Guidelines for Responsible Use of Explainable Machine Learning

©Patrick Hall*

H_2O.ai

July 30, 2019

*This material is shared under a CC By 4.0 license which allows for editing and redistribution, even for commercial purposes. However, any derivative work should attribute the author and H2O.ai.

Contents

What is explainable machine learning (ML)?

Variously defined along with aliases or similar concepts:

- "Towards a Rigorous Science of Interpretable Machine Learning" (Doshi-Velez and Kim [6])
- "Explaining Explanations" (Gilpin et al. [8])
- "A Survey Of Methods For Explaining Black Box Models" (Guidotti et al. [10])
- "The Mythos of Model Interpretability" (Lipton [11])
- *Interpretable Machine Learning* (Molnar [13])
- "Interpretable Machine Learning: Definitions, Methods, and Applications" (Murdoch et al. [14])
- "Challenges for Transparency" (Weller [21]).

What is explainable ML?

What do ***I*** mean by explainable ML?

Mostly post-hoc techniques used to enhance ***understanding*** of trained model mechansims and predictions, e.g. ...

- **Direct measures of global and local feature importance**:
  - Gradient-based feature attribution (Ancona et al. [2])
  - Shapley values (Lundberg and Lee [12])
- **Global and local surrogate models**:
  - Decision tree variants (Bastani, Pu, and Solar-Lezama [4], Craven and Shavlik [5])
  - Anchors (Ribeiro, Singh, and Guestrin [15])
  - Local interpretable model-agnostic explanations (LIME) (Ribeiro, Singh, and Guestrin [16])
- **Global and local visualizations of trained model predictions**:
  - Accumulated local effect (ALE) (Apley [3])
  - Partial dependence (Friedman, Hastie, and Tibshirani [7])
  - Individual conditional expectation (ICE) (Goldstein et al. [9])

Why explainable ML?

Responsible Use of Explainable ML can enable:

- Human learning from machine learning
- Human appeal of automated decisions
- Regulatory compliance
- White-hat hacking

Misuse and Abuse of Explainable ML can enable:

- Model and data stealing (Tramèr et al. [20], Shokri et al. [19], Shokri, Strobel, and Zick [18])
- False justification for black-boxes, e.g. "fairwashing" (Aïvodji et al. [1], Rudin [17])

Proposed Guidelines for Responsible Use

Explainable ML is already in-use: numerous open source[†] and commercial packages[‡] available today.

Best-practices are needed to prevent misuse and abuse. So, four basic guidelines are proposed here:

- Use explainable ML to enhance understanding.
- Learn how explainable ML is used for nefarious purposes.
- Augment surrogate models with direct explanations.
- Use highly transparent mechanisms for high-stakes applications.

[†] See: https://github.com/jphall663/awesome-machine-learning-interpretability

[‡] For instance Datarobot, H2O Driverless AI, SAS Visual Data Mining and Machine Learning, Zest AutoML

References

This presentation:
https://www.github.com/jphall663kdd_2019

Code examples for this presentation:
https://www.github.com/jphall663/interpretable_machine_learning_with_python
https://www.github.com/jphall663/responsible_xai

Associated texts:
https://arxiv.org/pdf/1810.02909.pdf
https://arxiv.org/pdf/1906.03533.pdf

References

[1] Ulrich Aïvodji et al. "Fairwashing: the Risk of Rationalization." In: *arXiv preprint arXiv:1901.09749* (2019). URL: https://arxiv.org/pdf/1901.09749.pdf.

[2] Marco Ancona et al. "Towards Better Understanding of Gradient-based Attribution Methods for Deep Neural Networks." In: *6th International Conference on Learning Representations (ICLR 2018)*. URL: https://www.research-collection.ethz.ch/bitstream/handle/20.500.11850/249929/Flow_ICLR_2018.pdf. 2018.

[3] Daniel W. Apley. "Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models." In: *arXiv preprint arXiv:1612.08468* (2016). URL: https://arxiv.org/pdf/1612.08468.pdf.

[4] Osbert Bastani, Yewen Pu, and Armando Solar-Lezama. "Verifiable Reinforcement Learning Via Policy Extraction." In: *Advances in Neural Information Processing Systems*. URL: http://papers.nips.cc/paper/7516-verifiable-reinforcement-learning-via-policy-extraction.pdf. 2018, pp. 2494–2504.

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

References

[6] Finale Doshi-Velez and Been Kim. "Towards a Rigorous Science of Interpretable Machine Learning." In: *arXiv preprint arXiv:1702.08608* (2017). URL: https://arxiv.org/pdf/1702.08608.pdf.

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

[8] Leilani H. Gilpin et al. "Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning." In: *arXiv preprint arXiv:1806.00069* (2018). URL: https://arxiv.org/pdf/1806.00069.pdf.

[9] Alex Goldstein et al. "Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation." In: *Journal of Computational and Graphical Statistics* 24.1 (2015). URL: https://arxiv.org/pdf/1309.6392.pdf.

[10] Riccardo Guidotti et al. "A Survey of Methods for Explaining Black Box Models." In: *ACM Computing Surveys (CSUR)* 51.5 (2018). URL: https://arxiv.org/pdf/1802.01933.pdf, p. 93.

[11] Zachary C. Lipton. "The Mythos of Model Interpretability." In: *arXiv preprint arXiv:1606.03490* (2016). URL: https://arxiv.org/pdf/1606.03490.pdf.

References

[12] Scott M. Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf. Curran Associates, Inc., 2017, pp. 4765–4774.

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

[14] W. James Murdoch et al. "Interpretable Machine Learning: Definitions, Methods, and Applications." In: *arXiv preprint arXiv:1901.04592* (2019). URL: https://arxiv.org/pdf/1901.04592.pdf.

[15] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Anchors: High-Precision Model-agnostic Explanations." In: *AAAI Conference on Artificial Intelligence*. URL: https://homes.cs.washington.edu/~marcotcr/aaai18.pdf. 2018.

[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*. URL: http://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf. ACM. 2016, pp. 1135–1144.

References

[17] Cynthia Rudin. "Please Stop Explaining Black Box Models for High Stakes Decisions." In: *arXiv preprint arXiv:1811.10154* (2018). URL: https://arxiv.org/pdf/1811.10154.pdf.

[18] Reza Shokri, Martin Strobel, and Yair Zick. "Privacy Risks of Explaining Machine Learning Models." In: *arXiv preprint arXiv:1907.00164* (2019). URL: https://arxiv.org/pdf/1907.00164.pdf.

[19] Reza Shokri et al. "Membership Inference Attacks Against Machine Learning Models." In: *2017 IEEE Symposium on Security and Privacy (SP)*. URL: https://arxiv.org/pdf/1610.05820.pdf. IEEE. 2017, pp. 3–18.

[20] Florian Tramèr et al. "Stealing Machine Learning Models via Prediction APIs." In: *25th {USENIX} Security Symposium ({USENIX} Security 16)*. URL: https://www.usenix.org/system/files/conference/usenixsecurity16/sec16_paper_tramer.pdf. 2016, pp. 601–618.

[21] Adrian Weller. "Challenges for Transparency." In: *arXiv preprint arXiv:1708.01870* (2017). URL: https://arxiv.org/pdf/1708.01870.pdf.