On Unhelpful eXplainable Machine Learning (XML) Misconceptions

Patrick Hall

October 21, 2018

1 Misconception: Post-Hoc XML Methods and Interpretable Models are Mutually Exclusive

You can, and likely should, use post-hoc XML methods on white-box models, for example consider:

- Locally explainable models + global explainability techniques: rule-based classifiers + partial dependence plots
- Globally explainable models + local explainability techniques: decision trees + Shapley explanations

Corollary Misconception: Post-Hoc XML Methods and Fairness Methods are Mutually Exclusive In banks, using partial dependence plots and disparate impact analysis together is common place.

2 Misconception: XML Methods Only Provide Cover For Government and Commercial Entities to Use Black-Box ML for Nefarious Purposes

XML probably does provide cover to certain extent, but XML methods can be used to crack-open those same black-boxes. Take Angwin et al. (2016) as evidence that this type of investigative analysis of black-box models is possible [1]. Such investigations would likely only be improved by newer explanatory, debugging, and fairness tools.

Moreover, many important technological advances present similar challenges, i.e. social media, strong encryption. People are just now starting to understand and debate these issues for XML. I'm not sure this is fine, but I'd dare to say it is typical.

3 Misconception: XML is Just Models of Models

Yes, models of models or surrogate models, can be helpful explanatory tools, but most serious practitioners I know understand surrogate models are usually approximate, or *low-fidelity* explainers. So, most serious practitioners I know are also aware of methods to increase the fidelity of their surrogate models, see for instance Bastani, Kim, and Bastani (2017) or Hu et al. (2018) [2], [3].

Also, one of the most important breakthroughs in XML is Shapley explanations. These are typically model-specific (i.e. not a surrogate model), rigorously defined, high-fidelity global and local variable importance measures [4]. There are many other model-specific XML methods.

For an ongoing list of many interpretable modeling, model debugging, and model-specific and model-agnostic XML techniques, please see:

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

Misconception Corollary: XML is just LIME. LIME, in it's most popular implementation, uses local linear surrogate models. See above. (It also turns out LIME can be a specific instance of Shapley explanations.)

4 Misconception: XML is a Totally New Thing

Somethings in XML are new, some are not. Many data practitioners in regulated industry and elsewhere have been using surrogate models, partial dependence plots, and white-box models for years. For instance:

- Serious references for surrogate models dating back to at least 1996 [5].
- Partial dependence plots for GBM being proposed at least as far back as 1999 [6].
- Breiman discussing using decision trees for serious commercial applications for a number of years in his seminal *Two Cultures* paper [7].
- Before LIME was presented at KDD 2016, Equifax was developing methods for creating reason codes using monotonically constrained neural networks [8], [9].
- ...

5 Misconception: All the Key Terms in XML are Hopelessly Undefined

There is a long way to go toward a science of interpretable machine learning, but two definitions I find helpful and use frequently are:

- **Interpretable**: The ability to explain or to present in understandable terms to a human [10].
- **A Good Explanation**: When you can no longer keep asking why. [11].

6 References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[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 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.