Toward Dispelling Unhelpful Explainable Machine Learning (ML) Misconceptions

Patrick Hall*
Washington, DC
patrick.hall@h2o.ai

Abstract

This short text presents counter-factual arguments, common sense proposals, and references to address recently uncovered misinformation and misconceptions about explainable machine learning in the data science community. It also argues that post-hoc explanatory methods are a viable tool in a holistic, interpretable approach to machine learning.

1 Introduction

"Please don't do explainable ML," extolled an important researcher in the title of a recent short talk. The same talk also included claims such as, "[Explainble ML] forces you to rely on two models instead of one," and intimated that explainable machine learning can be a foil for companies and governments to conduct unsavory deeds with unaccountable black-box models. [2]. Perhaps even more noteworthy was the online response, which included incorrect musings such as, "don't forget hidden assumptions for explainable ML (e.g., locally linear behavior near predictions)," and lamentations like, "no one has explained to me what 'explainable' or 'interpretable' is." [3]

In response, this short text aims to dispel several misconceptions about explainable ML and to fill in some obvious gaps in community knowledge. For clarity's sake and as illustrated in Figure 1, here explainble ML means post-hoc techniques used to understand trained model behavior or predictions. Examples of common explainable ML techniques include:

- Local and global feature importance methods, in particular Shapley values [13].
- Local and global model-agnostic surrogate models, such as surrogate decision trees and Local Interpretable Model-agnostic Explanations (LIME) [5], [3], [11], [14].
- Local and global visualizations of model predictions such as partial dependence and individual conditional expectation (ICE) plots [6], [8].

By presenting definitions for key terms and addressing misconceptions, this text builds a case for a holistic approach to ML that includes white-box models along with explanatory, debugging, and fairness techniques and also argues that ignoring an entire set of methods because *some subset* of the methods are approximate is akin to throwing the baby out with the bath water.

This text does not condone the use of black-box models with cursory applications of low fidelity post-hoc explanatory methods – that is likely lazy, irresponsible, and unethical, and also potentially dangerous.

*H2O.ai and George Washington University

[2]Statistics at a Crossroads, Webinar 2. URL: https://zoom.us/recording/play/0y-iI9HamgyDzzP2k_jiTu6jB7JgVVXnjWZKDMbnyRTn3FsxTDZy6Wkrj3_ekx4J?startTime=1538497702000

[3]Twitter thread: https://twitter.com/tdietterich/status/1052680788389507073

Preprint. Work in progress.

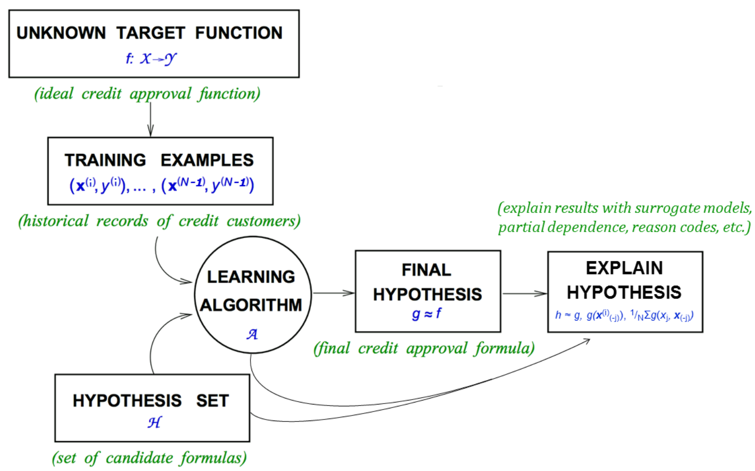

Figure 1: An augmented learning problem diagram in which several post-hoc techniques create explanations for a credit scoring model. Adapted from *Learning From Data* [1].

2 Misconception: All the Key Terms in Explainable ML are Undefined

While we are probably far from a true science, and concrete vocabulary, of interpretable machine learning, at least two helpful definitions have been coined previously.

- **Interpretable**: "The ability to explain or to present in understandable terms to a human" – in *Towards a Rigorous Science of Interpretable Machine Learning* by Finale-Doshi and Kim (2017).
- **A Good Explanation**: "When you can no longer keep asking why." – in *Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning* by Gilpin et al. (2018).

From a literal reading of these two well-founded definitions, it would certainly seem that explanations contribute to some process being interpretable.

3 Misconception: Explainable ML is Just Models of Models

Models of models, or surrogate models, can be helpful explanatory tools, but they are usually approximate, low-fidelity explainers. Aside from the facts that 1.) a course, global summary of a complex model provided by a surrogate model can be helpful and 2.) much work in explainable ML has been directed toward improving the fidelity and usefulness surrogate models [5], [3], [11], [15], *many explainable ML techniques have nothing to do with surrogate models!*

One of the most exciting breakthroughs in explainable ML is tree shap, a model-specific (i.e. not a surrogate model), rigorously defined, accurate and consistent, global and local feature importance measure [12]. There are many other model-specific explainable ML methods such as partial dependence and ICE plots [6], [8]. Also, surrogate models and model-specific explanatory techniques can be combined, for instance by using partial dependence, ICE, and surrogate decision trees to investigate and confirm modeled interactions [9].

For a curated list of many different types of white-box modeling, model debugging, and model-specific, model-agnostic, and surrogate model explainable ML techniques, please see:

`https://github.com/jphall663/awesome-machine-learning-interpretability`

Misconception Corollary: Explainable ML is just LIME. LIME, in it's most popular implementation, uses local linear surrogate models [14]. LIME is popular, important, and imperfect, but just one of many explainable ML tools. And again, LIME can sometimes be combined with model-specific methods to yield deeper insights. Consider that tree shap can provide locally accurate and consistent point estimates for local feature importance whereas LIME can provide information about modeled local linear trends around the same point.

4 Misconception: Explainable ML Methods Simply Provide Cover For Government and Commercial Entities to Use Black-Box ML for Nefarious Purposes

If used disingenuously, explainable ML methods probably do provide such cover, but explainable ML methods were designed to crack-open those same black-boxes. See Angwin et al. (2016) for evidence that this type of investigative analysis of commercial black-box models is possible [2]. Such investigations would likely only be improved by advances in explanatory, debugging, and fairness tools.

Additionally, many important computer-based technological advances present similar double-edged sword dilemmas, i.e. social media, strong encryption. Rarely does the ability of a tool to be misused for nefarious purposes disqualify it from being used as designed. Explainable AI methods need more debate and development. Dismissal and derision is unhelpful.

5 Misconception: Explainable ML Methods and White-box Models are Somehow Mutually Exclusive

Publications tend to focus on either white-box modeling techniques or on post-hoc explanations, but the two can be, and potentially should be, used together. Consider the seemingly useful case of augmenting globally interpretable models with local post-hoc explanations, or the converse, combining global explanatory methods with locally interpretable models.

- **Proposed globally interpretable model + local explainability method example**: Using a single pruned decision tree with local Shapley feature importance to see accurate feature contributions for each model prediction.
- **Proposed locally interpretable model + global explainability method example**: Combining a locally interpretable rule-based classifier, that produces a rule list for each prediction, with partial dependence plots to aid in understanding the complex rule-based response function w.r.t. to each model input or pairwise combination of inputs.

Corollary Misconception: Explainable ML Methods and Fairness Methods are Somehow Mutually Exclusive

Like white-box models, fairness methods are often presented in different articles than post-hoc explanatory methods. However, in banks, using partial dependence plots for model validation and disparate impact analysis for fair lending purposes for the same model is common place.

6 Conclusion

References

[1] Yaser S. Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. ***Learning from Data***. AMLBook, New York, 2012. URL: `https://work.caltech.edu/textbook.html`.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks. *ProPublica*, 2016. URL: `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

[3] Osbert Bastani, Carolyn Kim, and Hamsa Bastani. Interpreting Blackbox Models via Model Extraction. *arXiv preprint arXiv:1705.08504*, 2017. URL: `https://arxiv.org/pdf/1705.08504.pdf`.

[4] Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3), 2001. URL: `https://projecteuclid.org/euclid.ss/1009213726`.

[5] Mark W. Craven and Jude W. Shavlik. Extracting Tree-structured Representations of Trained Networks. *Advances in Neural Information Processing Systems*, 1996. URL: `http://papers.nips.cc/paper/1152-extracting-tree-structured-representations-of-trained-networks.pdf`.

[6] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. ***The Elements of Statistical Learning***. Springer, New York, 2001. URL: `https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12.pdf`.

[7] Jerome H. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 29:1189–1232, 1999. URL: `https://statweb.stanford.edu/~jhf/ftp/trebst.pdf`.

[8] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics*, 24(1), 2015. URL: `https://arxiv.org/pdf/1309.6392.pdf`.

[9] Patrick Hall. On the art and science of machine learning explanations. *arXiv preprint arXiv:1810.02909*, 2018. URL:`https://arxiv.org/pdf/1810.02909.pdf`.

[10] Patrick Hall, Wen Phan, and K Whitson. The Evolution of Analytics, 2016. URL: `https://pdfs.semanticscholar.org/cc62/c04074334d1d39b1c9f6a47b1ada99858529.pdf`.

[11] Linwei Hu, Jie Chen, Vijayan N. Nair, and Agus Sudjianto. Locally Interpretable Models and Effects Based on Supervised Partitioning (LIME-SUP). *arXiv preprint arXiv:1806.00663*, 2018. URL: `https://arxiv.org/ftp/arxiv/papers/1806/1806.00663.pdf`.

[12] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv preprint arXiv:1802.03888*, 2018. URL: `https://arxiv.org/pdf/1706.06060.pdf`.

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

[14] 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. URL: `http://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf`.

[15] Joel Vaughan, Agus Sudjianto, Erind Brahimi, Jie Chen, and Vijayan N Nair. Explainable Neural Networks Based on Additive Index Models. *arXiv preprint arXiv:1806.01933*, 2018. URL: `https://arxiv.org/pdf/1806.01933.pdf`.