision-making: white-box vs. black-box (i.e. showing the internal workings of an algorithm or not), and static vs. interactive (i.e. allowing users to explore an algorithm’s behavior through static visualizations or interactive interfaces). We conducted an online experiment where participants used four different explanation interfaces to understand an algorithm for making university admissions decisions. We developed measures to assess participants’ objective and self-reported understanding of the algorithm. Our results show that interactive explanations improved both objective and self-reported understanding of the algorithm, while “white-box” explanations only improved users’ objective understanding. Although the interactive approach is more effective for comprehension, it requires more of the user’s time. Surprisingly, we also found that users’ trust in algorithmic decisions was not affected by the explanation interface they are assigned to. The contributions of our work are three-fold. First, our work provides concrete recommendations for designing effective algorithm explanations. Second, our findings suggest nuanced trade-offs between different explanation strategies. Third, we provide a framework to evaluate algorithmic understanding with end users in a real world application. Future researchers can use and adapt our framework to evaluate algorithmic understanding in other domains.

RELATED WORK AND RESEARCH QUESTIONS

Algorithmic Decision-making

We define “algorithmic decision-making”, or simply “algorithm”, as the processing of input data to produce a score or a choice that is used to support decisions such as prioritization, classification, association, and filtering . In some settings, algorithmic decision-making systems have been used to completely replace human decisions. But in most real-world scenarios, there is a human operator involved in the final decision, who is influenced by the algorithm’s suggestions and nudging . In this paper, we focus on algorithms generated through supervised machine learning-based approaches. The first step is to define a prediction target, often a proxy for the actual evaluation outcome. With reference to the examples cited above, this might consist of whether a defendant will be charged with a crime if released, whether a child will be removed from their home and placed in care, or whether a job applicant will receive a job offer. The second step is to use labeled training data, often in large volumes, to train and validate machine learning models. Finally, validated models are applied to new data from incoming cases in order to generate predictive scores. Note that in this paper, the goal is to help users and other stakeholders understand the "algorithmic decision model", rather than the process of model training.

Explaining and Visualizing Machine Learning

Applied Machine Learning (ML) and visualization communities have long been working on developing techniques and tools to explain and visualize ML algorithms and models (e.g. [26]). However, there are two challenges in directly applying these techniques to help non-experts understand algorithmic decision-making, particularly profiling. First, the majority of these techniques and tools are designed to support expert users like data scientists and ML practitioners or serve educational purposes for people who are machine learning novices but often have good technical literacy. For instance, these tools often depend on performance measures (e.g. accuracy, precision, recall, confusion matrices, and area under the ROC curve measures) to help people understand and compare different models; these techniques might not help non-expert users, especially those with low technical literacy.

White-box vs. Black-box Explanation

There are two distinct approaches for explaining algorithms: the "white-box" approach (i.e. explaining the internal workings of the model) and the "black-box" approach (i.e. explaining how the inputs relate to the outputs without showing the internal workings of the model). Examples of the whitebox technique include showing probabilities of the nodes for Bayesian networks [4], projection techniques [9] and Nomograms [28] to see the "cut" in the data points for Support Vector Machines, and the visualization of the graph of a neural network. In contrast to the white-box approach, the black-box approach focuses on explaining the relationships between input and output, regardless of how complicated the model itself is. For example, Krause et al. design an analytics system , Prospector, to help data scientists understand how any given feature affects algorithm prediction overall. Plate et al. and Olden propose methods to show how input features influence the outcome of neural network classifications. Martens and Provost show removal-based explanations such as "the classification would change to [name of an alternative class] if the words [list of words] were removed from the document." However, we posit that the relative strengths and weaknesses of white-box and black-box approaches in helping non-expert users understand profiling algorithms remain unestablished. For example, one possible trade-off is that the white-box approach can give users a comprehensive understanding of the model, but might cause information overload and create barriers for users who are not technologically savvy .

Explanation Interface Prototypes

We created four interface prototypes (white-box interactive, white-box static, black-box interactive, and black-box static) to explain student admission algorithms (see Figure 1(a)). All versions of the interfaces followed the principles below:

(1) We used a “card”-based design. The interface presents a student’s fifteen attributes and corresponding values, as well as the algorithm’s decision (i.e. strong accept, weak accept, weak reject, and strong reject). The users could obtain a quick overview of all the information relevant to one student.

(2) We presented the student’s attributes in groups. Specifically, we categorized the fifteen attributes into four groups (test scores, academic performance, application materials and additional attributes). Detailed description of the attributes was provided when users hovered over their labels. Next, we describe how the interface of the tool varies between different explanation strategies.

Experimental Design

To evaluate the effectiveness of the four explanation interfaces, we conducted a randomized between-subject experiment on Mechanical Turk. We used a 2x2+1 design, resulting in five conditions: white-box interactive, white-box static, black-box interactive, black-box static, and control. In the first four conditions, participants were given access to the explanation interface of the respective condition. They were allowed to spend as much time as they needed to understand the algorithmic decision with help of the interface. In the control condition, participants were only provided a static webpage which displayed the list of attributes considered by the algorithm.

LIMITATIONS AND FUTURE WORK

As with any study, it is important to note the limitations of this work. One concern is the choice of using an experimental approach. While the experimental approach allows us to draw causal conclusions, it limits our ability to observe how the users actually interact with the explanation tools. In the future work, we will observe how students and admission committees actually use the admission algorithm and the explanation interfaces in real world settings, which can potentially complement our experimental findings. Supporting users’ “right to explanation” is an important issue in a wide variety of domains that involve algorithmic decision-making. We have developed and evaluated explanation strategies and interfaces in the specific context of student admission. Future work is needed to use different contexts to replicate and validate our findings. Through replication, we can either validate our findings, or better understand the circumstances in which these findings do or do not apply.

CONCLUSION

Artificial intelligence is rapidly shaping modern society towards increased automation, in some cases making important decisions that affect human welfare. We believe that HCI researchers should strive to find ways to help people understand the automated decisions that affect their livelihood. In this paper, we took steps toward that goal by examining user interface strategies for explaining profiling algorithms. We found that our experimental interfaces increased algorithm comprehension, and that features supporting interacting with and visualizing the inner workings of an algorithm help improve users’ objective comprehension.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. ACM, 582.

[2] Saleema Amershi, Max Chickering, Steven M Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. 2015. Modeltracker: Redesigning performance analysis tools for machine learning. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. ACM, 337–346.

[3] Julia Angwin. 2016. Make algorithms accountable. The New York Times 1 (2016), 168.

[4] Barry Becker, Ron Kohavi, and Dan Sommerfield. 2002. Visualizing the Simple Bayesian Classifier. In Information visualization in data mining and knowledge discovery. Morgan Kaufmann, 237–249.

[5] Victoria Bellotti and Keith Edwards. 2001. Intelligibility and accountability: human considerations in context-aware systems. Human– Computer Interaction 16, 2-4 (2001), 193–212.

[6] Or Biran and Courtenay Cotton. 2017. Explanation and justification in machine learning: A survey. In IJCAI-17 Workshop on Explainable AI (XAI). 8.

Authors

Hari Krishna Chilakala obtained his master degree in computer science from Jawaharlal Nehru Technological University, Kakinada in 2014. His research interest includes: Data Mining, Data Science and Machine Learning.

Lavanya Baviri obtained her master degree in computer science engineering from Jawaharlal Nehru Technological University, Kakinada in 2011. Her research interest includes: Data Mining, Artificial Intelligence.