Responsible Machine Learning*

Lecture 2: Post-hoc Explanations

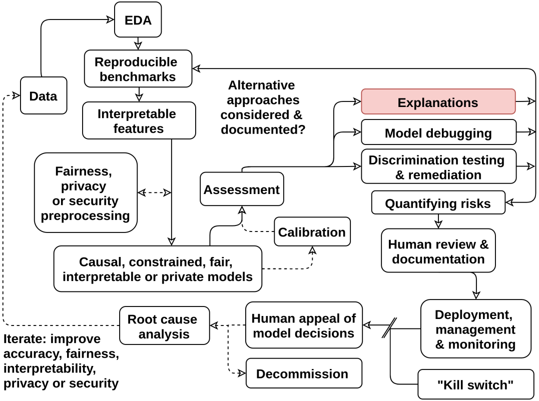
Patrick Hall

The George Washington University

May 29, 2020

Contents

A Responsible Machine Learning Workflow[2]


EDA
Reproducible benchmarks
Data
Interpretable features
Alternative approaches considered & documented?
Explanations
Model debugging
Discrimination testing & remediation
Fairness, privacy or security preprocessing
Assessment
Quantifying risks
Calibration
Human review & documentation
Causal, constrained, fair, interpretable or private models
Deployment, management & monitoring
Root cause analysis
Human appeal of model decisions
Iterate: improve accuracy, fairness, interpretability, privacy or security
Decommission
"Kill switch"

[2] *A Responsible Machine Learning Workflow*

What is an Explanation in Machine Learning (ML)?

> *"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."*
>
> — Sameer Singh, *UCI*

Variously defined along with aliases or similar concepts:

- "Towards a Rigorous Science of Interpretable Machine Learning" (Doshi-Velez and Kim [9])
- "Explaining Explanations" (Gilpin et al. [14])
- "A Survey Of Methods For Explaining Black Box Models" (Guidotti et al. [17])
- "The Mythos of Model Interpretability" (Lipton [25])
- *Interpretable Machine Learning* (Molnar [28])
- "Interpretable Machine Learning: Definitions, Methods, and Applications" (Murdoch et al. [30])
- "Challenges for Transparency" (Weller [43]).

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 [27], Shapley [35])
- **Global and local surrogate models**:
  - Decision tree variants (Bastani, Pu, and Solar-Lezama [6], Craven and Shavlik [8])
  - Anchors (Ribeiro, Singh, and Guestrin [32])
  - Local interpretable model-agnostic explanations (LIME) (Ribeiro, Singh, and Guestrin [33])
- **Global and local visualizations of trained model predictions**:
  - Accumulated local effects (ALE) (Apley [4])
  - Partial dependence (Friedman, Hastie, and Tibshirani [12])
  - Individual conditional expectation (ICE) (Goldstein et al. [15])

Shapley Value

Shapley explanations, including TreeSHAP and even certain implementations of LIME, are a class of additive, locally accurate feature contribution measures with long-standing theoretical support ([27]).

For some observation $\mathbf{x} \in \mathcal{X}$, Shapley explanations take the form:

$$\phi_j = \underbrace{\sum_{S \subseteq \mathcal{P} \setminus \{j\}} \frac{|S|!(\mathcal{P} - |S| - 1)!}{\mathcal{P}!}}_{\text{weighted average over all subsets in } \mathbf{X}} \underbrace{[(S \cup \{j\}) - g_x(S)]}_{g \text{ "without" } x_j} \tag{1}$$

$$g(\mathbf{x}) = \phi_0 + \sum_{j=0}^{j=\mathcal{P}-1} \phi_j \mathbf{z}_j \tag{2}$$

Surrogate Decision Trees (DT)

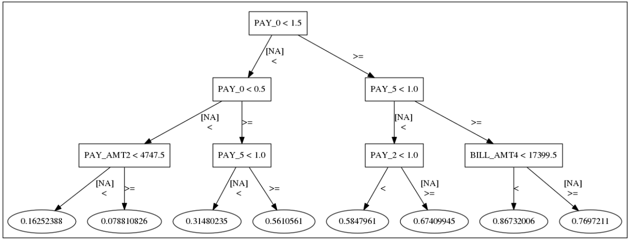


Figure: h_{tree} for Taiwanese credit card data [23], and for machine-learned GBM response function g.

- Given a learned function g and set of predictions $g(\mathbf{X})$, a surrogate DT can be trained: $\mathbf{X}, g(\mathbf{X}) \xrightarrow{\mathcal{A}_{\text{surrogate}}} h_{\text{tree}}$.
- h_{tree} displays a low-fidelity, high-interpretability flow chart of g's decision making process, and important features and interactions in g.

Surrogate Decision Trees (DT)

- Always use error measures to assess the trustworthiness of h_{tree}.
- Prescribed methods ([8]; [5]) for training h_{tree} do exist. In practice, straightforward cross-validation approaches are typically sufficient.
- Comparing cross-validated training error to traditional training error can give an indication of the stability of the single tree model, h_{tree}.
- Hu et al. (2018) use local linear surrogate models, h_{GLM}, in h_{tree} leaf nodes to increase overall surrogate model fidelity while also retaining a high degree of interpretability.

Local Interpretable Model-agnostic Explanations (LIME)

Ribeiro, Singh, and Guestrin (2016) define LIME for some observation $\mathbf{x} \in \mathcal{X}$:

$$\underset{h \in \mathcal{H}}{\arg\min}\, \mathcal{L}(g, h, \pi_{\mathbf{X}}) + \Omega(h)$$

Here g is the function to be explained, h is an interpretable surrogate model of g, often a linear model h_{GLM}, $\pi_{\mathbf{X}}$ is a weighting function over the domain of g, and $\Omega(h)$ limits the complexity of h.

Typically, h_{GLM} is constructed such that $\mathbf{X}^{(*)}, g(X^{(*)}) \xrightarrow{\mathcal{A}_{\text{surrogate}}} h_{\text{GLM}}$, where $\mathbf{X}^{(*)}$ is a generated sample, $\pi_{\mathbf{X}}$ weighs $\mathbf{X}^{(*)}$ samples by their Euclidean similarity to $\mathbf{x}$, local feature importance is estimated using $\beta_j x_j$, and L_1 regularization is used to induce a simplified, sparse h_{GLM}.

Local Interpretable Model-agnostic Explanations (LIME)

- LIME is ideal for creating low-fidelity, highly interpretable explanations for non-DT models and for neural network models trained on unstructured data, e.g. deep learning.
- Always use regression fit measures to assess the trustworthiness of LIMEs.
- LIME can be difficult to deploy, but there are highly deployable variants. ([19]; [18])
- Local feature importance values are offsets from a local intercept.
  - Note that the intercept in LIME can account for the most important local phenomena.
  - Generated LIME samples can contain large proportions of out-of-range data that can lead to unrealistic intercept values.

- To increase the fidelity of LIMEs, try LIME on discretized input features and on manually constructed interactions.
- Use cross-validation to construct standard deviations or even confidence intervals for local feature importance values.
- LIME can fail, particularly in the presence of extreme nonlinearity or high-degree interactions.

Partial Dependence (PD) and Individual Conditional Expectation (ICE)

- Following Friedman, Hastie, and Tibshirani (2001) a single feature $X_j \in \mathbf{X}$ and its complement set $X_{(-j)} \in \mathbf{X}$ (where $X_j \cup X_{(-j)} = \mathbf{X}$) is considered.
- PD(X_j, g) for a given feature X_j is estimated as the average output of the learned function g when all the components of X_j are set to a constant $x \in \mathcal{X}$ and $X_{(-j)}$ is left untouched.
- ICE$(X_j, \mathbf{x}^{(i)}, g)$ for a given observation $\mathbf{x}^{(i)}$ and feature X_j is estimated as the output of the learned function g when $x_j^{(i)}$ is set to a constant $x \in \mathcal{X}$ and $\mathbf{x}^{(i)} \in X_{(-j)}$ are left untouched.
- PD and ICE curves are usually plotted over some set of interesting constants $x \in \mathcal{X}$.

Partial Dependence (PD) and Individual Conditional Expectation (ICE)

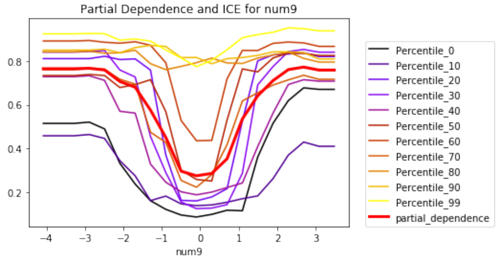


Figure: PD and ICE curves for $X_j = \text{num}_9$, for known signal generating function $f(\mathbf{X}) = \text{num}_1 * \text{num}_4 + |\text{num}_8| * \text{num}_9^2 + e$, and for machine-learned GBM response function g.

Overlaying PD and ICE curves is a succinct method for describing global and local prediction behavior and can be used to detect interactions. ([15])

Partial Dependence (PD) and Individual Conditional Expectation (ICE)

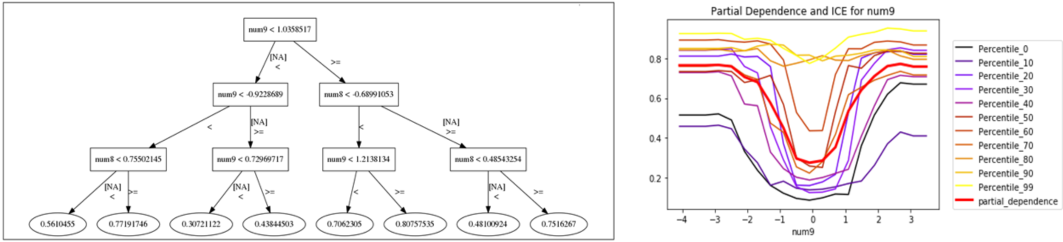


Figure: Surrogate DT, PD, and ICE curves for $X_j = \text{num}_9$, for known signal generating function $f(\mathbf{X}) = \text{num}_1 * \text{num}_4 + |\text{num}_8| * \text{num}_9^2 + e$, and for machine-learned GBM response function g.

Combining Surrogate DT models with PD and ICE curves is a convenient method for detecting, confirming, and understanding important interactions.

Why Explainable ML?

Responsible Use of Explainable ML can enable:

- Human learning from machine learning
- Human appeal of automated decisions
- Regulatory compliance[3]
- White-hat hacking and security audits of ML models

Even logistic regression is often "explained", or post-processed, for credit scoring, e.g. max. points lost method and adverse action notices.

[3] In the U.S., interpretable models, explanations, and the 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, the Fair Credit Reporting Act, the Fair Housing Act, Federal Reserve SR 11-7, and the European Union (EU) General Data Protection Regulation (GDPR) Article 22 [44].

Why Propose Guidelines?

Misuse and Abuse of Explainable ML can enable:

- Model and data stealing (Tramèr et al. [40], Shokri et al. [38], Shokri, Strobel, and Zick [37])
- False justification for harmful black-boxes, e.g. "fairwashing" (Aïvodji et al. [1], Rudin [34])

Explainable ML is already in-use:

- Numerous open source[4] and commercial packages[5] are available today.
- At least gradient-based feature attribution, partial dependence, and surrogate models are used for model validation in financial services today.[6,7]

Regulatory guidance is not agreed upon yet.[8]

[4] Please contribute: https://github.com/jphall663/awesome-machine-learning-interpretability.

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

[6] See: https://ww2.amstat.org/meetings/jsm/2019/onlineprogram/AbstractDetails.cfm?abstractid=303053.

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

[8] See: https://www.americanbanker.com/news/regulators-must-issue-ai-guidance-or-fdic-will-mcwilliams.

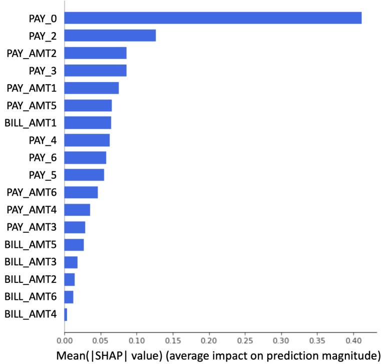
Guidelines for Responsible Use of Explainable ML

1. **Use explainable ML to enhance understanding.**
2. **Learn how explainable ML is used for nefarious purposes.**
3. **Augment surrogate models with direct explanations.**
4. **Use highly transparent mechanisms for high stakes applications (Rudin [34]).**

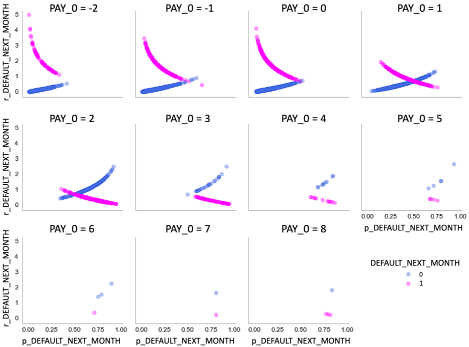
1: Use Explainable ML to **Enhance Understanding**

- Explanations enhance understanding **directly**, and increase trust as a **side-effect**.
- Models can be **understood and not trusted**, and **trusted but not understood**.
- Explanations **alone** are neither necessary nor sufficient for trust.
- Good explanations **enable human appeal** of model decisions.

Understanding Without Trust



g_{mono} monotonically-constrained probability of default (PD) classifier trained on the UCI credit card dataset over-emphasizes the most important feature, a customer's most recent repayment status, PAY_0 [23].



g_{mono} also struggles to predict default for favorable statuses, $-2 \leq \texttt{PAY\_0} < 2$, and often cannot predict on-time payment when recent payments are late, PAY_0 ≥ 2.

Trust Without 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 (Gopinathan et al. [16]). When these models performed well, they were trusted.[9] However, they were not explainable or well-understood by contemporary standards.

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

2: Explainable ML Can be Used for **Nefarious Purposes**

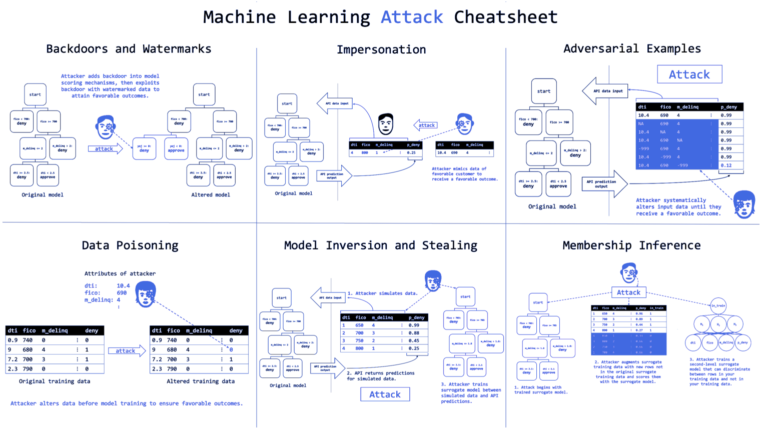
When unintentionally misused, explainable ML can act as a faulty safeguard for potentially harmful black-boxes.

When intentionally abused, explainable ML can be used for:

- Stealing data, models, or other intellectual property.
- *Fairwashing*, to mask the sociological biases of a discriminatory black-box.

AI Incident: ML Hacking

Many ML hacks use, or are exacerbated by, explainable ML techniques.[10]



Machine Learning Attack Cheatsheet

Backdoors and Watermarks

Attacker adds backdoor into model scoring mechanisms, then exploits backdoor with watermarked data to attain favorable outcomes.

Impersonation

dti	fico	m_delinq		p_deny
4	800	1	⋮	0.25

dti	fico	m_delinq	
10.4	690	4	⋮

Attacker mimics data of favorable customer to receive a favorable outcome.

Adversarial Examples

dti	fico	m_delinq		p_deny
10.4	690	4	⋮	0.99
NA	690	4	⋮	0.99
10.4	NA	4	⋮	0.99
10.4	690	NA	⋮	0.99
-999	690	4	⋮	0.99
10.4	-999	4	⋮	0.99
10.4	690	-999	⋮	0.12

Attacker systematically alters input data until they receive a favorable outcome.

Data Poisoning

Attributes of attacker

dti: 10.4
fico: 690
m_delinq: 4
⋮

dti	fico	m_delinq		deny
0.9	740	0	⋮	0
9	680	4	⋮	1
7.2	700	3	⋮	1
2.3	790	0	⋮	0

Original training data

dti	fico	m_delinq		deny
0.9	740	0	⋮	0
9	680	4	⋮	0
7.2	700	3	⋮	1
2.3	790	0	⋮	0

Altered training data

Attacker alters data before model training to ensure favorable outcomes.

Model Inversion and Stealing

1. Attacker simulates data.

dti	fico	m_delinq		p_deny
1	650	4	⋮	0.99
2	700	3	⋮	0.88
3	750	2	⋮	0.45
4	800	1	⋮	0.25

2. API returns predictions for simulated data.

3. Attacker trains surrogate model between simulated data and API predictions.

Membership Inference

1. Attack begins with trained surrogate model.

dti	fico	m_delinq		p_deny	in_train
1	650	4	⋮	0.96	1
2	700	3	⋮	0.89	1
3	750	2	⋮	0.44	1
4	800	1	⋮	0.27	1
[illegible]	[illegible]	[illegible]	⋮	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	⋮	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	⋮	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	⋮	[illegible]	[illegible]

2. Attacker augments surrogate training data with new rows not in the original surrogate training data and scores them with the surrogate model.

3. Attacker trains a second-level surrogate model that can discriminate between rows in your training data and not in your training data.

[10] See https://github.com/jphall663/secure_ML_ideas for full size image and more information.

Case 2.1: White-hat Attacks Can Crack Potentially Harmful Black-boxes

The flip-side of the dark side is community oversight of black-boxes.

Recent high profile analyses of commercial black-boxes, e.g. ...

- Propublica and COMPAS (Angwin et al. [3])[11]
- Gendershades and Rekognition (Buolamwini and Gebru [7], Raji and Buolamwini [31])

... **could** be characterized as white-hat attacks on proprietary black-boxes (respectively, model stealing and adversarial examples).

[11] This presentation makes no claim on the quality of the analysis in Angwin et al. (2016), which has been criticized, but is simply stating that such cracking is possible [3], [11].

Case 2.2: Explanation ***is Not*** a Front Line Fairness Tool

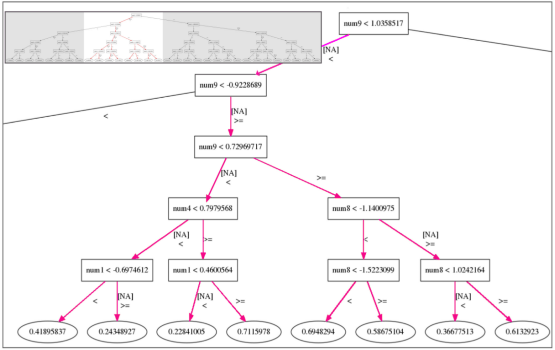
Use fairness tools, e.g. ...

- Disparate impact testing (Feldman et al. [10])
- Reweighing (Kamiran and Calders [20])
- Reject option based classification (Kamiran, Karim, and Zhang [21])
- Adversarial de-biasing (Zhang, Lemoine, and Mitchell [46])
- aequitas, AIF360, Themis, themis-ml

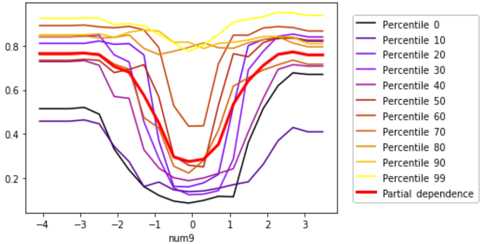
... for fairness tasks: bias testing, bias remediation, and to establish trust.

Explanations can be used to understand and augment such results.

3: Augment **Surrogate Models** with **Direct Explanations**



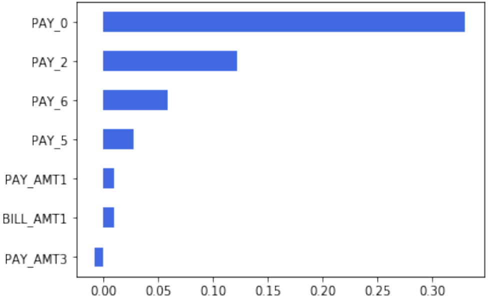
Naïve h_{tree}, *a surrogate model*, forms an approximate overall flowchart for the explained model, g_{GBM}.



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

h_{tree} displays known interactions in $f = X_{\text{num1}} * X_{\text{num4}} + |X_{\text{num8}}| * X_{\text{num9}}^2$ 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 also highlighted by the divergence of partial dependence and ICE curves for $\sim -1 < X_{\text{num9}} <\sim 1$.

Example 3.1: Augment LIME with Direct Explanations



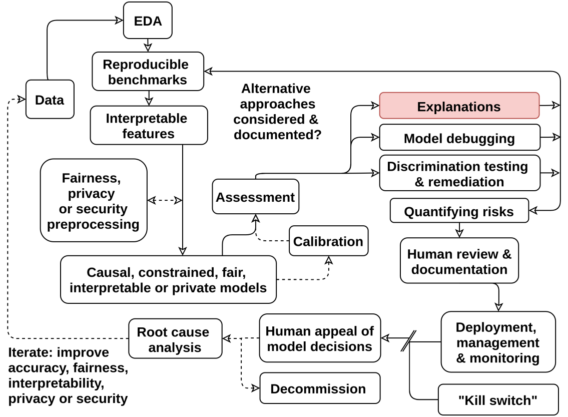
Locally accurate Shapley contributions for a high risk individual's probability of default as predicted by a simple decision tree model, $g_{\mathbf{tree}}$. See slide 29 for a directed graph representation of $g_{\mathbf{tree}}$.

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

Coefficients for a local linear interpretable model, $h_{\mathbf{GLM}}$, with an intercept of 0.77 and an R^2 of 0.73, trained between the original inputs and predictions of $g_{\mathbf{tree}}$ for a segment of the UCI credit card dataset with late most recent repayment statuses, $\mathbf{X}_{PAY\_\mathbf{0}>\mathbf{1}}$.

Because h_{GLM} is relatively well-fit and has a logical intercept, 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.

4: Use Highly Transparent Mechanisms for **High Stakes Applications**


EDA
Reproducible benchmarks
Data
Interpretable features
Alternative approaches considered & documented?
Explanations
Model debugging
Discrimination testing & remediation
Quantifying risks
Fairness, privacy or security preprocessing
Assessment
Calibration
Human review & documentation
Causal, constrained, fair, interpretable or private models
Root cause analysis
Human appeal of model decisions
Deployment, management & monitoring
Iterate: improve accuracy, fairness, interpretability, privacy or security
Decommission
"Kill switch"

A diagram of a proposed workflow in which explanations 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.

Case 4.1: Use **Interpretable Models** for High Stakes Applications (Rudin [34])

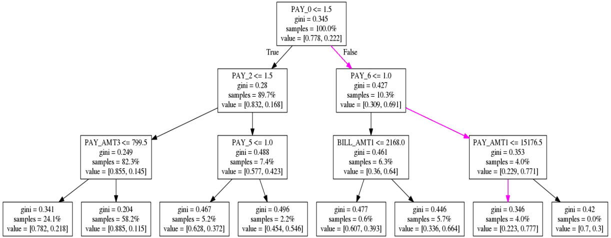
In addition to penalized GLM, decision trees, and conventional rule-based models, many other types of accurate and interpretable models are available today, e.g. ...

- Explainable boosting machine (EBM)
- Monotonic GBM in h2o or XGBoost
- RuleFit (Friedman and Popescu [13])
- Super-sparse linear integer model (SLIM) (Ustun and Rudin [41])
- Explainable neural network (XNN) (Vaughan et al. [42])
- Scalable Bayesian rule list (Yang, Rudin, and Seltzer [45])

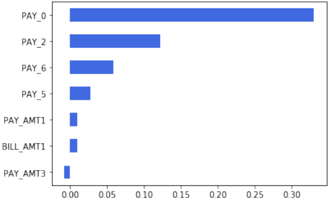
... use them for human-centered or other high stakes ML applications.[12]

[12] There are shades of interpretability in models. Interpretability is probably not a binary, on-off quality. For instance see Figure 3: https://arxiv.org/pdf/1904.03867.pdf [29].

Case 4.2: Explanations and Interpretable Models are **Not Mutually Exclusive**



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 high risk individuals is highlighted in fuschia.



Locally accurate Shapley contributions for the highlighted individual's probability of default. See slide 26 for LIMEs for the high risk customers in g_{tree}.

The Shapley values are helpful because they highlight the local importance of features not on the decision path, which could be underestimated by examining the decision policy alone.

Interlude: An Ode to the Shapley Value

1. **In the beginning**: A Value for N-Person Games, 1953 [35]
2. **Nobel-worthy contributions**: *The Shapley value: Essays in honor of Lloyd S. Shapley*, 1988 [36]
3. **Shapley regression**: Analysis of Regression in Game Theory Approach, 2001 [24]
4. **First reference in ML?** Fair Attribution of Functional Contribution in Artificial and Biological Networks, 2004 [22]
5. **Into the ML research mainstream, i.e. JMLR**: An Efficient Explanation of Individual Classifications using Game Theory, 2010 [39]
6. **Into the real-world data mining workflow ... *finally***: Consistent Individualized Feature Attribution for Tree Ensembles, 2017[13] [26]
7. **Unification**: A Unified Approach to Interpreting Model Predictions, 2017[14] [27]

[13] See h2o, LightGBM, or XGBoost for implementation.

[14] See shap for implementation.

Case 4.3: Explanation and Fairness Techniques are **Not Mutually Exclusive**

	Adverse Impact Disparity	Accuracy Disparity	TPR Disparity	TNR Disparity	FPR Disparity	FNR Disparity
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

Basic group disparity metrics across different marital statuses for monotonically constrained GBM model, g_{mono}, trained on the UCI credit card dataset. See slide 19 for global Shapley feature importance for g_{mono} and slide 24 for an important note about explanation and fairness techniques.

Many fairness toolkits are available today: aequitas, AIF360, Themis, themis-ml.

Traditional disparate impact testing tools are best-suited for constrained models because average group metrics cannot reliably identify local instances of discrimination that can occur when using complex, unconstrained models.

Acknowledgments

Some of the best engineers, researchers, and business leaders in the world!

Christoph Molnar, Doug Deloy, Josephine Wang, Kerry O'Shea, Ladislav Ligart, Leland Wilkinson, Mark Chan, Martin Dvorak, Mateusz Dymczyk, Megan and Michal Kurka, Mike Williams, Navdeep Gill, Pramit Choudhary, Przemyslaw Biecek, Sameer Singh, Sri Ambati, Wen Phan, Zac Taschdjian, and Lisa Song.[15]

[15] My world anyway ... and in alphabetical order by first name.

References

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

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

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.

Julia Angwin et al. "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks.." In: *ProPublica* (2016). URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

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.

Osbert Bastani, Carolyn Kim, and Hamsa Bastani. "Interpreting Blackbox Models via Model Extraction." In: *arXiv preprint arXiv:1705.08504* (2017). URL: https://arxiv.org/pdf/1705.08504.pdf.

References

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.

Joy Buolamwini and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In: *Conference on fairness, accountability and transparency*. URL: http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf. 2018, pp. 77–91.

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.

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.

Michael Feldman et al. "Certifying and Removing Disparate Impact." In: *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. URL: https://arxiv.org/pdf/1412.3756.pdf. ACM. 2015, pp. 259–268.

References

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." In: *Fed. Probation* 80 (2016). URL: https://bit.ly/2Gesf9Y, p. 38.

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.

Jerome H. Friedman, Bogdan E Popescu, et al. "Predictive Learning via Rule Ensembles." In: *The Annals of Applied Statistics* 2.3 (2008). URL: https://projecteuclid.org/download/pdfview_1/euclid.aoas/1223908046, pp. 916–954.

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.

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.

References

Krishna M. Gopinathan et al. *Fraud Detection using Predictive Modeling*. US Patent 5,819,226. URL: https://patents.google.com/patent/US5819226A. 1998.

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.

Patrick Hall et al. *Machine Learning Interpretability with H2O Driverless AI*. 2017. URL: http://docs.h2o.ai/driverless-ai/latest-stable/docs/booklets/MLIBooklet.pdf.

Linwei Hu et al. "Locally Interpretable Models and Effects Based on Supervised Partitioning (LIME-SUP)." In: *arXiv preprint arXiv:1806.00663* (2018). URL: https://arxiv.org/ftp/arxiv/papers/1806/1806.00663.pdf.

Faisal Kamiran and Toon Calders. "Data Preprocessing Techniques for Classification Without Discrimination." In: *Knowledge and Information Systems* 33.1 (2012). URL: https://link.springer.com/content/pdf/10.1007/s10115-011-0463-8.pdf, pp. 1–33.

Faisal Kamiran, Asim Karim, and Xiangliang Zhang. "Decision Theory for Discrimination-aware Classification." In: *2012 IEEE 12th International Conference on Data Mining*. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.722.3030&rep=rep1&type=pdf. IEEE. 2012, pp. 924–929.

References

Alon Keinan et al. "Fair Attribution of Functional Contribution in Artificial and Biological Networks." In: *Neural Computation* 16.9 (2004). URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.436.6801&rep=rep1&type=pdf, pp. 1887–1915.

M. Lichman. *UCI Machine Learning Repository*. URL: http://archive.ics.uci.edu/ml. 2013.

Stan Lipovetsky and Michael Conklin. "Analysis of Regression in Game Theory Approach." In: *Applied Stochastic Models in Business and Industry* 17.4 (2001), pp. 319–330.

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

Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. "Consistent Individualized Feature Attribution for Tree Ensembles." In: *Proceedings of the 2017 ICML Workshop on Human Interpretability in Machine Learning (WHI 2017)*. Ed. by Been Kim et al. URL: https://openreview.net/pdf?id=ByTKSo-m-. ICML WHI 2017, 2017, pp. 15–21.

References

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.

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

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

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.

Inioluwa Deborah Raji and Joy Buolamwini. "Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products." In: *AAAI/ACM Conf. on AI Ethics and Society*. Vol. 1. URL: http://www.aies-conference.com/wp-content/uploads/2019/01/AIES-19_paper_223.pdf. 2019.

References

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.

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.

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.

Lloyd S. Shapley. "A Value for N-Person Games." In: *Contributions to the Theory of Games* 2.28 (1953). URL: http://www.library.fa.ru/files/Roth2.pdf#page=39, pp. 307–317.

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

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.

References

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.

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

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.

Berk Ustun and Cynthia Rudin. "Supersparse Linear Integer Models for Optimized Medical Scoring Systems." In: *Machine Learning* 102.3 (2016). URL: https://users.cs.duke.edu/~cynthia/docs/UstunTrRuAAAI13.pdf, pp. 349–391.

Joel Vaughan et al. "Explainable Neural Networks Based on Additive Index Models." In: *arXiv preprint arXiv:1806.01933* (2018). URL: https://arxiv.org/pdf/1806.01933.pdf.

References

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

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

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

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. "Mitigating Unwanted Biases with Adversarial Learning." In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. URL: https://arxiv.org/pdf/1801.07593.pdf. ACM. 2018, pp. 335–340.