Does Stated Accuracy Affect Trust in Machine Learning Algorithms?
(Extended Abstract)

Ming Yin [1] [2] Jennifer Wortman Vaughan [1] Hanna Wallach [1]

In recent years, machine learning (ML) has become increasingly important as a tool to aid human decision making. Researchers have trained deep neural networks to help dermatologists identify skin cancer (Esteva et al., 2017), while political strategists regularly leverage forecasts produced by ML models when determining their next moves (Nickerson & Rogers, 2014). Prompted in part by this increase in human use of ML predictions and in part by new regulations such as the EU GDPR, researchers have turned their attention to the interpretability of ML systems (e.g., Ribeiro et al., 2016; Lipton, 2016; Doshi-Velez & Kim, 2017).

To date, most of the work on interpretability has focused explicitly on ML models themselves, asking questions about people's abilities to understand model internals or the way that a particular model maps inputs to outputs. However, the model is just one component of the ML pipeline, which spans data collection, training algorithms and procedures, model evaluation, and ultimately, deployment. One particularly under-explored aspect of interpretability in the model evaluation and deployment stages of the pipeline is the interpretability of performance metrics. For example, how well do people understand the relationship between a model's performance on held-out data and the model's expected performance post deployment? And how does such understanding influence people's willingness to trust a model?

In this paper, we focus on this under-explored aspect of interpretability. We take an experimental approach to measure the impact of one particular performance metric—a model's stated accuracy on held-out data—on people's trust in the model. We report the results of a large-scale randomized human subject experiment in which subjects made predictions about the outcome of speed dating events with the help of an ML model. Subjects were first shown information about a speed dating participant and his or her date and asked to predict whether or not the participant would want to see the date again. They were then shown a prediction from the model and given the option to revise their prediction.

Subjects were randomized into one of ten treatments, which differed along two dimensions. The first was the stated accuracy of the model. Some subjects were given no background information on its accuracy, while others were told that the model's accuracy on some held-out data was either 60%, 70%, 90%, or 95%. Halfway through the experiment, all subjects were given feedback on both own accuracy and on the model's accuracy (always 80% by design) on the first half of the prediction tasks. This design allowed us to rigorously test whether the stated accuracy of the model influenced people's trust in the model, both before and after they observed the model's performance in practice.

The second dimension was whether or not subjects were rewarded for making correct predictions. In the high-stakes treatments, subjects received a flat "base" payment as well as a bonus of $0.10 for each correct prediction, while in the low-stakes treatments they received only the base payment for completing the experiment. This design allowed us to test whether subjects would be more or less likely to follow the model when they had more "skin in the game."

We found that the stated accuracy of the model did have a significant effect on the extent to which people trust the model, as measured by both the frequency with which subjects adjusted their predictions to match those of the model and by subjects' self-reported levels of trust in the model, although the effect size was smaller after subjects received feedback about the model's performance. We did not detect a significant effect of prediction stakes on trust. Finally, subjects consistently chose to revise their predictions to match those of the model more often and reported higher levels of trust in the model after receiving feedback on the model's performance, regardless of whether the observed accuracy was higher or lower than the initial stated accuracy. We conjecture that this is because the model's observed accuracy (80%) was much higher than most subjects' accuracies.

These results highlight the need for developers of ML models to clearly and responsibly communicate their expectations about model performance since this information shapes the extent to which people trust a model, both before and after they are able to interact with it and observe its perfor-

[1] Microsoft Research, New York, NY, USA [2] Department of Computer Science, Purdue University, West Lafayette, IN, USA. Correspondence to: Ming Yin <mingyin@purdue.edu>.

2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018), Stockholm, Sweden. Copyright by the author(s).

mance first-hand. Our results also reveal that people put substantial weight on their own experiences with a model when deciding how much they should trust it. Of course, proper caution should be used when generalizing our results to other settings. In particular, although we did not observe a significant effect of prediction stakes on trust, it is entirely possible that there would be an effect when stakes are sufficiently high (e.g., doctors making life-or-death decisions).

References

Doshi-Velez, Finale and Kim, Been. Towards a rigorous science of interpretable machine learning. *CoRR arXiv:1702.08608*, 2017.

Esteva, Andre, Kuprel, Brett, Novoa, Roberto A, Ko, Justin, Swetter, Susan M, Blau, Helen M, and Thrun, Sebastian. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

Lipton, Zachary C. The mythos of model interpretability. In *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*, pp. 96–100, 2016.

Nickerson, David W and Rogers, Todd. Political campaigns and big data. *Journal of Economic Perspectives*, 28(2): 51–74, 2014.

Ribeiro, Marco Tulio, Singh, Sameer, and Guestrin, Carlos. Why should I trust you?: Explaining the predictions of any classifier. In *KDD*, 2016.