Manipulating and Measuring Model Interpretability

Forough Poursabzi-Sangdeh[1], Daniel G. Goldstein[2], Jake M. Hofman[2], Jennifer Wortman Vaughan[2], and Hanna Wallach[2]

[1]*University of Colorado Boulder*
[2]*Microsoft Research*
[1]*forough.poursabzisangdeh@colorado.edu*
[2]*{dgg,jmh,jenn,wallach}@microsoft.com*

Machine learning models are often treated as black boxes and evaluated only in terms of their performance (e.g., accuracy) on held-out data sets. However, good performance on a held-out data set is seldom sufficient for people to trust a model and deploy it to make real-world decisions, and there is widespread consensus that people's failure to understand a model can be problematic [1, 2]. In response to these concerns, there is a new line of research that focuses on developing *interpretable* methods for machine learning, either by developing new models that are inherently simple to understand [3] or by providing explanations or interpretations of existing complex models [4, 5, 6]. Despite the popularity of this line of research, there is no clear, agreed-upon definition of interpretability. Defining and quantifying interpretability therefore remains an open question.

Through large-scale randomized experiments, we vary factors that should make models more or less interpretable and, in turn, measure how these changes impact people's decision making. We ask each participant to predict apartment prices in New York City, with the help of a model (linear regression). We show participants models that receive the same inputs and produce the same outputs, manipulating only the presentation of the models. We vary the number of features (two vs. eight) and the visibility of the model internals (clear vs. black box) in a 2×2 between-subject study. As a baseline, we also ask participants to predict apartment prices without the help of a model. For each experimental condition, we show participants a set of apartments, the model's price predictions for those apartments, and the apartments' sale prices. Next, we show participants a new set of apartments. For each one, we ask them to guess the model's prediction. We then show them the model's prediction and ask them to guess the sale price. Drawing on previous work [7, 8], we measure three different proxies for interpretability: 1) Simulation error (the participant's error in guessing the model's prediction); 2) trust (the participant's confidence that the model has made the right prediction); and 3) prediction error (the participant's error in guessing the sale price).

Our preliminary results indicate that, on average, participants in the two-feature, clear-model-internals experimental condition have lower simulation error. Interestingly, participants in the eight-feature, black-box-model-internals experimental condition do as well as participants in the eight-feature, clear-model-internals experimental condition. This result suggests that the number of features affects model interpretability. Despite these differences in simulation error, we find that participants' prediction error is comparable across all four experimental conditions—i.e., participants appear to trust the models similarly.

We see this as the first of many possible experiments to guide the development of interpretable machine learning methods.

References

1. Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 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*, pages 1721–1730. ACM, 2015.
2. Been Kim. *Interactive and interpretable machine learning models for human machine collaboration*. PhD thesis, Massachusetts Institute of Technology, 2015.
3. Jongbin Jung, Connor Concannon, Ravi Shro, Sharad Goel, and Daniel G. Goldstein. Simple rules for complex decisions. *arXiv preprint arXiv:1702.04690*, 2017.
4. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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*, pages 1135–1144. ACM, 2016.
5. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Nothing else matters: Model-agnostic explanations by identifying prediction invariance. *arXiv preprint arXiv:1611.05817*, 2016.
6. Brian Y Lim, Anind K Dey, and Daniel Avrahami. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2119–2128. ACM, 2009.
7. Zachary C Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.
8. Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. 2017.