Toward Dispelling Unhelpful Explainable Machine Learning (ML) Misconceptions

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

Pramit Chaudhary[†]
Los Angeles, CA
pramit.chaudhary@h2o.ai

Abstract

This short text presents arguments, proposals, and references to address recently uncovered misinformation and misconceptions about explainable machine learning. It also argues that post-hoc explanatory methods are one of several viable types of tools in a holistic, interpretable approach to machine learning.

1 Introduction

"Please stop doing explainable ML," extolled one of the brightest minds in machine learning with the title of a recent, well-reasoned, and admittedly controversial short talk. The same talk also included points such as, "[Explainable ML] forces you to rely on two models instead of one," and argued that explainable machine learning can be a foil for companies and governments to conduct unsavory or negligent deeds with black-box models. [3] Perhaps even more noteworthy was the online response, which included 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." [4] Strangely, neither the talk nor the follow-up discussions seemed to allow for combining white-box models and post-hoc explanatory techniques. This article aims to clear the misconceptions and fill the gaps in community knowledge exposed by these recent discussions.

To avoid ambiguity and as illustrated in the figure below, here explainable 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 [17], [11], [18], [14].
- Local and global model-agnostic surrogate models, such as surrogate decision trees and Local Interpretable Model-agnostic Explanations (LIME) [4], [3], [10], [16].
- Local and global visualizations of model predictions such as 1- and 2-dimensional partial dependence and individual conditional expectation (ICE) plots [6], [8].

By presenting definitions for key terms and at the same time 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. The article also implies that ignoring an entire set of methods

[*] H2O.ai and George Washington University

[†] H2O.ai

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

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

Preprint. Work in progress.

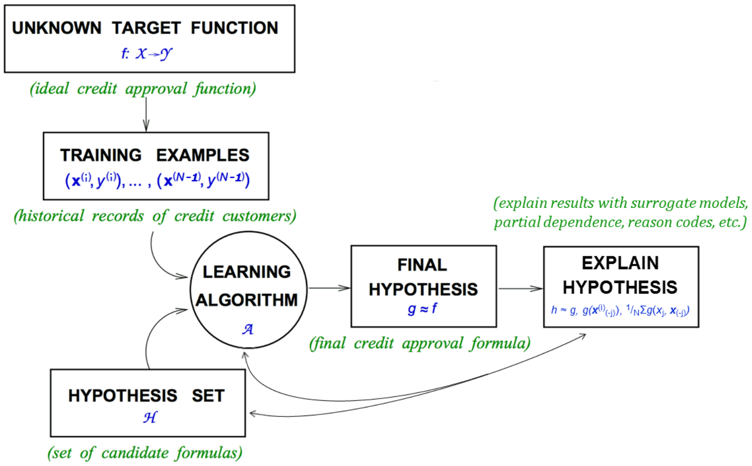
UNKNOWN TARGET FUNCTION
$f: \mathcal{X} \rightarrow \mathcal{Y}$
(ideal credit approval function)
TRAINING EXAMPLES
$(\mathbf{x}^{(i)}, y^{(i)}), \ldots, (\mathbf{x}^{(N-1)}, y^{(N-1)})$
(historical records of credit customers)
LEARNING ALGORITHM
$\mathcal{A}$
FINAL HYPOTHESIS
$g \approx f$
(final credit approval formula)
EXPLAIN HYPOTHESIS
$h \approx g,\ g(\mathbf{x}^{(i)}_{(-j)}),\ {}^{1}/_{N}\Sigma g(x_j, \mathbf{x}_{(-j)})$
(explain results with surrogate models, partial dependence, reason codes, etc.)
HYPOTHESIS SET
$\mathcal{H}$
(set of candidate formulas)

Figure: An augmented learning problem diagram in which several post-hoc techniques create explanations for a credit scoring model g. Explanations can increase understanding of g and help improve subsequent applications of $\mathcal{H}$ and $\mathcal{A}$. Adapted from *Learning From Data* [1].

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 only cursory applications of low fidelity post-hoc explanatory methods – that practice is likely lazy, irresponsible, and unethical, and also potentially dangerous.

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

While far from a complete vocabulary, at least two helpful definitions that apply to explainable ML have been put forward.

- **Interpretable**: "The ability to explain or to present in understandable terms to a human" – in *Towards a Rigorous Science of Interpretable Machine Learning* by Doshi-Velez and Kim (2017) [5].
- **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) [7]. (Gilpin et al. also provide several clear constructs for describing more specific types of explanations.)

From a literal reading of these two well-founded definitions, it would certainly appear that explanations contribute to some process being interpretable. Moreover, other authors have made significant attempts to grapple with a variety of concepts related to interpretability and explanations, including Zachary Lipton's *The Mythos of Model Interpretability* (2016) and Christoph Molnar's *Interpretable Machine Learning* (2018) [12], [15].

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 of surrogate models [4], [3], [10], [19], *many explainable ML techniques have nothing to do with surrogate models!*

One of the most exciting breakthroughs for supervised learning problems in explainable ML is the application of a coalitional game theory concept, Shapley values, to compute feature contributions which are consistent globally and accurate locally using the trained model itself [18], [14]. An extension of this idea, called Tree SHAP, has already been implemented for popular tree ensemble methods [13]. There are many other explainable ML methods that operate on trained models directly such as partial dependence and ICE plots [6], [8]. Notably, surrogate models and explanatory techniques that operate directly on trained models 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 [16]. LIME is 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 approximate 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 specifically to crack open those same nefarious and complex black-boxes. See Angwin et al. (2016) for evidence that 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 or strong encryption. Rarely does the ability of a tool to be misused for malicious purposes disqualify it from being used as designed. Many explainable ML methods have already been implemented into open source software and are being used somewhat widely. The techniques need more public debate, scrutiny, and development, not dismissal and derision.

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 cases 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, numeric feature contributions for each model prediction in addition to the entire directed graph of the decision tree.
- **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

This short text is not an attack on any party. Much credible work has been done in several disciplines to make machine learning more interpretable, and thus better as a science. All of that work is likely less valuable in silos, and likely more valuable and practical when used in combination. Because work in white-box models, or in explanations, or in fairness, or in model debugging all typically requires both a deep understanding of machine learning and of other branches of science, say data visualization, optimization, or sociology, others might consider penning this type of short FAQ or summary article to elucidate details of their specialty to the broader community.

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] 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`.

[5] Finale Doshi-Velez and Been Kim. Towards a Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv:1702.08608*, 2017. URL: `https://arxiv.org/pdf/1702.08608.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] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning. *arXiv preprint arXiv:1806.00069*, 2018. URL: `https://arxiv.org/pdf/1806.00069.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] 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`.

[11] Alon Keinan, Ben Sandbank, Claus C. Hilgetag, Isaac Meilijson, and Eytan Ruppin. Fair Attribution of Functional Contribution in Artificial and Biological Networks. *Neural Computation*, 16(9):1887–1915, 2004. URL: `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.436.6801&rep=rep1&type=pdf`.

[12] Zachary C. Lipton. The Mythos of Model Interpretability. *arXiv preprint arXiv:1606.03490*, 2016. URL: `https://arxiv.org/pdf/1606.03490.pdf`.

[13] 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`.

[14] 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`.

[15] Christoph Molnar. *Interpretable Machine Learning*. christophm.github.io/interpretable-ml-book, 2018. URL: `https://christophm.github.io/interpretable-ml-book/`.

[16] 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`.

[17] Lloyd S. Shapley, Alvin E. Roth, et al. *The Shapley value: Essays in honor of Lloyd S. Shapley*. Cambridge University Press, 1988. URL: `http://www.library.fa.ru/files/Roth2.pdf`.

[18] Erik Strumbelj and Igor Kononenko. An Efficient Explanation of Individual Classifications using Game Theory. *Journal of Machine Learning Research*, 11(Jan):1–18, 2010. URL: `http://www.jmlr.org/papers/volume11/strumbelj10a/strumbelj10a.pdf`.

[19] 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`.