

Hands-On Tutorial: Guidelines for Responsible Use of Explainable Machine Learning

Patrick Hall
patrick.hall@h2o.ai
H2O.ai

George Washington University

Navdeep Gill
navdeep.gill@h2o.ai
H2O.ai

Nicholas Schmidt
nschmidt@bldsllc.com
BLDS, LLC

INTRODUCTION & AGENDA

Explainable machine learning (ML) enables human learning from ML, human appeal of automated model decisions, regulatory compliance, and white-hat hacking and security audits of ML models.¹

² Explainable ML (i.e. explainable artificial intelligence or XAI) has been implemented in numerous open source and commercial packages and explainable ML is also an important, mandatory, or embedded aspect of commercial predictive modeling in industries like financial services.^{3,4,5} However, like many technologies, explainable ML can be misused and abused, particularly as a faulty safeguard for harmful black-boxes, e.g. *fairwashing*, and for other malevolent purposes like stealing models or sensitive training data [1], [34], [37], [39]. In an effort to begin establishing best practices for this already in-flight technology and also to promote nuanced discussion, this tutorial presents:

- Definitions and examples related to explainable ML. (**Section 1; 20 mins.**)
- Responsible use guidelines and corollaries:
 - Use explanations to enable understanding directly (and trust as a side-effect). (**Section 2.1; 40 mins.**)
 - Learn how explainable ML is used for nefarious purposes. (**Section 2.2; 30 mins.**)
 - Augment surrogate models with direct explanations. (**Section 2.3; 30 mins.**)
 - Use fully transparent ML mechanisms for high-stakes applications. (**Section 2.4; 50 mins.**)
- Conclusion: a holistic approach to ML (**Section 3; 10 mins.**)

Total time: 180 mins.

¹In the U.S., interpretable models, explainable ML, and model documentation they enable may be required under the Civil Rights Acts of 1964 and 1991, the Americans with Disabilities Act, the Genetic Information Nondiscrimination Act, the Health Insurance Portability and Accountability Act, the Equal Credit Opportunity Act (ECOA), the Fair Credit Reporting Act (FCRA), the Fair Housing Act, Federal Reserve SR 11-7, and the European Union (EU) Greater Data Privacy Regulation (GDPR) Article 22 [43].

²For security applications, see for instance: <https://www.oreilly.com/ideas/proposals-for-model-vulnerability-and-security>.

³Like H2O-3, XGBoost, and various other Python and R packages. See: <https://github.com/jphall663/awesome-machine-learning-interpretability> for a longer, curated list of relevant open source software packages.

⁴For instance Datarobot, H2O Driverless AI, SAS Visual Data Mining and Machine Learning, Zest AutoML, and likely several others.

⁵For instance, “Deep Insights into Explainability and Interpretability of Machine Learning Algorithms and Applications to Risk Management,” Jie Chen, Wells Fargo Corporate Model Risk, <https://www2.amstat.org/meetings/jsm/2019/onlineprogram/AbstractDetails.cfm?abstractid=303053>.

1 DEFINITIONS & EXAMPLES

While explainable ML practitioners have seemingly not yet adopted a clear taxonomy of concepts or a precise vocabulary, many authors have grappled with a variety of general concepts related to interpretability and explanations, e.g. Guidotti et al. [18], Lipton [27], Molnar [30], Murdoch et al. [32]), and Weller [42]. To avoid ambiguity, this section addresses the terms and phrases *interpretable*, *explanation*, *explainable ML*, *interpretable models*, *model debugging techniques*, and *fairness techniques*.

1.1 Interpretable & Explanation

This text selects for an internal definition of *interpretable*, “the ability to explain or to present in understandable terms to a human,” from Doshi-Velez and Kim [9]. In the context of machine learning, Sameer Singh, co-inventor of the popular local interpretable model-agnostic explanations (LIME) technique, defines an *explanation* as, “a collection of visual and/or interactive artifacts that provide a user with sufficient description of the model behavior to accurately perform tasks like evaluation, trusting, predicting, or improving the model.”⁶ From Gilpin et al. [15], “when you can no longer keep asking why,” will serve as an internal working definition of a *good explanation*. These three thoughtful characterizations link explanation to some ML process being interpretable and also provide an abstract objective for any given explanatory task.

1.2 Explainable ML & Interpretable Models

Herein *explainable ML* means mostly post-hoc analysis and techniques used to understand trained model mechanisms or predictions. Examples of common explainable ML techniques include:

- Local and global feature importance methods, e.g. Shapley values and derivative-based feature attribution [3] [24], [29], [35], [38].
- Local and global model-agnostic surrogate models, e.g. surrogate decision trees and LIME [6], [7], [8], [21], [33], [41].
- Local and global visualizations of model predictions, e.g. accumulated local effect (ALE) plots, 1- and 2-dimensional partial dependence plots, and individual conditional expectation (ICE) plots [5], [14], [16].

Although difficult to quantify, credible research efforts into measures of interpretability are underway [13], [31]. The ability to measure degrees of interpretability implies it’s not a binary, on-off quantity.⁷ Here unconstrained, traditional black-box ML models, such as multilayer perceptron (MLP) neural networks and gradient

⁶“Proposed Guidelines for Responsible Use of Explainable Machine Learning,” Patrick Hall, H2O.ai, https://github.com/jphall663/kdd_2019.

⁷See: Figure 3 in “Quantifying Interpretability of Arbitrary Machine Learning Models Through Functional Decomposition” (Molnar et al. [31]).

boosting machines (GBMs), are said to be directly uninterpretable, potentially unsafe for use on humans, but not necessarily completely unexplainable. In this tutorial *interpretable models* or “white-box” models will include linear models, decision trees, rule-based models, constrained or Bayesian variants of traditional black-box ML models, or novel types of models designed to be directly interpretable. Examples of sophisticated and interpretable-by-design modeling techniques include explainable neural networks (XNNs), explainable boosting machines (EBMs, a.k.a. GA2M), monotonically constrained GBMs, scalable Bayesian rule lists, or super-sparse linear integer models (SLIMs), [40], [41], [44].^{8,9,10}

1.3 Model Debugging Techniques

Herein *model debugging techniques* refer to methods for testing ML models that increase trust in model mechanisms and predictions. Examples of model debugging techniques include model assertions, ML security audits, variants of sensitivity (i.e. “what-if?”) analysis, variants of residual analysis, and units test used to verify the accuracy or security of ML models [2], [23].¹¹ Model debugging should also include remediating any discovered errors or vulnerabilities.

1.4 Fairness Techniques

For this tutorial, *fairness techniques* are used to diagnose and remediate unwanted sociological bias in ML models. Diagnosis approaches include disparate impact testing and other tests for unwanted sociological bias [11]. Remediation methods tend to involve model selection by minimization of bias, preprocessing training data, e.g. reweighing (Kamiran and Calders [22]), training unbiased models, e.g. adversarial de-biasing (Zhang et al. [45]), or post-processing model predictions, e.g. by equalizing odds (Hardt et al. [20]).¹² i have like 300

2 GUIDELINES

Four guidelines are presented and discussed in Sections 2.1 – 2.4 to assist practitioners in avoiding any unintentional misuse or in identifying any intentional abuse of explainable ML techniques. Important corollaries to the guidelines are also highlighted and simple, reproducible software examples accompany the guidelines to avoid hypothetical reasoning whenever possible.

2.1 Guideline: Use Explanations to Enable Understanding Directly.

While they are likely necessary for trust in many cases, explanations are certainly not sufficient for trust in all cases. Explanation, as a general idea, is related more directly to understanding and transparency than to trust.¹³ Explanations promote understanding directly, and only increase trust if the explanation-facilitated understanding of a model is acceptable to human consumers. Simply put,

one can understand and explain a model without trusting it. One can also trust a model and not be able to understand or explain it.

- **Explanation & understanding without trust:** In Figure 1, global Shapley explanations and residual analysis identify a pathology in an unconstrained GBM model, g_{GBM} , trained on the UCI credit card dataset [25]. g_{GBM} over-emphasizes the input feature PAY_0, or a customer’s most recent repayment status. Due to over-emphasis of PAY_0, g_{GBM} is often unable to predict on-time payment if recent payments are delayed ($\text{PAY}_0 > 1$), causing large negative residuals. g_{GBM} is also often unable to predict default if recent payments are made on-time ($\text{PAY}_0 \leq 1$), causing large positive residuals. In this example scenario, g_{GBM} is explainable, but not trustworthy.
- **Trust without explanation & understanding:** Years before reliable explanation methods were widely acknowledged and available, black-box models, such as autoencoder and MLP neural networks, were used for fraud detection in the financial services industry [17]. When these models performed well, they were trusted.¹⁴ However, they were not explainable or well-understood by contemporary standards.

If trust in models is your goal, then explanations alone are not sufficient. However, as discussed in Sections 2.4 and 3 and illustrated in Figure 4, in an ideal scenario, explanation techniques would be used with a wide variety of other methods to increase accuracy, fairness, interpretability, privacy, security, and trust in ML models.

2.2 Guideline: Learn How Explainable ML is Used for Nefarious Purposes.

When used disingenuously, explainable ML methods can provide cover for misused or intentionally abusive black-boxes [1], [34]. Explainable ML methods can also enable hacking or stealing of data and models through public prediction APIs or other endpoints [37], [39]. Moreover, explainable ML methods are likely to be used for other nefarious purposes in the future and may be used for unknown destructive purposes now.

2.2.1 Corollary: Explainable ML Can be Used to Crack Nefarious Black-boxes. Used as white-hat hacking tools, explainable ML methods can draw attention to potential fairness or accuracy problems in proprietary black-boxes. See Angwin et al. [4] for evidence that cracking of commercial black-box models for oversight purposes is possible.¹⁵ Such investigations would likely only be improved by advances in explanatory and fairness tools.

2.2.2 Corollary: Explainable ML is a Privacy Vulnerability. Recent research has shown that providing explanations along with model predictions eases attacks that can compromise sensitive training data [36]. When issuing predictions with explanations, reveal only as much information as needed to appeal erroneous model decisions and protect prediction APIs and other endpoints with access control and security measures.

⁸As implemented in the interpret library: <https://github.com/microsoft/interpret>.

⁹As implemented in XGBoost (<https://xgboost.readthedocs.io/en/latest/tutorials/monotonic.html>) or H2O-3 (https://github.com/h2oai/h2o-3/blob/master/h2o-py/demos/H2O_tutorial_gbm_monotonicity.ipynb).

¹⁰And other methods, e.g. <https://users.cs.duke.edu/~cynthia/papers.html>.

¹¹And similar methods, e.g. <https://debug-ml-iclr2019.github.io/>.

¹²And similar methods, e.g. <http://www.fatml.org/resources/relevant-scholarship>.

¹³The Merriam-Webster definition of *explain*, accessed May 8th 2019, does not mention *trust*: <https://www.merriam-webster.com/dictionary/explain>.

¹⁴See: https://www.sas.com/en_ph/customers/hsbc.html, <https://www.kdnuggets.com/2011/03/sas-patent-fraud-detection.html>.

¹⁵This text makes no claim on the quality of the analysis in Angwin et al. (2016), which has been criticized [12]. This now infamous analysis is presented only as evidence that motivated activists can crack commercial black-boxes using surrogate models and other explanatory techniques.

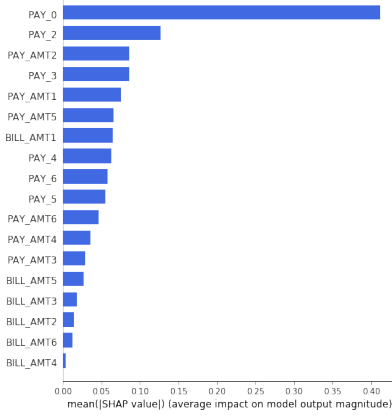
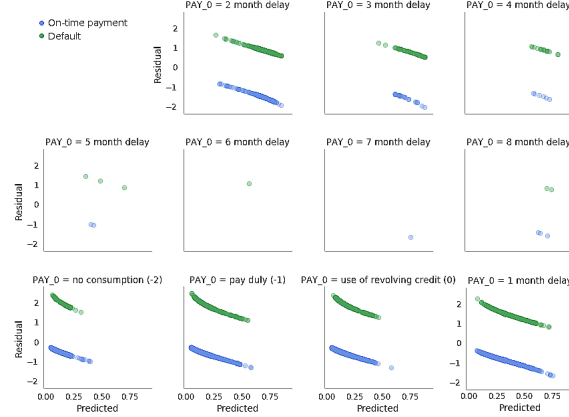
(a) Global Shapley feature importance for g_{GBM} .(b) g_{GBM} deviance residuals and predictions by PAY_0 .

Figure 1: An unconstrained GBM probability of default model, g_{GBM} , generally over-emphasizes the importance of the input feature PAY_0 , a customer’s most recent repayment status. g_{GBM} produces large positive residuals when PAY_0 indicates on-time payments ($PAY_0 \leq 1$) and large negative residuals when PAY_0 indicates late payments ($PAY_0 > 1$). Combining explanatory and debugging techniques shows that g_{GBM} is explainable, but probably not trustworthy. Code to replicate Figure 1 is available here: https://github.com/h2oai/xai_guidelines.

2.3 Augment Surrogate Models with Direct Explanations.

Models of models, or surrogate models, can be helpful explanatory tools, but they are usually approximate, low-fidelity explainers. Aside from 1.) a global or local summary of a complex model provided by a surrogate model can be helpful sometimes and 2.) much work in explainable ML has been directed toward improving the fidelity and usefulness of surrogate models [6], [7], [8], [21], [41], *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 attributions which are consistent globally and accurate locally using the trained model itself [29], [38]. An extension of this idea, called tree SHAP, has already been implemented for popular tree ensemble methods [28].

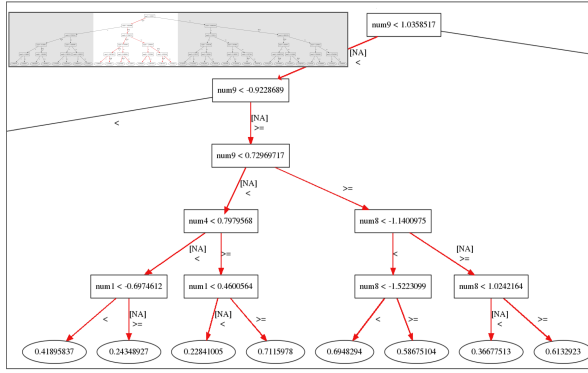
Many other explainable ML methods operate on trained models directly such as partial dependence, ALE, and ICE plots [5], [14], [16]. Surrogate models and explanatory techniques that operate directly on trained models can also be combined, for instance by using partial dependence, ICE, and surrogate decision trees to investigate and confirm modeled interactions [19]. In Figure 2, an unconstrained g_{GBM} models a known signal generating function f :

$$f(\mathbf{X}) = \begin{cases} 1 & \text{if } X_{num1} * X_{num4} + |X_{num8}| * X_{num9}^2 + e \geq 0.42 \\ 0 & \text{if } X_{num1} * X_{num4} + |X_{num8}| * X_{num9}^2 + e < 0.42 \end{cases} \quad (1)$$

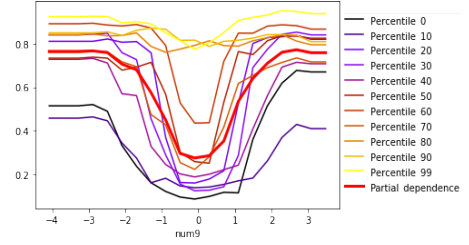
where e signifies random noise in the form of label switching for roughly 15% of the training and validation observations. g_{GBM} is then trained such that $g_{GBM}(\mathbf{X}) \approx f(\mathbf{X})$ in training and validation data. h_{tree} , displayed in Figure 2a, is extracted such that $h_{tree}(\mathbf{X}) \approx g_{GBM}(\mathbf{X}) \approx f(\mathbf{X})$ in validation data. Partial dependence and ICE plots are generated directly for g_{GBM} in the same validation data and

overlaid in Figure 2b. The parent-child node relationship displayed between X_{num9} and X_{num8} for $\sim -0.923 < X_{num9} < \sim 1.04$ in 2a and the divergence of ICE and partial dependence curves in 2b for $\sim -1 < X_{num9} < \sim 1$ help confirm and understand how g_{GBM} learned the interaction between X_{num8} and X_{num9} in f . As in Figure 1, combining different approaches provided additional, beneficial information about a complex ML model.

2.3.1 Corollary: Augment LIME with Direct Explanations. LIME is important, imperfect (like every other ML technique), and one of many explainable ML tools. LIME, in its most popular implementation, uses local linear surrogate models to explain regions of complex, machine-learned response functions [33]. Like other surrogate models, LIME can 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 as in 3b. LIME can then provide approximate information about modeled local linear trends around the same point. Table 1 contains LIME h_{GLM} coefficients for a local region of a validation set sampled from the UCI credit card data defined by $PAY_0 > 1$, or customers with a fairly high risk of default due to late most recent payments. h_{GLM} models the predictions of a simple interpretable decision tree model, g_{tree} , displayed in 3a. h_{GLM} coefficients show linear trends between features in the sampled set $X_{PAY_0 > 1}$ and $g_{tree}(X_{PAY_0 > 1})$. Because h_{GLM} is relatively well-fit ($0.73 R^2$) and has a logical intercept (0.77), it can be used along with Shapley values to reason about the modeled average behavior for risky customers and to differentiate the behavior of any one specific risky customer from their peers under the model. This additional information is often useful for debugging and compliance purposes.



(a) Naïve h_{tree} , a surrogate model, forms an approximate overall flowchart for the explained model, g_{GBM} .



(b) Partial dependence and ICE curves generated directly from the explained model, g_{GBM} .

Figure 2: h_{tree} displays known interactions in $f = X_{num1} * X_{num4} + |X_{num8}| * X_{num9}^2$ for $-0.923 < X_{num9} < 1.04$. Modeling of the known interaction between X_{num9} and X_{num8} in f by g_{GBM} is confirmed by the divergence of partial dependence and ICE curves for $-1 < X_{num9} < 1$. Explanations from a surrogate model have augmented and confirmed findings from a direct model visualization technique. Code to replicate Figure 2 is available here: https://github.com/h2oai/xai_guidelines.

Table 1: Coefficients for a local linear interpretable model, h_{GLM} , with an intercept of 0.77 and an R^2 of 0.73. h_{GLM} is trained on a segment of the UCI credit card dataset containing higher-risk customers with late most recent repayment statuses, $X_{PAY_0} > 1$, and the predictions of a simple decision tree, $g_{tree}(X_{PAY_0} > 1)$. Code to replicate Table 1 is available here: https://github.com/h2oai/xai_guidelines.

h_{GLM} Feature	h_{GLM} Coefficient
PAY_0 == 4	0.0009
PAY_2 == 3	0.0065
PAY_5 == 2	-0.0006
PAY_6 == 2	0.0036
BILL_AMT1	3.4339e-08
PAY_AMT1	4.8062e-07
PAY_AMT3	-5.867e-07

2.4 Use Fully Transparent ML Mechanisms for High-Stakes Applications.

Explanation, along with white-box models, model debugging, disparate impact analysis, and the documentation they enable, are often required under numerous regulatory statutes in the U.S. and E.U., and explainable ML tools are already used to document, understand, and validate different types of models in the financial services industry [21], [41].^{2, 5} Moreover, adverse action notices are mandated under ECOA and FCRA for many credit lending, employment, and insurance decisions in the United States.¹⁶ When ML is used for such decisions today it must be explained in terms of adverse action notices.¹⁷ Shapley values, and other local feature importance approaches, provide a convenient methodology to rank the contribution of input features to model decisions and potentially generate customer-specific adverse action notices.

¹⁶ See: <https://consumercomplianceoutlook.org/2013/second-quarter/adverse-action-notice-requirements-under-ecoa-fcra/>.

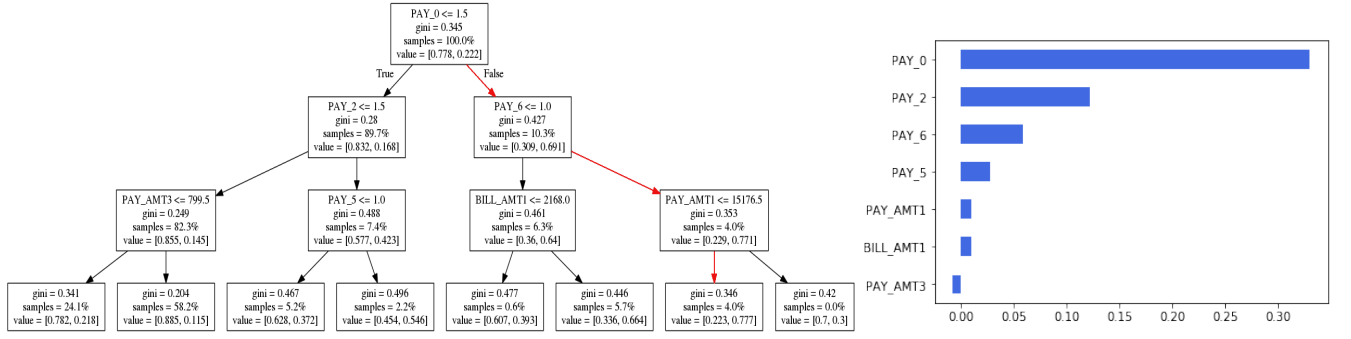
¹⁷ This is already happening: <https://www.prnewswire.com/news-releases/new-patent-pending-technology-from-eqifax-enables-configurable-ai-models-300701153.html>.

Aside from regulatory mandates, explanation enables logical appeal processes for automated decisions made by ML models. Consider being negatively impacted by an erroneous black-box model decision, say for instance being mistakenly denied a loan or parole. How would you argue your case for appeal without knowing how model decisions were made? According to the New York Times, a man named Glenn Rodriguez found himself in this unfortunate position in a penitentiary in Upstate New York in 2016.¹⁸

Some may argue that, outside of regulated dealings, for a model with little or no impact on humans and that has been thoroughly and responsibly tested by knowledgeable practitioners, that explanation is *really* unnecessary. While that statement appears technically true, a counter argument centers on human learning from ML models. Explanatory techniques allow users to gain insights from complex models about nonlinear or faint phenomena and complex interactions – information that may sometimes be unlearnable by linear models. Why go through weeks, months, or years of training and deploying a production ML system, and not take a small amount of that time to learn about the model’s findings?

2.4.1 Corollary: Use Interpretable Models Along with Explanation Techniques. A few well-known publications have focused either on white-box modeling techniques (e.g. Ustun and Rudin [40], Yang et al. [44]) or on post-hoc explanations (e.g. Lundberg and Lee [29], Ribeiro et al. [33]), but the two can be used together in the context of a holistic ML workflow as illustrated in Figure 4. Consider the seemingly useful example case of augmenting globally interpretable models with local post-hoc explanations. A practitioner could train a single decision tree, a globally interpretable model, then apply local explanations in the form of Shapley feature importance as illustrated in Figure 3. This enables the practitioner to see accurate numeric feature contributions for each model prediction in addition to the entire directed graph of the tree. Even for interpretable models, such as linear models and decision trees, it has been shown

¹⁸ This too is happening today: <https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html>.



(a) Simple decision tree, g_{tree} , trained on the UCI credit card data to predict default with validation AUC of 0.74. The decision policy for a high-risk individual is highlighted in red.

(b) Locally-accurate Shapley contributions for the highlighted individual's probability of default.

Figure 3: A simple decision tree, g_{tree} , is trained on the UCI credit card dataset to predict probability of default. g_{tree} has a validation AUC of 0.74. The decision-policy for a high-risk customer is highlighted in 3a and the locally-accurate Shapley contributions for this same individual's predicted probability are displayed in 3b. The Shapley values are helpful because they highlight the local importance of features not on the decision path in this particular encoding of the unknown signal-generating function, i.e. g_{tree} , which could be underestimated by examining the decision policy alone. Code to replicate Figure 3 is available here: https://github.com/h2oai/xai_guidelines.

that Shapley values present accuracy and consistency advantages over standard feature attribution methods [26], [28], [29].

2.4.2 Corollary: Use Explanations Along with Bias Testing and Remediation. Like white-box models, fairness methods (e.g. Feldman et al. [11], Hardt et al. [20]) are often presented in different articles than post-hoc explanatory methods. However, in banks, using post-hoc explanatory tools along with disparate impact analysis is necessary to comply with model documentation guidance and with fair lending regulations.^{5,19,20} To clarify, explanatory techniques should *not* replace fairness methods, but in general, explanations increase transparency and understanding of model mechanisms and predictions while bias auditing and remediation increases trust that model predictions are as fair as possible. As in previous sections, trust and understanding are different but complimentary goals achieved by combining multiple approaches.

Table 2: Basic group disparity metrics across marital statuses for monotonically constrained GBM model, g_{mono} , trained on the UCI credit card dataset. Code to replicate Table 2 is available here: https://github.com/h2oai/xai_guidelines.

	Adverse Impact Disparity	TPR Disparity	FPR Disparity	FNR Disparity
married	1.00	1.00	1.00	1.00
single	0.89	0.99	0.85	1.01
divorced	1.01	0.81	1.25	1.22
other	0.26	0.62	0	1.44

Table 2 displays basic group disparity metrics for a monotonically constrained GBM model, g_{mono} , trained on the UCI credit card data. In this example scenario, g_{mono} displays group parity according to the four-fifths rule with married as the reference level

for single customers, but exposes potential disparate impact for divorced customers and customers with marital status of other (for which there is very little training data). Alternative models with less disparate impact or other remediation processes should be considered in such cases to increase trust ML systems.

2.4.3 Corollary: Explanation is Not a Frontline Fairness Tool. In many high-stakes and commercially viable applications of explainable ML in credit lending, insurance, and employment in the U.S. that fall under FCRA, ECOA, or other applicable regulations, demographic attributes cannot be used in predictive models and thus their contribution to model predictions cannot be explained using common explainable ML techniques. Even when demographic attributes can be used in predictive models, it has been shown that explanations may not detect unwanted social bias [1]. Given these known drawbacks, it is recommended that practitioners use fairness techniques to test for and remediate unwanted sociological bias, and use explanations to augment and understand bias when appropriate.

2.4.4 Corollary: Use Bias Testing Along with Constrained Models. Because unconstrained ML models have the ability to treat similar individuals differently based on small differences in their input data values, unconstrained models can cause local bias that is not detectable with standard bias testing methods that analyze group fairness [10]. To minimize local unwanted sociological bias when using machine learning, and to ensure standard bias testing methods are most effective, pair bias testing techniques with constrained models.

3 CONCLUSION

ML systems are used today to make life-altering decisions about employment, bail, parole, and lending.²¹ The scope of decisions

¹⁹See: <https://www.aba.com/Compliance/Documents/FairLendingWhitePaper2017Apr.pdf>.

²⁰See: <https://www.govinfo.gov/content/pkg/FR-1994-04-15/html/94-9214.htm>.

²¹ICLR 2019 model debugging workshop CFP: <https://debug-ml-iclr2019.github.io/>.

delegated to ML systems seems likely only to expand in the future. Many researchers and practitioners are tackling fairness, inaccuracy, privacy violations, and security vulnerabilities with a number of brilliant, but sometimes siloed, approaches. By presenting some explainable ML guidelines, this tutorial also gives examples of combining innovations from several sub-disciplines of ML research to train explainable, fair, and trustable predictive modeling systems. As proposed in Figure 4, using these techniques together can create a new and more holistic approach to ML potentially better-suited for use in business- and life-critical decision support than conventional methods.

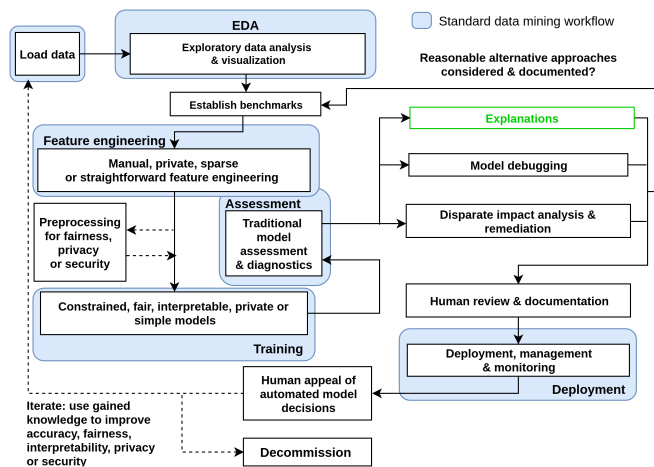


Figure 4: A diagram of a proposed human-centered ML workflow in which explanations (highlighted in green) are used along with interpretable or white-box models, disparate impact analysis and remediation techniques, and other review and appeal mechanisms to create a fair, accountable, and transparent ML system.

SOFTWARE RESOURCES

This tutorial uses Jupyter notebooks and Python code stored in a public GitHub repository with an Apache 2.0 license https://github.com/h2oai/xai_guidelines. Notebooks will be deployed in H2O.ai's free educational cloud environment, Aquarium: <http://aquarium.h2o.ai>. Attendees will only need an email address (to receive a password after Aquarium registration) and to bring their laptop to access and execute the materials.

ACKNOWLEDGEMENTS

The author thanks Przemyslaw Biecek, Pramit Choudhary, Christoph Molnar, and numerous members of the online data science community for their helpful input and insights.

REFERENCES

- [1] Ulrich Aivodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. 2019. Fairwashing: the Risk of Rationalization. *arXiv preprint arXiv:1901.09749* (2019). URL: <https://arxiv.org/pdf/1901.09749.pdf>.
- [2] Saleema Amershi, Max Chickering, Steven M Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. 2015. Modeltracker: Redesigning Performance Analysis Tools for Machine Learning. In *Proceedings of the 33rd Annual*

- ACM Conference on Human Factors in Computing Systems*. ACM, 337–346. URL: <https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/amershi.CHI2015.ModelTracker.pdf>.
- [3] Marco Ancona, Enea Ceolini, Cengiz Oztireli, and Markus Gross. 2018. 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.
- [4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. 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>.
- [5] Daniel W. Apley. 2016. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *arXiv preprint arXiv:1612.08468* (2016). URL: <https://arxiv.org/pdf/1612.08468.pdf>.
- [6] Osbert Bastani, Carolyn Kim, and Hamsa Bastani. 2017. Interpreting Blackbox Models via Model Extraction. *arXiv preprint arXiv:1705.08504* (2017). URL: <https://arxiv.org/pdf/1705.08504.pdf>.
- [7] Osbert Bastani, Yewen Pu, and Armando Solar-Lezama. 2018. Verifiable Reinforcement Learning Via Policy Extraction. In *Advances in Neural Information Processing Systems*. 2494–2504. URL: <http://papers.nips.cc/paper/7516-verifiable-reinforcement-learning-via-policy-extraction.pdf>.
- [8] Mark W. Craven and Jude W. Shavlik. 1996. 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>.
- [9] Finale Doshi-Velez and Been Kim. 2017. Towards a Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv:1702.08608* (2017). URL: <https://arxiv.org/pdf/1702.08608.pdf>.
- [10] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness Through Awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. ACM, 214–226. URL: <https://arxiv.org/pdf/1104.3913.pdf>.
- [11] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 259–268. URL: <https://arxiv.org/pdf/1412.3756.pdf>.
- [12] Anthony W. Flores, Kristin Bechtel, and Christopher T. Lowenkamp. 2016. False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks. *Fed. Probation* 80 (2016), 38. URL: <https://bit.ly/2Gesf9Y>.
- [13] Sorelle A. Friedler, Chitradheep Dutta Roy, Carlos Scheidegger, and Dylan Slack. 2019. Assessing the Local Interpretability of Machine Learning Models. *arXiv preprint arXiv:1902.03501* (2019). URL: <https://arxiv.org/pdf/1902.03501.pdf>.
- [14] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2001. *The Elements of Statistical Learning*. Springer, New York. URL: https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12.pdf.
- [15] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning. *arXiv preprint arXiv:1806.00069* (2018). URL: <https://arxiv.org/pdf/1806.00069.pdf>.
- [16] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. 2015. 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>.
- [17] Krishna M. Gopinathan, Louis S. Bifore, William M. Ferguson, Michael A. Lazarus, Anu K. Pathria, and Allen Jost. 1998. Fraud Detection using Predictive Modeling. US Patent 5,819,226. URL: <https://patents.google.com/patent/US5819226A>.
- [18] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys (CSUR)* 51, 5 (2018), 93. URL: <https://arxiv.org/pdf/1802.01933.pdf>.
- [19] Patrick Hall. 2018. On the Art and Science of Machine Learning Explanations. In *JSM Proceedings, Statistical Computing Section*. American Statistical Association, 1781–1799. URL: https://github.com/jphall663/jsm_2018_paper.
- [20] Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*. 3315–3323. URL: <http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf>.
- [21] Linwei Hu, Jie Chen, Vijayan N. Nair, and Agus Sudjianto. 2018. 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>.
- [22] Faisal Kamiran and Toon Calders. 2012. Data Preprocessing Techniques for Classification Without Discrimination. *Knowledge and Information Systems* 33, 1 (2012),

- 1–33. URL: <https://link.springer.com/content/pdf/10.1007/s10115-011-0463-8.pdf>.
- [23] Daniel Kang, Deepti Raghavan, Peter Bailis, and Matei Zaharia. [n. d.]. Debugging Machine Learning Models via Model Assertions. URL: https://debug-ml-iclr2019.github.io/cameraready/DebugML-19_paper_27.pdf.
- [24] Alon Keinan, Ben Sandbank, Claus C. Hilgetag, Isaac Meilijson, and Eytan Ruppin. 2004. Fair Attribution of Functional Contribution in Artificial and Biological Networks. *Neural Computation* 16, 9 (2004), 1887–1915. URL: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.436.6801&rep=rep1&type=pdf>.
- [25] M. Lichman. 2013. UCI Machine Learning Repository. URL: <http://archive.ics.uci.edu/ml>.
- [26] Stan Lipovetsky and Michael Conklin. 2001. Analysis of Regression in Game Theory Approach. *Applied Stochastic Models in Business and Industry* 17, 4 (2001), 319–330.
- [27] Zachary C. Lipton. 2016. The Mythos of Model Interpretability. *arXiv preprint arXiv:1606.03490* (2016). URL: <https://arxiv.org/pdf/1606.03490.pdf>.
- [28] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. 2017. Consistent Individualized Feature Attribution for Tree Ensembles. In *Proceedings of the 2017 ICML Workshop on Human Interpretability in Machine Learning (WHI 2017)*, Been Kim, Dmitry M. Malioutov, Kush R. Varshney, and Adrian Weller (Eds.). ICML WHI 2017, 15–21. URL: <https://openreview.net/pdf?id=ByTKSo-m->.
- [29] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4765–4774. URL: <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>.
- [30] Christoph Molnar. 2018. *Interpretable Machine Learning*. christophm.github.io. URL: <https://christophm.github.io/interpretable-ml-book/>.
- [31] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. 2019. Quantifying Interpretability of Arbitrary Machine Learning Models Through Functional Decomposition. *arXiv preprint arXiv:1904.03867* (2019). URL: <https://arxiv.org/pdf/1904.03867.pdf>.
- [32] W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Interpretable Machine Learning: Definitions, Methods, and Applications. *arXiv preprint arXiv:1901.04592* (2019). URL: <https://arxiv.org/pdf/1901.04592.pdf>.
- [33] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. 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*. ACM, 1135–1144. URL: <http://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf>.
- [34] Cynthia Rudin. 2018. Please Stop Explaining Black Box Models for High Stakes Decisions. *arXiv preprint arXiv:1811.10154* (2018). URL: <https://arxiv.org/pdf/1811.10154.pdf>.
- [35] Lloyd S. Shapley, Alvin E. Roth, et al. 1988. *The Shapley value: Essays in honor of Lloyd S. Shapley*. Cambridge University Press. URL: <http://www.library.fu.ru/files/Roth2.pdf>.
- [36] Reza Shokri, Martin Strobel, and Yair Zick. 2019. Privacy Risks of Explaining Machine Learning Models. *arXiv preprint arXiv:1907.00164* (2019). URL: <https://arxiv.org/pdf/1907.00164.pdf>.
- [37] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership Inference Attacks Against Machine Learning Models. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 3–18. URL: <https://arxiv.org/pdf/1610.05820.pdf>.
- [38] Erik Strumbelj and Igor Kononenko. 2010. An Efficient Explanation of Individual Classifications using Game Theory. *Journal of Machine Learning Research* 11, Jan (2010), 1–18. URL: <http://www.jmlr.org/papers/volume11/strumbelj10a/strumbelj10a.pdf>.
- [39] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. 2016. Stealing Machine Learning Models via Prediction APIs. In *25th {USENIX} Security Symposium ({USENIX} Security 16)*. 601–618. URL: https://www.usenix.org/system/files/conference/usenixsecurity16/sec16_paper_tramer.pdf.
- [40] Berk Ustun and Cynthia Rudin. 2016. Supersparse Linear Integer Models for Optimized Medical Scoring Systems. *Machine Learning* 102, 3 (2016), 349–391. URL: <https://users.cs.duke.edu/~cynthia/docs/UstunTrRuAAAI13.pdf>.
- [41] Joel Vaughan, Agus Sudjianto, Erind Brahimi, Jie Chen, and Vijayan N. Nair. 2018. Explainable Neural Networks Based on Additive Index Models. *arXiv preprint arXiv:1806.01933* (2018). URL: <https://arxiv.org/pdf/1806.01933.pdf>.
- [42] Adrian Weller. 2017. Challenges for Transparency. *arXiv preprint arXiv:1708.01870* (2017). URL: <https://arxiv.org/pdf/1708.01870.pdf>.
- [43] Mike Williams et al. 2017. *Interpretability*. Fast Forward Labs. URL: <https://www.cloudera.com/products/fast-forward-labs-research.html>.
- [44] Hongyu Yang, Cynthia Rudin, and Margo Seltzer. 2017. Scalable Bayesian Rule Lists. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*. URL: <https://arxiv.org/pdf/1602.08610.pdf>.
- [45] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 335–340. URL: <https://arxiv.org/pdf/1801.07593.pdf>.