On Explainable Machine Learning Misconceptions & A More Human-Centered Machine Learning*

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

May 23, 2019

Introduction

Due to obvious community and commercial demand, explainable machine learning (ML) methods have already been implemented in popular open source software and in commercial software.[1,2] Yet, as someone who has been involved in the implementation of explainable ML software for the past three years, I find a lot of what I read about the topic confusing and detached from my personal, hands-on experiences. This short opinion paper presents arguments, proposals, and references to address some observed explainable ML misconceptions. Please note that this text is not an attack on any party, and disagreements with any previous credible opinion pieces are mostly minor and technical.[3] This text is instead a call for pragmatism from the trenches. It seeks to promote nuanced debate, responsible use, and continued development of explainable ML, and not dismissal or derision of the discipline. Moreover, this text builds the seemingly natural case for a holistic approach to ML that includes interpretable (i.e. "white-box") models along with explanatory, debugging, and disparate impact analysis techniques.

To avoid ambiguity, several internal definitions and accompanying examples are put forward before discussing misconceptions. Here *explainable ML* means mostly post-hoc techniques used to understand trained model mechanisms or predictions. Examples of common explainable ML techniques include:

- Local and global feature importance methods, in particular Shapley values [20], [25], [30], [31].
- Local and global model-agnostic surrogate models, such as surrogate decision trees and Local Interpretable Model-agnostic Explanations (LIME) [4], [5], [6], [19], [28].
- Local and global visualizations of model predictions such as accumulated local effect (ALE) plots, 1- and 2-dimensional partial dependence plots, and individual conditional expectation (ICE) plots [3], [11], [13].

In this text *model debugging* refers to testing ML models to increase trust in model mechanisms and predictions. Examples of model debugging techniques include variants of sensitivity (i.e. "what-if?"), residual analysis, assertions, and units test used to verify the accuracy or security of ML models.[4] Model debugging should also include remediating any discovered errors or vulnerabilities. Herein *fairness* techniques refer to

†H2O.ai and George Washington University

[1]Like 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.

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

[3]For instance, say "Please Stop Explaining Black Box Models for High Stakes Decisions" [29].

[4]And other similar techniques, say those mentioned here: https://debug-ml-iclr2019.github.io/.

disparate impact analysis, model selection by minimization of disparate impact, and remediation techniques such as disparate impact removal preprocessing or equalized odds post-processing [8], [18].[5] In this text *interpretable* or *white-box* models will include linear models, decision trees, constrained or Bayesian variants of traditional black-box ML models, or novel types of models designed to be directly interpretable. (Additional examples of interpretable modeling techniques include explainable neural networks (XNNs), monotonically constrained gradient boosting machines (GBMs)[6], scalable Bayesian rule lists, or super-sparse linear integer models (SLIMs), [32], [33], [36].[7]) Herein unconstrained, traditional black-box ML models, such as multilayer perceptron (MLP) neural networks and GBMs, are said to be directly uninterpretable, potentially unsafe for use on humans, but not necessarily completely unexplainable.

1 Misconception: Explanations are Necessary and Sufficient to Establish Trust in ML

Explanations are likely necessary for trust in many cases, but certainly not sufficient for trust in all cases. Explanation, as a general concept, is related more directly to understanding and transparency than to trust.[8] 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. Consider the following example scenarios.

- **Explanation and 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 [21].[9] 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 (`PAY_0` > 1.5), causing large negative residuals. g_{GBM} is also often unable to predict default if recent payments are made on-time (`PAY_0` ≤ 1.5), causing large positive residuals. In this example scenario, g_{GBM} is explainable, but not trustworthy.
- **Trust without explanation and understanding**: Years before reliable explanation techniques were widely acknowledged and available, black-box predictive models, such as autoencoder and MLP neural networks, were used for fraud detection in the financial services industry [14]. When these models performed well, they were trusted.[10] 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 Section 6 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 Misconception: Transparency in ML is Unnecessary

Explainable ML tools like surrogate models, partial dependence plots, and global feature importance are already used to document, understand, and validate different types of models in the financial services industry [19], [33].[11] Furthermore, adverse action notices are mandated under the Equal Credit Opportunity Act (ECOA) and the Fair Credit Reporting Act (FCRA) for many credit lending, employment, and insurance decisions in the United States.[12] If machine learning is used for such decisions it must be explained in

[5] And other similar techniques, say those mentioned here: http://www.fatml.org/resources/relevant-scholarship.

[6] As 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).

[7] And other similar techniques, say those mentioned here: https://users.cs.duke.edu/~cynthia/papers.html.

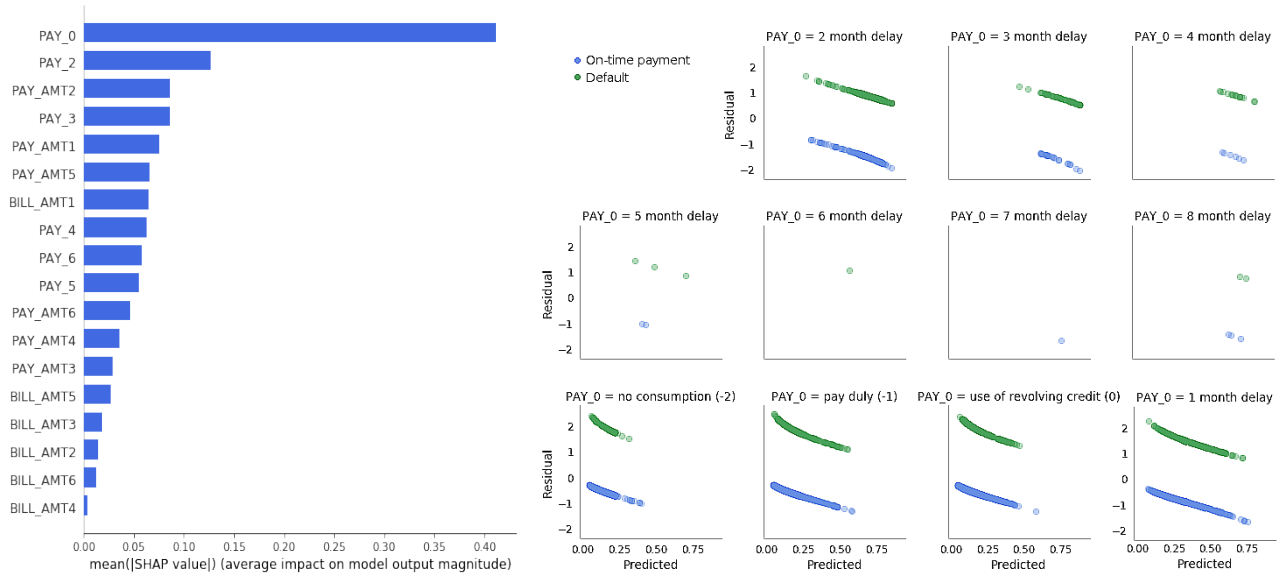
[8] The Merriam-Webster definition of *explain*, accessed May 8th 2019, does not mention *trust*: https://www.merriam-webster.com/dictionary/explain.

[9] Code to replicate Figure 1 is available here: https://github.com/jphall663/xai_misconceptions.

[10] For example: https://www.sas.com/en_ph/customers/hsbc.html, https://www.kdnuggets.com/2011/03/sas-patent-fraud-detection.html.

[11] Working paper: "SR 11-7, Validation and Machine Learning Models". Tony Yang, CFA, CPA, FRM. KPMG USA.

[12] See: https://consumercomplianceoutlook.org/2013/second-quarter/adverse-action-notice-requirements-under-ecoa-fcra/.



(a) Consistent global Shapley feature importance values for g_{GBM}.

(b) g_{GBM} deviance residuals and predictions by `PAY_0`.

Figure 1: An unconstrained GBM probability of default model, g_{GBM}, 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` ≤ 1) and large negative residuals when `PAY_0` indicates late payments (`PAY_0` > 1). g_{GBM} is explainable, but probably not trustworthy.

terms of adverse action notices.[13] Shapley values, and other local feature importance approaches, provide a convenient methodology to rank the direct contribution of input features to final model decisions and potentially generate customer-specific adverse action notices. In a number of other application domains, broader interpretability is also legal necessity. Explanation, along with white-box models, model debugging, disparate impact analysis, and the documentation they enable, can also 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 Fair Housing Act, Federal Reserve SR 11-7, the European Union (EU) Greater Data Privacy Regulation (GDPR) Article 22, and other regulatory statutes [35].

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 the New York Times, a man named Glenn Rodríguez found himself in this unfortunate position in Upstate New York in 2016.[14]

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, the counter argument in this case centers on human learning from ML models. Explanatory techniques allow us to gain insights from complex models about nonlinear phenomena and complex interactions – information that may sometimes be unlearnable by linear models. Why go through the weeks, months, or years of training and deploying a production ML system, and not take a small percentage of that time to learn about the model's findings?

[13] This is apparently *already* happening: https://www.prnewswire.com/news-releases/new-patent-pending-technology-from-equifax-enables-configurable-ai-models-300701153.html.

[14] This too is happening *today*: https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html.

3 Misconception: All the Key Ideas in Explainable ML are Undefined

Helpful definitions that apply to explainable ML have been put forward, including:

- **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) [7].
- **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) [12]. (Gilpin et al. also provide several clear constructs for describing more specific types of explanations.)

While the explainable ML field is far from embracing a clear and accepted taxonomy of concepts or an exhaustive and precise vocabulary, these two thoughtful definitions appear to link explanations to some ML process being interpretable. Moreover, many authors have made significant attempts to grapple with a variety of general concepts related to interpretability and explanations, including "A Survey Of Methods For Explaining Black Box Models" by Guidotti et al. (2018), Zachary Lipton's "The Mythos of Model Interpretability" (2016), Christoph Molnar's *Interpretable Machine Learning* (2018), and Adrian Weller's "Challenges for Transparency" (2017) [16], [23], [26], [34].

Misconception Corollary: Interpretability is Unquantifiable. While interpretability is difficult to quantify, credible research efforts into scientific measures of interpretability are underway [10], [27]. Furthermore, the ability to measure degrees of interpretability implies it's not a binary, on-off quantity.[15]

4 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 1.) a global 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 [4], [5], [6], [19], [33], **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 [25], [31]. An extension of this idea, called tree SHAP, has already been implemented for popular tree ensemble methods [24].

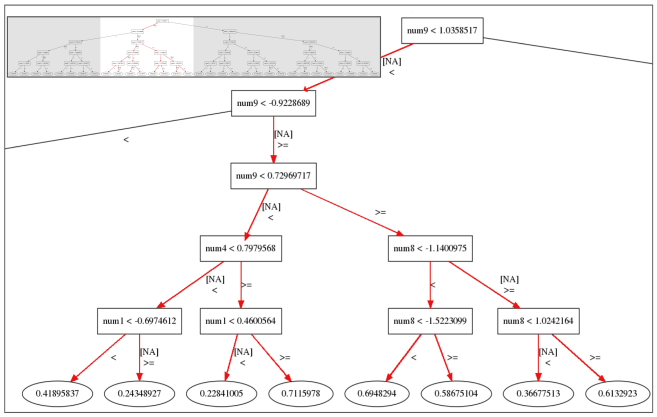
There are many other explainable ML methods that operate on trained models directly such as partial dependence, ALE, and ICE plots [3], [11], [13]. Surrogate models and explanatory techniques that operate directly on trained models can also be combined, for instance by using partial dependence, ICE, and surrogate decision trees to investigate and confirm modeled interactions [17]. In Figure 2, an unconstrained GBM, g_{GBM}, models a known signal generating function f:

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

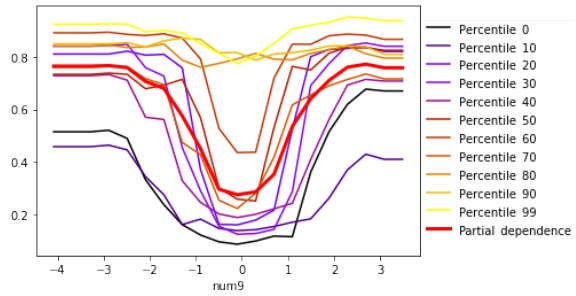
where e signifies the injection of random noise in the form of label switching for roughly 15% of the training and validation observations.[16] g_{GBM} is then trained such that $g_{\text{GBM}}(\mathbf{X}) \approx f(\mathbf{X})$ in training and validation data. h_{tree}, displayed in Figure 2a, is extracted such that $h_{\text{tree}}(\mathbf{X}) \approx g_{\text{GBM}}(\mathbf{X}) \approx f(\mathbf{X})$ in validation data. Partial dependence and ICE plots are generated directly for g_{GBM} in the same validation

[15] For instance, see Figure 3 in "Quantifying Interpretability of Arbitrary Machine Learning Models Through Functional Decomposition" [27].

[16] Code to replicate Figure 2 is available here: https://github.com/h2oai/mli-resources/tree/master/lime_shap_treeint_compare.



(a) Naïve h_{tree} is an approximate overall flowchart for g_{GBM}.



(b) Partial dependence and ICE curves generated directly for g_{GBM}.

Figure 2: h_{tree} displays known interactions between X_{num9} and X_{num8} (along with potential noise interactions) for $\sim -0.923 < X_{\text{num9}} <\sim 1.04$. 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 $\sim -1 < X_{\text{num9}} <\sim 1$.

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

Misconception Corollary: Explainable ML is just LIME. LIME is important, imperfect, and one of many explainable ML tools. LIME, in it's most popular implementation, uses local linear surrogate models to explain regions of complex, machine-learned response functions [28]. 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 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 > 1.5`, or customers with a fairly high risk of default due to late most recent payments.[17] 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 $\mathbf{X}_{\text{PAY\_0>1.5}}$ and $g_{\text{tree}}(\mathbf{X}_{\text{PAY\_0>1.5}})$. 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 can be useful for model validation and compliance purposes.

5 Misconception: Explainable ML Methods Simply Provide Cover to Use Black-Box ML for Nefarious Purposes

If used disingenuously, explainable ML methods probably do provide such cover [1]. 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 hacking or stealing of commercial black-box models for oversight purposes is possible [2].[18] Such investigations would likely only be improved by advances in explanatory and fairness tools. Additionally, many important computer-based technological advances present similar double-edged sword dilemmas, e.g. social media or strong encryption. Rarely does the ability of a tool to be misused for

[17] Code to replicate Table 1 is available here: https://github.com/jphall663/xai_misconceptions.

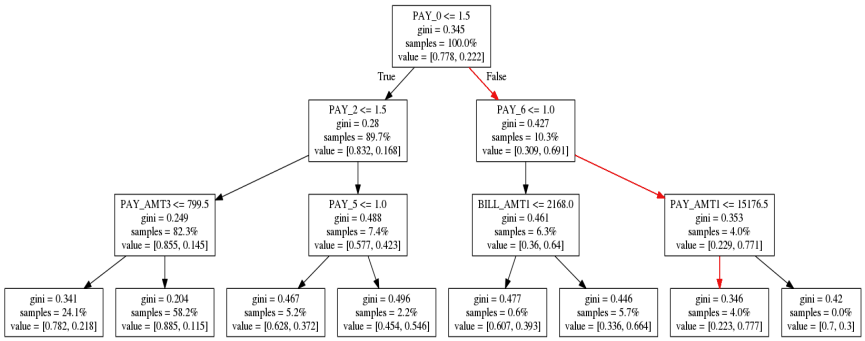
[18] This text makes no claim on the quality of the analysis in Angwin et al. (2016), which has been criticized [9]. This now infamous analysis is presented only as evidence that motivated activists can hack or steal commercial black-boxes.

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, $\mathbf{X}_{PAY\_0>1.5}$, and the predictions of simple decision tree, $g_{\text{tree}}(\mathbf{X}_{\text{PAY\_0>1.5}})$.

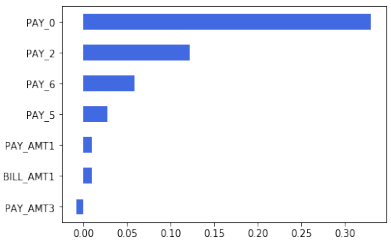
h_{GLM} Feature	h_{GLM} Coefficient
`PAY_0 == 4`	0.0009
`PAY_2 == 3`	0.0065
`PAY_5 == 2`	−0.0006
`PAY_6 == 2`	0.0036
`BILL_AMT1`	3.4339e−08
`PAY_AMT1`	4.8062e−07
`PAY_AMT3`	3.4338e−08

malicious purposes disqualify it from being used as designed. Given that explainable ML techniques have already been released in popular open source software, it seems there is a need for agreed-upon best practices and education on responsible use.

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



(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 few well-known publications have focused either on white-box modeling techniques (e.g. [32], [36]) or on post-hoc explanations (e.g. [25], [28]), but the two can be used together in the context of a broader and more human-centered machine learning 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.[19] This enables the practitioner to see accurate numeric feature contributions for each model prediction in addition to the entire directed graph of the decision tree. Even for interpretable models, such as linear models and decision trees, it has been shown that Shapley values present accuracy and consistency advantages over standard feature attribution methods [22], [24], [25]. In Figure 3, a simple decision tree, g_{tree}, is trained on the UCI credit card data set 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

[19] Code to replicate Figure 3 is available here: https://github.com/jphall663/xai_misconceptions.

displayed in 3b. The Shapley values are helpful because they highlight the local importance of `PAY_2`, the individual's second most recent repayment status, which could be underestimated by examing the decision policy alone. The Shapley values also enable the ranking of input features for each model decision, which is likely helpful for FCRA and ECOA compliance. Another twist on the idea of combining explainable AI methods and white-box models is described in "Surrogate Assisted Feature Extraction for Machine Learning (SAFE ML)" [15]. In the SAFE ML approach, features learned by more complex models are extracted and used in an explainable fashion to increase the accuracy of more interpretable models. Aren't either of these augmented processes more desirable than either a white-box model or post-hoc explanations alone?

Misconception Corollary: Explainable ML Methods and Fairness Methods are Somehow Mutually Exclusive: Like white-box models, fairness methods (e.g. [8], [18]) are often presented in different articles than post-hoc explanatory methods. However, in banks, using post-hoc explanatory tools such as partial dependence plots to comply with model documentation guidance often goes hand-in-hand with using disparate impact analysis to comply with fair lending regulations.[20,21,22]

Table 2: Basic group disparity metrics across different marital statuses for monotonically constrained GBM model, g_{mono}, trained on the UCI credit card dataset.

	Adverse Impact Disparity	Accuracy Disparity	TPR Disparity	TNR Disparity	FPR Disparity	FNR Disparity
`married`	1.00	1 .00	1.00	1.00	1.00	1.00
`single`	0.89	1.03	0.99	1.03	0.85	1.01
`divorced`	1.01	0.93	0.81	0.96	1.25	1.22
`other`	0.26	1.12	0.62	1.17	0	1.44

Table 2 displays basic group parity metrics for a monotonically constrained GBM model, g_{mono}, trained on the UCI credit card data.[23] 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 martial status of `other` (for which there is very little training data).

Conclusion

Machine learning systems are used today to make life-altering decisions about employment, bail, parole, and lending.[24] The scope of decisions delegated to machine learning systems seems likely only to expand in the future. Many researchers and practitioners are tackling disparate impact, inaccuracy, privacy violations, and security vulnerabilities with a number of brilliant, but perhaps siloed, approaches. By addressing some common explainable ML misconceptions, this short text also gives examples of combining innovations from several sub-disciplines of machine learning 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 human-centered type of machine learning potentially suitable for use in business- and life-critical decision support.

[20] Working paper: SR 11-7, Validation and Machine Learning Models. Tony Yang, CFA, CPA, FRM. KPMG USA.
[21] White paper: https://www.aba.com/Compliance/Documents/FairLendingWhitePaper2017Apr.pdf
[22] Policy Statement on Discrimination in Lending: https://www.govinfo.gov/content/pkg/FR-1994-04-15/html/94-9214.htm
[23] Code to replicate Table 2 is available here: https://github.com/jphall663/xai_misconceptions.
[24] ICLR 2019 model debugging workshop CFP: https://debug-ml-iclr2019.github.io/

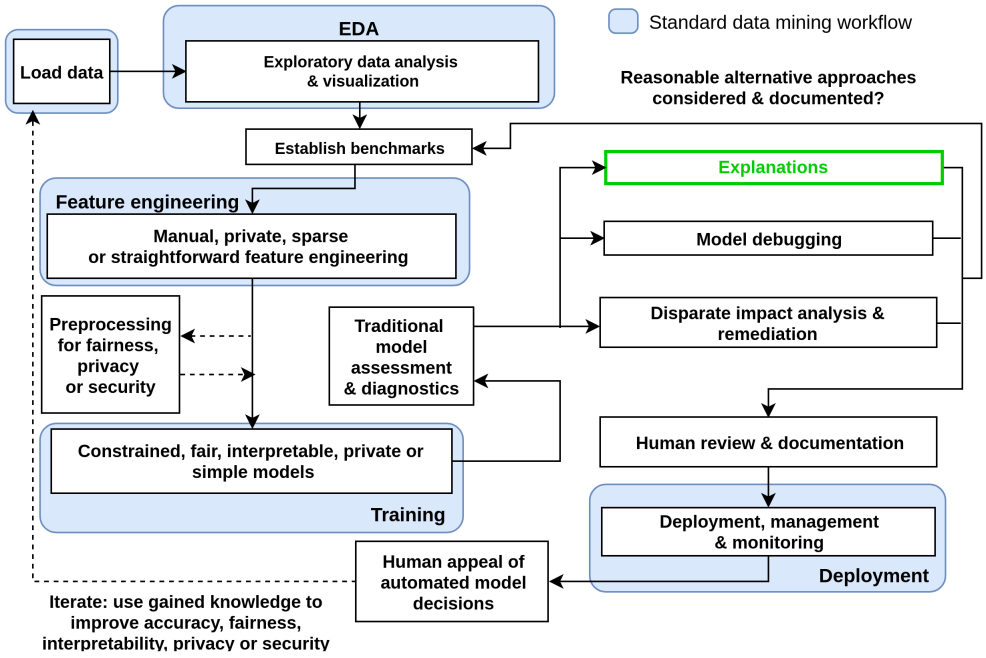


Figure 4: A diagram of a proposed human-centered machine learning workflow in which explanations (highlighted in green) are used along with interpretable models, disparate impact analysis and remediation techniques, and other review and appeal mechanisms to create a fair, accountable, and transparent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

Acknowledgemnts

The author thanks Przemyslaw Biecek, Pramit Choudhary, Navdeep Gill, and Christoph Molnar for their helpful input and insights.

References

[1] Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. Fairwashing: the Risk of Rationalization. *arXiv preprint arXiv:1901.09749*, 2019. URL: https://arxiv.org/pdf/1901.09749.pdf.

[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] Daniel W. Apley. 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015. URL: https://arxiv.org/pdf/1412.3756.pdf.

[9] Anthony W. Flores, Kristin Bechtel, and Christopher T. Lowenkamp. 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*, 80:38, 2016. URL: https://www.researchgate.net/profile/Christopher_Lowenkamp/publication/306032039_False_Positives_False_Negatives_and_False_Analyses_A_Rejoinder_to_Machine_Bias_There%27s_Software_Used_Across_the_Country_to_Predict_Future_Criminals_And_it%27s_Biased_Against_Blacks/links/57ab619908ae42ba52aedbab/False-Positives-False-Negatives-and-False-Analyses-A-Rejoinder-to-Machine-Bias-Theres-Software-Used-Across-th pdf.

[10] Sorelle A. Friedler, Chitradeep Dutta Roy, Carlos Scheidegger, and Dylan Slack. Assessing the Local Interpretability of Machine Learning Models. *arXiv preprint arXiv:1902.03501*, 2019. URL: https://arxiv.org/pdf/1902.03501.pdf.

[11] 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.

[12] 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.

[13] 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.

[14] Krishna M. Gopinathan, Louis S. Biafore, William M. Ferguson, Michael A. Lazarus, Anu K. Pathria, and Allen Jost. Fraud Detection using Predictive Modeling, October 6 1998. US Patent 5,819,226. URL: https://patents.google.com/patent/US5819226A.

[15] Alicja Gosiewska, Aleksandra Gacek, Piotr Lubon, and Przemyslaw Biecek. SAFE ML: Surrogate Assisted Feature Extraction for Model Learning. *arXiv preprint arXiv:1902.11035*, 2019. URL: https://arxiv.org/pdf/1902.11035v1.pdf.

[16] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys (CSUR)*, 51(5):93, 2018. URL: https://arxiv.org/pdf/1802.01933.pdf.

[17] Patrick Hall. On the Art and Science of Machine Learning Explanations. In *JSM Proceedings, Statistical Computing Section*, pages 1781–1799. American Statistical Association, 2018. URL: https://github.com/jphall663/jsm_2018_paper.

[18] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016. URL: http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf.

[19] 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.

[20] 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.

[21] M. Lichman. UCI Machine Learning Repository, 2013. URL: http://archive.ics.uci.edu/ml.

[22] Stan Lipovetsky and Michael Conklin. Analysis of Regression in Game Theory Approach. *Applied Stochastic Models in Business and Industry*, 17(4):319–330, 2001.

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

[24] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. Consistent Individualized Feature Attribution for Tree Ensembles. In Been Kim, Dmitry M. Malioutov, Kush R. Varshney, and Adrian Weller, editors, *Proceedings of the 2017 ICML Workshop on Human Interpretability in Machine Learning (WHI 2017)*, pages 15–21. ICML WHI 2017, 2017. URL: https://openreview.net/pdf?id=ByTKSo-m-.

[25] 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.

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

[27] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. Quantifying Interpretability of Arbitrary Machine Learning Models Through Functional Decomposition. *arXiv preprint arXiv:1904.03867*, 2019. URL: https://arxiv.org/pdf/1904.03867.pdf.

[28] 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.

[29] Cynthia Rudin. Please Stop Explaining Black Box Models for High Stakes Decisions. *arXiv preprint arXiv:1811.10154*, 2018. URL: https://arxiv.org/pdf/1811.10154.pdf.

[30] 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.

[31] 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.

[32] Berk Ustun and Cynthia Rudin. Supersparse Linear Integer Models for Optimized Medical Scoring Systems. *Machine Learning*, 102(3):349–391, 2016. URL: https://users.cs.duke.edu/~cynthia/docs/UstunTrRuAAAI13.pdf.

[33] 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.

[34] Adrian Weller. Challenges for Transparency. *arXiv preprint arXiv:1708.01870*, 2017. URL: https://arxiv.org/pdf/1708.01870.pdf.

[35] Mike Williams et al. *Interpretability*. Fast Forward Labs, 2017. URL: https://www.cloudera.com/products/fast-forward-labs-research.html.

[36] Hongyu Yang, Cynthia Rudin, and Margo Seltzer. Scalable Bayesian Rule Lists. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017. URL: https://arxiv.org/pdf/1602.08610.pdf.