Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning*

Harmanpreet Kaur,[1] Harsha Nori,[2] Samuel Jenkins,[2]
Rich Caruana,[2] Hanna Wallach,[2] Jennifer Wortman Vaughan[2]
[1]University of Michigan, [2]Microsoft Research
harmank@umich.edu,{hanori,sajenkin,rcaruana,wallach,jenn}@microsoft.com

ABSTRACT

Machine learning (ML) models are now routinely deployed in domains ranging from criminal justice to healthcare. With this newfound ubiquity, ML has moved beyond academia and grown into an engineering discipline. To that end, interpretability tools have been designed to help data scientists and machine learning practitioners better understand how ML models work. However, there has been little evaluation of the extent to which these tools achieve this goal. We study data scientists' use of two existing interpretability tools, the InterpretML implementation of GAMs and the SHAP Python package. We conducted a contextual inquiry (N=11) and a survey (N=197) of data scientists to observe how they use interpretability tools to uncover common issues that arise when building and evaluating ML models. Our results indicate that data scientists over-trust and misuse interpretability tools. Furthermore, few of our participants were able to accurately describe the visualizations output by these tools. We discuss implications for researchers and tool designers, and contextualize our findings in the social science literature.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Human-centered computing** → **User studies**.

KEYWORDS

interpretability, machine learning, user-centric evaluation

1 INTRODUCTION

Machine learning (ML) has become ubiquitous in our everyday lives in domains ranging from criminal justice and public policy to healthcare and education. Modern ML models show promise on problems in computer vision and natural language processing for which solutions were only recently out of reach. Moreover, while ML was once confined to the academic community, it has now grown into a full engineering discipline.

These developments create countless opportunities for impact, but with these opportunities come new challenges. ML models have been found to amplify societal biases in datasets and lead to unfair outcomes [1, 5, 10]. When ML models have the potential to affect people's lives, it is critical that their developers are able to understand and justify their behavior. More generally, data scientists and machine learning practitioners cannot debug their models if they do not understand their behavior. Yet the behavior of complex ML models like deep neural networks and random forests is notoriously difficult to understand.

Faced with these challenges, the ML community has turned its attention to the design of techniques aimed at *interpretability*[1] [7, 12]. These techniques generally take one of two approaches. First, there are ML models that are designed to be inherently interpretable, often due to their simplicity, such as point systems [9, 21] or generalized additive models (GAMs) [6]. Second, there are techniques that provide post-hoc explanations for the predictions made by complex models, such as local interpretable model-agnostic explanations (LIME) [16] and Shapley additive explanations (SHAP) [14].

Although the number of proposed techniques continues to grow, there has been little evaluation of whether they help stakeholders achieve their desired goals. User studies of interpretability are extremely challenging for a number of reasons. They require expertise in the mathematics underlying ML models and in human–computer interaction (HCI), as well as knowledge of both the academic literature and day-to-day engineering practices. To paint a full picture, studies must rely on qualitative methods to understand the nuances of how tools are used in context, and quantitative methods to scale up findings. They must also mimic realistic settings, yet not be too cumbersome (e.g., take over an hour to complete).

We study the effectiveness of interpretability tools for one key stakeholder group: data scientists and machine learning practitioners. (For simplicity, we refer to this group simply as "data scientists" throughout the paper.) We perform a human-centric evaluation of two existing interpretability tools, the InterpretML implementation of GAMs and the SHAP Python package, in the context of building and evaluating ML models. Below, we present a high-level overview of our three-prong evaluation methodology, followed by a summary of our results and a discussion of implications for future work.

2 METHODOLOGY

Our evaluation methodology is comprised of three components that build on each other: 1) a series of pilot interviews ($N = 6$) to identify common issues faced by data scientists in their day-to-day work; 2) a contextual inquiry ($N = 11$) to observe data scientists' abilities to use interpretability tools to uncover these issues, and 3) a large-scale survey ($N = 197$) to scale up and quantify the main findings from our contextual inquiry.

Pilot Interviews. To better understand the issues that data scientists face in their data-to-day work—i.e., the setting in which interpretability tools would be used—we first conducted semi-structured interviews with six data scientists at a large technology company. The interview protocol was designed to surface common issues

*This is a shortened version of a CHI 2020 paper. The full paper is available here: https://doi.org/10.1145/3313831.3376219

[1]There is not yet consensus within the research community on the distinction between the terms *interpretability*, *intelligibility*, and *explainability*, and they are often, though not always, used interchangeably. Throughout this paper, we stick with *interpretability*, which is more commonly used within the machine learning community.

that arise when building and evaluating ML models. Based on an inductive thematic analysis of the interview transcripts, conducted via open coding followed by affinity diagramming [3], we identified six themes capturing common issues faced by data scientists. Five of these correspond to issues with data itself: missing values, temporal changes in the data, duplicate data masked as unique, correlated features, and ad-hoc categorization. The sixth theme relates to the difficulty of trying to debug or identify potential improvements to an ML model based on only a small number of data points. With only six interviews, we cannot claim that this list is at all exhaustive, but it is consistent with previous research on ML.

Contextual Inquiry. With these common issues in mind, we designed a contextual inquiry, intended to put data scientists in a realistic setting: exploring a dataset and an ML model in a hands-on fashion. We recruited eleven participants, each of whom was given a Jupyter notebook that included a dataset, an ML model that we had trained using that dataset, an interpretability tool (either the InterpretML implementation of GAMs or the SHAP Python package used to explain a LightGBM model, selected at random) that we had set up, and several questions to answer. Our goal was to observe whether they were able to use the interpretability tool to uncover the issues identified via our pilot interviews.

Participants explored the dataset, model, and interpretability tool on their own. For each tool, we provided examples of all three types of visualization—i.e., global explanations, additive components (GAMs) or dependence plots (SHAP), and local explanations. Participants answered ten questions about the dataset and model. Four were general questions about the visualizations (e.g., "What are the most important features that affect the output, according to the explanation above?"), while the rest were designed to get at the issues identified in our pilot. Answering these questions required the use of all three types of visualization. For each question, we also asked each participant to rate their confidence in their understanding of the visualizations and their confidence that these explanations were reasonable, on a scale of 1 (not at all) to 7 (extremely).

Survey. We designed a survey to scale up and quantify our findings from the contextual inquiry and shed light on data scientists' mental models of interpretability tools. Similar to our contextual inquiry, the survey placed data scientists in a realistic setting. The dataset, models, and interpretability tools used were identical to those used in the contextual inquiry.

As in our contextual inquiry, each participant used only one interpretability tool, selected at random. For the survey, we also showed participants either "normal" or "manipulated" visualizations, again selected at random. In the normal-visualization condition, we showed participants the visualizations output by the interpretability tools. However, in the manipulated-visualization condition, we instead showed participants global and local explanations where the input feature names had been rearranged, resulting in the input features with smallest contributions to the predictions being displayed as the most important, and vice versa. We designed this manipulation to test the extent to which participants' perception and use of the interpretability tools depend on how reasonable their explanations are (versus the mere existence of visualizations).

We asked each participant to answer four blocks of questions about the dataset and the model, covering global feature importance, the relationship between one specific feature and the output variable, the local explanation for a correctly classified data point, and the local explanation for a misclassified data point, respectively. Each of these blocks contained seven questions: (1) a multiple-choice question with a ground-truth correct answer, which was designed to quantify the participants' accuracy at reading the visualizations (e.g., "Which is the 3rd most important feature for the underlying model, according to the explanation system?"); (2) an open-ended question designed to test how well participants understood the visualizations and whether any suspicions arose; (3) a question about which visualizations they had used to answer the previous questions; (4) their stated confidence in their understanding of the visualizations (on a scale of 1–7) ; (5) their stated confidence that these explanations were reasonable (on a scale of 1–7); (6) their stated confidence that the underlying model was reasonable (on a scale of 1–7) ; and (7) an optional open-ended text field for comments or concerns.

3 RESULTS AND DISCUSSION

Our results indicate that the visualizations output by interpretability tools can sometimes—though not always—help data scientists to uncover the kinds of issues that they deal with on a day-to-day basis, such as missing values in a dataset that have been filled in incorrectly. We found that the choice of interpretability tool matters, with participants performing better when using the InterpretML implementation of GAMs. However, for both tools, the existence of visualizations and the fact that they were publicly available led to cases of over-trust and misuse. Furthermore, despite being provided with standard tutorials, few of participants were able to accurately describe what the visualizations were showing. Participants were also biased toward model deployment, despite recognizing suspicious aspects of the ML models. This was true even when we showed them manipulated, nonsensical explanations, though we observed this less with data scientists who were more experienced.

Interpretability is typically viewed as being unidirectional, with tools providing information to user. However, it may be better to design interpretability tools that allow back-and-forth communication [2]. Social science and HCI research consider this kind of back-and-forth to be a key factor in making explanations accessible to people with different levels of expertise [8, 15]. Weld and Bansal [18] propose interactive interpretability tools that allow users to dig deeper into explanations or to compare explanations from multiple different interpretability techniques. One might also imagine a tool that could update its mode of interactivity based on users' perceptions [13]. More generally, interpretability tools should be designed to adapt to users' expectations.

Research on interpretability in the ML and HCI communities has evolved somewhat independently [19]. Our results highlight the value of user studies for evaluating interpretability techniques from the ML community with stakeholders, marrying the goals and methods of both communities. Ideally, the HCI and ML communities should work together from the start, with HCI methodologies applied at all stages of interpretability tool development: supporting need-finding studies (e.g., [4, 17]), designing tools that can be understood by users with different backgrounds (e.g., [11]), and undertaking user studies at each stage of tool development (e.g., [20]).

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Kirchner Lauren. 2016. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica, May* 23 (2016), 2016. http://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[2] Dean C Barnlund. 2017. A transactional model of communication. In *Communication theory, Second edition*, C. David Mortensen (Ed.). Routledge, 47–57.

[3] Virginia Braun and Victoria Clarke. 2012. Thematic analysis. In *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.* American Psychological Association, Washington, DC, US, 57–71. https://doi.org/10.1037/13620-004

[4] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 104 (Nov. 2019), 24 pages. https://doi.org/10.1145/3359206

[5] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186. https://doi.org/10.1126/science.aal4230 arXiv:https://science.sciencemag.org/content/356/6334/183.full.pdf

[6] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney, NSW, Australia) *(KDD '15)*. ACM, New York, NY, USA, 1721–1730. https://doi.org/10.1145/2783258.2788613

[7] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[8] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. 2019. Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, Article 579, 13 pages. https://doi.org/10.1145/3290605.3300809

[9] Jongbin Jung, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G Goldstein. 2017. Simple rules for complex decisions. *Available at SSRN 2919024* (2017). http://dx.doi.org/10.2139/ssrn.2919024

[10] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. ACM, New York, NY, USA, 3819–3828. https://doi.org/10.1145/2702123.2702520

[11] Himabindu Lakkaraju and Osbert Bastani. 2020. " How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA) *(AIES '20)*. ACM, New York, NY, USA.

[12] Zachary C Lipton. 2016. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490* (2016).

[13] Stine Lomborg and Patrick Heiberg Kapsch. 2019. Decoding algorithms. *Media, Culture & Society* (2019). https://doi.org/10.1177/0163443719855301

[14] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[15] Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. In *IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI)*.

[16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. ACM, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[17] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, Article 601, 15 pages. https://doi.org/10.1145/3290605.3300831

[18] Daniel S. Weld and Gagan Bansal. 2019. The Challenge of Crafting Intelligible Intelligence. *Commun. ACM* 62, 6 (May 2019), 70–79. https://doi.org/10.1145/3282486

[19] Qian Yang, Nikola Banovic, and John Zimmerman. 2018. Mapping Machine Learning Advances from HCI Research to Reveal Starting Places for Design Innovation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 130, 11 pages. https://doi.org/10.1145/3173574.3173704

[20] Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. 2018. Investigating How Experienced UX Designers Effectively Work with Machine Learning. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) *(DIS '18)*. ACM, New York, NY, USA, 585–596. https://doi.org/10.1145/3196709.3196730

[21] Jiaming Zeng, Berk Ustun, and Cynthia Rudin. 2017. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180, 3 (2017), 689–722. https://doi.org/10.1111/rssa.12227