Explainable AI in Healthcare

Muhammad Aurangzeb Ahmad, Carly Eckert, Ankur Teredesai, Vikas Kumar
KenSci Inc
“Does your car have any idea why my car pulled it over?”
Learning Objectives

• Why do we need explanations in AI and Machine Learning in Healthcare?
• What are Explanations in AI and Machine Learning in Healthcare?
• How does one choose between machine learning algorithms when explanations are needed?
• What are the different types of interpretable machine learning models?
• What does the future of explainable AI looks like in healthcare?
Terminology

Explain: Make (an idea or situation) clear to someone by describing it in more detail or revealing relevant facts.

Interpret: Explain the meaning of (information or actions)

Understand: Perceive the intended meaning of (words, a language, or a speaker)

Comprehend: Grasp mentally; understand

[Oxford Dictionary]
Interpretable machine learning refers to giving explanations of machine learning models to humans with domain knowledge.

Explanation: Why is the prediction being made?

Explanation to Human: The explanation should be comprehensible to humans in (i) natural language (ii) easy to understand representations

Domain Knowledge: The explanation should make sense to a domain expert.
Explainable ML is more than models

Each element constituent of the machine learning solution process needs to be explainable for the solution to be truly explainable

Machine Learning Solution
- Features
- Algorithm
- Model Parameters
- Model

Machine Learning User
- Cognitive Capacity
- Domain Knowledge
- Explanation Granularity
<table>
<thead>
<tr>
<th>Data type</th>
<th>Models/Tools</th>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>-EHR data - Insurance claims data</td>
<td>ML(logistic regression, XGBoost)</td>
<td>Predict outcomes (disease, death, readmission etc.)</td>
</tr>
<tr>
<td>-Clinical notes - Conversation text data</td>
<td>-Rule based approach (regular expression) - Deep learning approach</td>
<td>-Extract concepts from clinical notes - Knowledge graphs - Chat-bot - QA system</td>
</tr>
<tr>
<td>Medical image data (X-ray, CT, OCR image etc.)</td>
<td>CNN</td>
<td>-Detection: diagnosis of skin cancer lung nodule or diabetic reinopathy - Segmentation of tumor, histopathology</td>
</tr>
<tr>
<td>Time series data (EEG, ECG, vital sign data etc.)</td>
<td>HMM, RNN, CNN</td>
<td>-Heart disease - Sleep disorder (apnea) - ICU monitoring</td>
</tr>
<tr>
<td>Genomics data</td>
<td>GATK, QIIME</td>
<td>-Cancer mutation identification - Biomarker identification - Drug discovery</td>
</tr>
<tr>
<td>Other data (hospital operational data)</td>
<td>-ML(regression) - Queueing model</td>
<td>-Reduce operational cost - Improve patient experience - ER wait time and queueing</td>
</tr>
</tbody>
</table>
Need for Explanations in Machine Learning
When do we need explanations?

When fairness is critical:
• Any context where humans are required to provide explanations so that people cannot hide behind machine learning models

When consequences are far-reaching:
• Predictions can have far reaching consequences e.g., recommend an operation, recommend sending a patient to hospice etc.

When the cost of a mistake is high:
• Ex: misclassification of a malignant tumor can be costly and dangerous

When a new/unknown hypothesis is drawn:
• “It's not a human move. I've never seen a human play this move.” (Fan Hui)
• Pneumonia patients with asthma had lower risk of dying (Caruana et al. 2015)
When do we need explanations?

• When performance is critical:
• When compliance is required:
  • GDPR
  • Right to explanation
• When trust is necessary:
  • Predictive performance is not enough
Why do we need explanations now?

Sampling of headlines about failures of AI in healthcare?

More Implications (known/unknown)

More AI

More Data

Source: Scopus.com
Need for ML Explanation in Healthcare

- Algorithms to predict which pneumonia patients should be admitted to hospital for treatment
- Neural nets were far more accurate than classical statistical methods
- The regression and the neural net inferred that asthma patients treated for pneumonia had a lower mortality risk, and therefore, should not be admitted
- In fact, due to their underlying lung condition, these patients were usually admitted directly to the ICU, treated aggressively, and survived
Problems in Patient Flow
Characteristics of Explainable AI in Healthcare

- Trust
- Transparency
- Fidelity
- Domain Sense
- Consistency
- Generalizability
- Parsimony
Characteristics of Explainable AI in Patient Flow

- Transparency
- Domain sense
- Consistency
- Generalizability
- Parsimony
- Fidelity
- Trust
Operationalizing AI in Healthcare

Only a small fraction of real-world machine learning systems actually constitutes machine learning code.
Data in Operationalized AI in Healthcare

• Syntactic Correctness
  • Is the data in the correct format

• Morphological Correctness
  • Is the data within the range of possible values e.g., a blood pressure of 500 does not make sense

• Semantic Correctness
  • Do the variables actually correspond to what the semantics that are being ascribed to them
The Problem of Point Solutions

Anyone can do the math. An ML model alone doesn’t solve a healthcare problem...

...and lots of models becomes a problem very quickly.
Admission Prediction
What is the likelihood of the patient being admitted to the hospital

Transparency
Ability of the machine learning algorithm, model and the features to be understandable by the user of the system
The ML model for predicting Katherine’s likelihood of admission gives her a high likelihood (0.62).

Katherine’s physician has noted her age, health history, and vital signs and is surprised by this elevated risk score.

The physician knows that the risk model is a deep learning model so he cannot understand how it is working.

But, he can examines the top factors associated with prediction.
Transparent to whom?

- Transparency may mean different things to different people.
- Understanding Model Outputs:
  \[ y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, n \]
- Understanding Algorithms:
  \[ a_j^l = \sigma \left( \sum_k w_{jk}^l a_k^{l-1} + b_j^l \right) \]
- Understanding the algorithm may not be sufficient.
• The whole model must be understandable simultaneously

• One must be able to look at the model and understand the whole model without too much cognizing about the model

• Example: While both Decision Trees are explainable, *Decision Tree B* has the property of Simultability but *Decision Tree A* does not

(Lipton 2016)
Transparency: Decomposability

• Each component should also admit to an easy/intuitive explanation

• A linear model with highly engineered features vs. a linear model with simple feature

• Example: Model A is decomposable but Model B is not

Regression Variables
• Age
• Gender
• Race
• Diabetic
• Smoker

Target Variable
• Length of Stay

Regression Model A

Regression Variables
• $e^{\lambda t} (\cos(\mu t) + i \sin(\mu t))$
• \[ \sum_{n=1}^{\infty} c_n (z - a)^n \]
• $\frac{1}{z} + g'(z) + \sum_{n=0}^{\infty} \left( \frac{1}{z - n} + \frac{1}{n} \right)$

Target Variable
• Length of Stay

Regression Model B
• Guarantee that a model will converge
• Models like Regression Models, SVM etc. have this property
• Deep Learning does not have this property
Feedback Transparency refers to how change in the model will affect the model prediction.

- How do multi-objective optimization models affect each other e.g., optimizing reduction of risk of readmission and reduction of length of stay.

- Not an absolute requirement but rather nice to have property.

- Especially applicable to Scrutable machine learning systems.
Transparency: Examples

• **Transparent**
  - GAM
  - GA2M
  - Naïve Bayes
  - Regression Models
  - Falling Rule Lists
  - SLIM

• **Semi-Transparent**
  - Shallow Ensembles

• **Non-Transparent**
  - Deep Learning
  - SVM
  - Gradient Boosting Models
Locally Interpretable Model Explanations

Figure 4: Explaining an image classification prediction made by Google’s Inception network, highlighting positive pixels. The top 3 classes predicted are “Electric Guitar” ($p = 0.32$), “Acoustic guitar” ($p = 0.24$) and “Labrador” ($p = 0.21$).

Original Image $P(\text{tree frog}) = 0.54$
Shapley Values

• Game Theoretic Method for determining feature contributions
• Each feature is a ‘player’ in a game where the prediction is the payout
• The Shapley value - a method from coalitional game theory - tells us how to fairly distribute the ‘payout’ among the features.

\[
\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[ f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right].
\]
Domain Sense
The explanation should make sense in the domain of application and to the user of the system

ED Census Prediction
Predict the number of patients in the ED at a given time
Katherine and her physician discuss the problem of ED census prediction i.e., predicting the number of patients expected in the emergency department in a given time frame. The top features are temporal features (day of the week, month) which are not really helpful. The explanations are reconfigured to surface only factors that are modifiable. When de, a nurse sees the new dashboard but the modifiable factors are not really helpful to her. Explanation dashboard is customized for Whende. In the aggregate the top reason for high ED census corresponds to people getting drunk because of a college football game. Proper staffing and proper stocking of supplies can be done with relevant explanations.
• A physician requires different explanations as compared to a staffing planner
• Explanations need to be in the right language and also in the right context
• If actionability is required then the factors for explanations should reflect that
# Taxonomy of Explainable Factors

<table>
<thead>
<tr>
<th>Mutability</th>
<th>Interveniability</th>
<th>Actionability</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Immutable</td>
<td>Non-Intervenial</td>
<td>Actionable</td>
<td>Age, Sex, Ethnicity</td>
</tr>
<tr>
<td>Mutated</td>
<td>Intervenial</td>
<td>Intrinsic Heart rate variability Marital Status</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intervention</td>
<td>Temperature (in Apendectomy)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Appendicitis</td>
</tr>
<tr>
<td></td>
<td>Post-Intervenial</td>
<td>Actionable</td>
<td>Immunization</td>
</tr>
</tbody>
</table>

*Figure 1: Factors for Predicting Risk of Readmission*
• Actionability is context dependent and role dependent
• Actionability is NOT causality
• There may be circumstances were actionability is not possible
• Generalized Additive Models with pairwise interactions
• Accounts for interactions between the target variable and the feature space
• Possible to visualize the relationship between the target variable to gain insights into why the prediction is being made

(Carauna 2017)
Consistency

The explanation should be consistent across different models and across different runs of the model.

LWBS

Left without being seen refers to a patient leaving the facility without being seen by a physician.
• Dr. Marcos is examining the reasons for a patient who is predicted to leave without being seen and notices that the explanation for them leaving are different from what observed 4 hours ago.

• Upon investigation Katherine determines that the LIME model is being used for prediction which is non-deterministic, hence differences in explanations.

• Having different explanations for the same instance can be confusing for users, hence the need for consistency.
Evaluating Consistency

• Kendall’s Tau

\[ \tau = \frac{\text{(number of concordant pairs)} - \text{(number of discordant pairs)}}{n(n-1)/2} \]

• Kendall’ W

<table>
<thead>
<tr>
<th>W</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>W \leq 0.3</td>
<td>Weak agreement</td>
</tr>
<tr>
<td>0.3 &lt; W \leq 0.5</td>
<td>Moderate agreement</td>
</tr>
<tr>
<td>0.5 &lt; W \leq 0.7</td>
<td>Good agreement</td>
</tr>
<tr>
<td>W &gt; 0.7</td>
<td>Strong agreement</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Judges</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>8</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>7</td>
<td>2</td>
<td>6</td>
<td>4.5</td>
<td>4.5</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>5.5</td>
<td>7</td>
<td>5.5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>6.5</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>6.5</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>4.5</td>
<td>2</td>
<td>7</td>
<td>3</td>
<td>8</td>
<td>1</td>
<td>4.5</td>
</tr>
<tr>
<td>6</td>
<td>7.5</td>
<td>6</td>
<td>7.5</td>
<td>2.5</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>2.5</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>7</td>
<td>4</td>
<td>8</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>50.5</td>
<td>44.5</td>
<td>32.5</td>
<td>38</td>
<td>28.5</td>
<td>33</td>
<td>12</td>
<td>13</td>
</tr>
</tbody>
</table>
• Given the same dataset, multiple machine learning algorithms can be constructed with similar performance

• The explanations that are produced by multiple explainable algorithms should be very similar if not the same

• Wide divergence in explanations is a sign of problem with explanations or with the algorithm(s)
Rebuttal – explanation not always needed
Examples of when explanation is not needed

• Whende and Katherine are looking at ED arrivals prediction
• Kate shows the explanation factors to Whende but the performance of the corresponding model is relatively low (Precision = 0.34, Recall = 0.46)
• Whende states that getting explanations for ED arrivals prediction is not very important, performance takes precedence
Exceptions to Explanations

A. HEALTHY

B. DISEASED

Hemorrhages
“You can ask a human, but, you know, what cognitive psychologists have discovered is that when you ask a human you’re not really getting at the decision process. They make a decision first, and then you ask, and then they generate an explanation and that may not be the true explanation.”

- Peter Norvig
Machine Learning is used in problems where the size of the data and/or the number of variables is too large for humans to analyze.

What if the most parsimonious model is indeed too complex for humans to analyze or comprehend?

Ante-Hoc explanations may be impossible and post-hoc explanations would be incorrect.

Explanations may not be possible in some cases.
Parsimony
The explanation should be as simple as possible

Admission Disposition
Where in the hospital the patient should go once they are admitted
Parsimony

- MDL (Minimum Description Length) and Occam’s Razor
- Occam’s Razor in Machine Learning
  - Occam’s First Razor
  - Occam’s Second Razor
- Occam’s Razor in Interpretable Machine Learning
- The simplest explanation is not always the best one
Hickman’s Dictum and Chattam’s Anti-Razor

• Applying Occam’s Razor can be counter productive
• Occam’s Razor is context free
• Healthcare in Practice
  • Multiple competing hypothesis
• Heliocentric vs. Geocentric Models
• Hiccam’s Dictim
• Chattam’s Anti-Razor
Decision Lists

- Falling Rule Lists (FRL) [1]
- Bayesian Rule List (BRL) [2]
- Interpretable Decision Sets (IDS) [3]

<table>
<thead>
<tr>
<th>Conditions</th>
<th>Probability</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>If IrregularShape AND Age ≥ 60</td>
<td>85.22%</td>
<td>230</td>
</tr>
<tr>
<td>ELSE IF SpiculatedMargin AND Age ≥ 45</td>
<td>78.13%</td>
<td>64</td>
</tr>
<tr>
<td>ELSE IF IllDefinedMargin AND Age ≥ 60</td>
<td>69.23%</td>
<td>39</td>
</tr>
<tr>
<td>ELSE IF IrregularShape</td>
<td>63.46%</td>
<td>153</td>
</tr>
<tr>
<td>ELSE IF LobularShape AND Density ≥ 2</td>
<td>39.68%</td>
<td>63</td>
</tr>
<tr>
<td>ELSE IF RoundShape AND Age ≥ 60</td>
<td>26.09%</td>
<td>46</td>
</tr>
<tr>
<td>ELSE</td>
<td>10.38%</td>
<td>366</td>
</tr>
</tbody>
</table>

**Table 1**

Decision list for mammographic mass dataset.

Bayesian Rule Lists

• BRLs are decision lists—-a series of if-then statements

• BRLs discretize a high-dimensional, multivariate feature space into a series of simple, readily interpretable decision statements.

• Experiments show that BRLs have predictive accuracy on par with the current top ML algorithms (approx. 85-90% as effective) but with models that are much more interpretable

• if hemiplegia and age > 60
  • then stroke risk 58.9% (53.8%–63.8%)

• else if cerebrovascular disorder
  • then stroke risk 47.8% (44.8%–50.7%)

• else if transient ischaemic attack
  • then stroke risk 23.8% (19.5%–28.4%)

• else if occlusion and stenosis of carotid artery without infarction
  • then stroke risk 15.8% (12.2%–19.6%)

• else if altered state of consciousness and age > 60
  • then stroke risk 16.0% (12.2%–20.2%)

• else if age ≤ 70
  • then stroke risk 4.6% (3.9%–5.4%)

• else stroke risk 8.7% (7.9%–9.6%)

(Letham 2015)
Generalizability
The explanation should be generalizable

Length of Stay
How long is the patient going to stay in the facility
• Dr. Marcos is looking at a patient’s info for length of stay prediction. She observes that many of the instances have similar explanations.

• However the explanations do not appear to be generalizable to the whole population.
• **Local Models:** Models that give explanations at the level of an instance

• *Examples:* LIME, Shapley Values etc.

• **Global Models:** Models that give explanations

• *Examples:* Decision Trees, Rule Based Models etc.

• **Cohort Level Models:** A type of global models where the explanations are generated at the level of cohort

• *Examples:* Same as global models
Algorithm Generalizability

• Are the interpretable/explainable algorithms generalizable to all predictive algorithm, a particular class of algorithms or tied to a particular algorithm

• Model Agnostic Explanations:
  • Examples: LIME, Shapley Values etc.

• Model Class Specific Explanations:
  • Examples: Tree Explainers

• Model Specific Explanations:
  • Examples: CENs, Decision Trees, SLIM etc.
Dimensions of Explanations

• Data
  • What variables or features are most relevant for the prediction of length of stay?

• Prediction
  • Explain why certain patients are being predicted to have long length of stays

• Model
  • What are the patterns belonging to a particular category (long length of stay) typically look like?
• Explain the model prediction for one instance by measuring the difference between the original prediction and the one made with omitting a set of features

• Problem: Imputation may be needed in case of missing values

Robnik-Sikonja and Kononenko (2008)
**Rule Based Models**

**Rules for Length of Stay Prediction**

IF age > 55 AND gender = male AND condition = ‘COPD’ AND complication = ‘YES’

THEN

Length of stay = long (> 7 days)

**Relative Variable Importance**

**Top 5 Variables for Length of Stay Prediction**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>0.45</td>
</tr>
<tr>
<td>Gender</td>
<td>0.37</td>
</tr>
<tr>
<td>Diabetic</td>
<td>0.32</td>
</tr>
<tr>
<td>Race</td>
<td>0.21</td>
</tr>
<tr>
<td>Smoker</td>
<td>0.14</td>
</tr>
</tbody>
</table>

**Case Based Models**

**Example cases for Length of Stay Prediction**

Patient X is predicted to have a length of stay of 20 days because he is most similar to these 5 patients who on average had length of stay of 5 days.
Trust / Performance
The expectation that the corresponding predictive algorithm for explanations should have a certain performance

ICU Transfer Prediction
Predict if the patient will be transferred to the ICU
• The model has at least parity with the performance of human practitioners
• Example: Model B has human performance parity

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physician’s Prediction</td>
<td>0.73</td>
<td>0.71</td>
<td>0.72</td>
<td>0.60</td>
</tr>
<tr>
<td>Model A</td>
<td>0.82</td>
<td>0.68</td>
<td>0.74</td>
<td>0.65</td>
</tr>
<tr>
<td>Model B</td>
<td>0.83</td>
<td>0.81</td>
<td>0.82</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Results for Length of Stay Prediction (Long vs. Short Stays)
Performance vs. Explainability

Image Source: Easy Solutions
• Performance vs. Explanation vs. Risk
• A CNN is trained to recognize objects in images
• A language generating RNN is trained to translate features of the CNN into words and captions
Hendricks et al created a system to generate explanations of bird classifications. The system learns to:

- Classify bird species with 85% accuracy
- Associate *image descriptions* (discriminative features of the image) with *class definitions* (image-independent discriminative features of the class)

**Limitations**
- Limited (indirect at best) explanation of internal logic
- Limited utility for understanding classification errors
Explainability and Adversarial ML

(Szegedy et al., 2013, Ghorbani 2017)
Fidelity
The expectation that the explanation and the predictive model align well with one another

Risk of Readmission
Predict if the patient will be readmitted within a particular span in time
• Dr. Marcos examines some explanations for risk of readmission prediction and discovers that the explanations do not appear to be correct
• After ruling out non-determinism and thus lack of consistency as an explanation, Katherine examines the data
Data Provenance

- The Explanation is going to be as good as the data
- Caruana’s Mortality Prediction Example
- Low Quality data
  - Instrumentation problems
  - Censoring
- Incorrect explanations may be given because of problems in the data
• Fidelity with the underlying phenomenon

• Readmission model which uses lunar cycles as a feature, the model may have good predictive power and the feature may even be quite helpful but the model does not have fidelity with the underlying phenomenon
• An explanation is **Sound** if it adheres to how the model actually works
• An Explanation is **Complete** if it encompasses the complete extent of the model
• Soundness and Completeness are relative
• The soundness of an explanation can also vary depending upon on the constraints on the explanations
• Ante-Hoc models are perfectly sound by definition, Post-Hoc models can have varying leveling of soundness
• Fidelity with the prediction model i.e., explanations should align as close to the predictive model as possible (Soundness)
• Ante-Hoc: Models where the predictive model and the explanation model is the same
• Post-Hoc Models: Models where the predictive model and the explanation model are different
• Special Case: Mimic Models
Mimic Models

- Also known as Shadow, Surrogate or Student Models
- Use the output (instead of the true labels) from the complex model and the training data to train an model which is explainable
- The performance of the student model is usually quite good
- Example: Given a highly accurate SVM, train a decision tree on the predicted label of the SVM and the original data

(Tan 2017)
Mimic Models for Deep Learning
• For each decision that a tree (or a forest) makes there is a path (or paths) from the root of the tree to the leaf
• The path consists of a series of decisions, corresponding to a particular feature, each of which contribute to the final predictions
• Variable importance can be computed as the contribution of each node for a particular decision for all the decision trees in the random forest

(Saabas 2014)
• Black Box Explanation through Transparent Approximations
• BETA learns a compact two-level decision set in which each rule explains part of the model behavior unambiguously
• Novel objective function so that the learning process is optimized for:
  • **High Fidelity** (high agreement between explanation and the model)
  • **Low Unambiguity** (little overlaps between decision rules in the explanation)
  • **High Interpretability** (the explanation decision set is lightweight and small)

(Lakkaraju, Bach & Leskovec, 2016)
<table>
<thead>
<tr>
<th>Technique</th>
<th>Composition</th>
<th>Performance*</th>
<th>Model Fidelity</th>
<th>Model Specificity</th>
<th>Explanation Type</th>
<th>Scalable**</th>
<th>Scope</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayesian Rule List</td>
<td>Ante-Hoc</td>
<td>H</td>
<td>Yes</td>
<td>Self</td>
<td>Relative Importance</td>
<td>Yes</td>
<td>Global</td>
</tr>
<tr>
<td>BETA**</td>
<td>Ante-Hoc</td>
<td>H</td>
<td>Yes</td>
<td>Self</td>
<td>Relative Importance</td>
<td>Yes</td>
<td>Global</td>
</tr>
<tr>
<td>Decision Trees</td>
<td>Ante-Hoc</td>
<td>M</td>
<td>Yes</td>
<td>Self</td>
<td>Rules</td>
<td>Yes</td>
<td>Global</td>
</tr>
<tr>
<td>Falling Rule Lists</td>
<td>Ante-Hoc</td>
<td>H</td>
<td>Yes</td>
<td>Self</td>
<td>Relative Importance</td>
<td>Yes</td>
<td>Global</td>
</tr>
<tr>
<td>GAM</td>
<td>Ante-Hoc</td>
<td>L</td>
<td>Yes</td>
<td>Self</td>
<td>Relative Importance</td>
<td>Yes</td>
<td>Global</td>
</tr>
<tr>
<td>GA2M</td>
<td>Ante-Hoc</td>
<td>M</td>
<td>Yes</td>
<td>Self</td>
<td>Graphs</td>
<td>Yes</td>
<td>Global</td>
</tr>
<tr>
<td>ICE Plots</td>
<td>Post-Hoc</td>
<td>N/A</td>
<td>No</td>
<td>Agnostic</td>
<td>Graphs</td>
<td>Yes</td>
<td>Global</td>
</tr>
<tr>
<td>Interpretable Decision Sets</td>
<td>Ante-Hoc</td>
<td>H</td>
<td>Yes</td>
<td>Self</td>
<td>Relative Importance</td>
<td>Yes</td>
<td>Global</td>
</tr>
<tr>
<td>k-LIME</td>
<td>Post-Hoc</td>
<td>N/A</td>
<td>No</td>
<td>Agnostic</td>
<td>Relative Importance</td>
<td>Yes</td>
<td>Local</td>
</tr>
<tr>
<td>LIME***</td>
<td>Post-Hoc</td>
<td>N/A</td>
<td>No</td>
<td>Agnostic</td>
<td>Relative Importance</td>
<td>Data size</td>
<td>Local</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>Ante-Hoc</td>
<td>M</td>
<td>Yes</td>
<td>Self</td>
<td>Relative Importance</td>
<td>Yes</td>
<td>Global</td>
</tr>
<tr>
<td>Model Distillation</td>
<td>Post-Hoc</td>
<td>M-H</td>
<td>Yes</td>
<td>Agnostic</td>
<td>Any</td>
<td>Yes</td>
<td>Global</td>
</tr>
<tr>
<td>Partial Dependence Plots</td>
<td>Post-Hoc</td>
<td>N/A</td>
<td>No</td>
<td>Agnostic</td>
<td>Graphs</td>
<td>Yes</td>
<td>Global</td>
</tr>
<tr>
<td>RF Explainer</td>
<td>Post-Hoc</td>
<td>H</td>
<td>No</td>
<td>Random Forest</td>
<td>Relative Importance</td>
<td>Yes</td>
<td>Local</td>
</tr>
<tr>
<td>Relative Baseline Contributions***</td>
<td>Post-Hoc</td>
<td>N/A</td>
<td>No</td>
<td>Agnostic</td>
<td>Relative Importance</td>
<td>Yes</td>
<td>Local</td>
</tr>
<tr>
<td>Right for the Right Reasons ^</td>
<td>Post-Hoc</td>
<td>H</td>
<td>No</td>
<td>Agnostic</td>
<td>Relative Importance</td>
<td>Yes</td>
<td>Local</td>
</tr>
<tr>
<td>Shapley Values</td>
<td>Post-Hoc</td>
<td>N/A</td>
<td>No</td>
<td>Agnostic</td>
<td>Graphs</td>
<td>Number of Features</td>
<td>Local</td>
</tr>
<tr>
<td>SLIM</td>
<td>Ante-Hoc</td>
<td>H</td>
<td>Yes</td>
<td>Self</td>
<td>Relative Importance</td>
<td>Yes</td>
<td>Global</td>
</tr>
<tr>
<td>XGB Explainer</td>
<td>Post-Hoc</td>
<td>H</td>
<td>No</td>
<td>XG Boost</td>
<td>Relative Importance</td>
<td>Yes</td>
<td>Local</td>
</tr>
</tbody>
</table>
Ahmad 2018  Muhammad Aurangzeb Ahmad, Carly Eckert, Greg McKelvey, Kiyana Zolfagar, Anam Zahid, Ankur Teredesai. Death vs. Data Science: Predicting End of Life IAII February 2-6, 2018


References


Gunning 2016 David Gunning Explainable Artificial Intelligence (XAI) DARPA/I2O 2016


Morstatter 2016 Fred Morstatter and Huan Liu Measuring Topic Interpretability with Crowdsourcing KDD Nuggets November 2016

Nott 2017 Nott, George "Google’s research chief questions value of 'Explainable AI" Computer World 23 June, 2017


Ribeiro 2016b Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin Introduction to Local Interpretable Model-Agnostic Explanations (LIME) August 12, 2016 Oriely Media

Ross 2018 Ross, Casey IBM’s Watson supercomputer recommended ‘unsafe and incorrect’ cancer treatments, internal documents show Stat News July 25, 2018


