Towards Interpretable and Explainable AI

The LNM Institute of Information Technology Jaipur

April 18, 2024

Supervisor : Dr Lal Upendra Pratap Singh

Team Members

- **Snehil Singh Solanki - 21UCC097**
- **Virender Kumar - 21UEC144**
- **Suman Kumar Singh - 21UCC099**

LNMIIT
The LNM Institute of Information Technology

Introduction

- Interpretable Machine Learning (IML) refers to the ability of machine learning models to provide explanations or justifications for their predictions or decisions in a human-understandable manner.
- It aims to balance the accuracy of complex models with the need for transparency and interpretability.
- IML enables users to trust, understand, and potentially act upon model outputs.

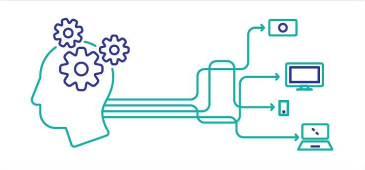
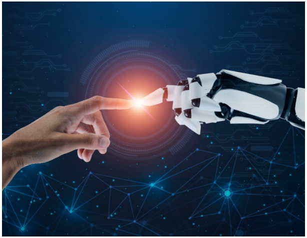


Figure: Explainable AI/Interpretability

https://christophm.github.io/interpretable-ml-book/

Motivation

- Machine learning systems' efficiency and accuracy drive their widespread adoption.
- Complex Neural Network (NN) and Deep Learning (DL) models, like NASNet, often lack interpretability.
- Interpretable machine learning addresses black-box model challenges.
- It harmonizes contradictions, supports emerging fields, and enhances model credibility, crucial in high-stakes domains like Autonomous Vehicles and Medical AI.



https://drive.google.com

Mathematical Modeling

- The easiest way to achieve interpretability is to use only a subset of algorithms that create interpretable models. Linear regression, logistic regression and the decision tree are commonly used interpretable models.
- linear regression, logistic regression, other linear regression extensions, decision trees, decision rules and the RuleFit algorithm in more detail. It also lists other interpretable models.

Algorithm	Linear	Monotone	Interaction	Task
Linear regression	Yes	Yes	No	regr
Logistic regression	No	Yes	No	class
Decision trees	No	Some	Yes	class,regr
RuleFit	Yes	No	Yes	class,regr
Naive Bayes	No	Yes	No	class
k-nearest neighbors	No	No	No	class,regr

https://christophm.github.io/interpretable-ml-book

Mathematical Modeling

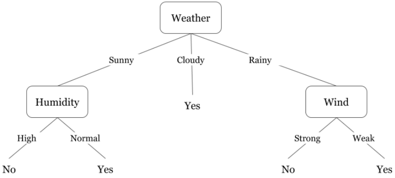
Self Explanatory Model

Decision Tree

- Decision trees are effective for capturing nonlinear relationships and feature interactions.
- They split data recursively to minimize variance or the Gini index.
- Feature importance is determined by how much each feature reduces impurity.
- Decision trees provide intuitive interpretation through visualization and individual prediction explanations.

$$\hat{y} = \hat{f}(x) = \sum_{m=1}^{M} c_m I\{x \in R_m\}$$

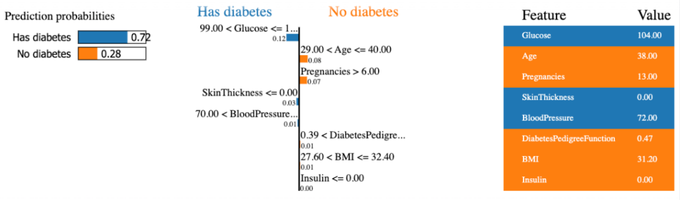
$$\hat{f}(x) = \bar{y} + \sum_{d=1}^{D} \text{split.contrib(d,x)} = \bar{y} + \sum_{j=1}^{p} \text{feat.contrib(j,x)}$$



Mathematical Modeling

LIME

- LIME (Local Interpretable Model-agnostic Explanations) explains black-box ML models.
- It fits interpretable models locally around specific examples.
- Aim: Making models easy to understand and faithful to original outputs.



- However, LIME is just one possible way to solve for feature attribution scores, and not necessarily the best way.

Mathematical Modeling

SHAP

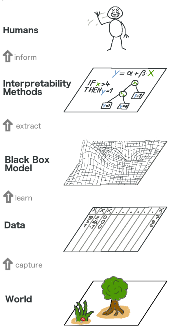
- SHAP(Shapley Additive Explanations) addresses inconsistencies in feature attribution.
- LIME may violate local accuracy and consistency.
- Shapley values ensure local accuracy, missingness, and consistency.
- Originating from game theory, Shapley values average marginal contributions of features.

$$\begin{aligned}
\Omega(g) &= 0, \\
\pi_{x'}(z') &= \frac{(M-1)}{(M \textit{ choose } |z'|)|z'|(M-|z'|)}, \\
L(f, g, \pi_{x'}) &= \sum_{z' \in Z} \left[f(h_x(z')) - g(z')\right]^2 \pi_{x'}(z'),
\end{aligned}$$

Literature Survey

Interpretable ML Methods

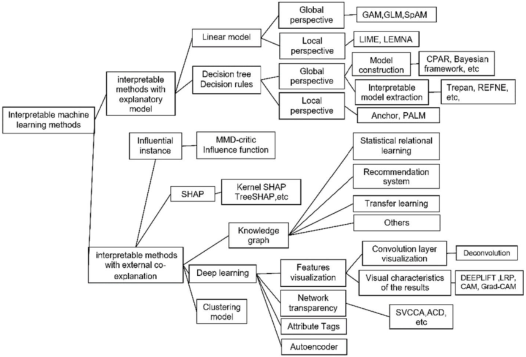
- Interpretable ML features linear regression and decision trees for transparency.
- New methods: example-based explanations, SHAP, KG-based interpretability, and DL model exploration.
- Model-specific methods limit flexibility and hinder switching.
- Model-agnostic systems offer flexibility in model, explanation, and representation.



Literature Survey

Global perspective

- Linear models provide global and local interpretability in ML. Techniques like SpAM and tree additive models enhance global interpretability.
- Local methods like LIME and LEMNA explain individual predictions. Local approximation offers better accuracy, often requiring additional methods for boundary determination.
- Decision tree/rule-based methods offer interpretable models globally and locally. CPAR and unsupervised binary trees create interpretable structures.



Literature Survey

Interpretable Methods with External Co-explaination(SHAP)

- Co-explanation methods, like SHAP and KG-based approaches, enhance interpretability. SHAP techniques, including KernelSHAP and TreeSHAP, offer individual prediction insights.
- KG-based explanations improve interpretability across various domains, including statistical relational learning and recommendation systems.

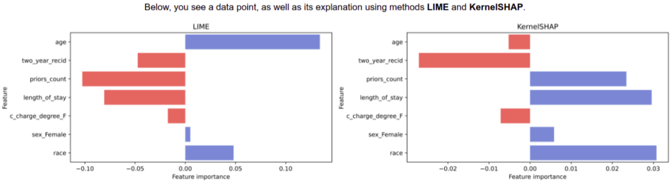


Figure: Explanation methods may disagree

Literature Survey

Feature Visualization

- Visualization:* Deconvolution and Grad-CAM highlight CNN features and decision-making factors.
- Transparency:* SVCCA dissects networks, while attribute tags offer human-friendly insights.
- Autoencoders:* Combined with SHAP, they improve anomaly detection and model understanding.

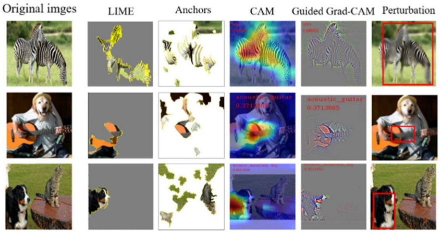


FIGURE 4. Illustration of interpretations given by different methods under three different interpretation ideas for the model explanation. Each row represents an explanation idea. The left column is the selected original images under three ideas and for other columns, each column represents an explainable method. The disturbing parts based on the meaningful perturbation method are marked with red boxes.

Implementation

- LIME Implementation
https://colab.research.google.com/drive/1bHOwywRnCF2QA25HzPjLKokgwy-7HbPA?usp=sharing

- Decision Tree
https://colab.research.google.com/drive/1JFJ3K6xRZitGIa4gykjjDkjgRLF1fKb1?usp=sharing

Requirements

Hardware and Software

Software Used:

- Google Colab notebook for machine learning
- Python language implementation;
- Relevant libraries like LIME, SHAP, sklearn, Numpy, Matplotlib, Keras, Pandas, Pytorch will be used

Hardware Used:

- - CPU:* AMD Ryzen 5 4000 series
- *RAM:* 8 GB type: DDR4 -
- *GPU :* AMD Radeon Graphics
- *Storage:* 512 GB SSD type
- Colab Configurations:- Latest Version

Towards Interpretable and Explainable AI

Team Members

Introduction

Motivation

Mathematical Modeling

Literature Survey

Hardware and Software Requirements

References

References

Published Research Paper

- Review study of Interpretation methods for Future Interpretable Machine Learning - Jian Xun Mi(member IEEE) AN-Di Li and LI-FANG ZHOU (Date of current version November 2nd 2020)
- Explainable AI - A review of Machine Learning interpretability methods - Panteilis Linardatos , Vasilis Papastefanopoulos and Sotiris Kotsianis **published** December 25th 2020
- Interpretable Machine Learning A Guide for Making Black Box Models Explainable Christoph Molnar 2023-08-21 **Chritoph Molner Link**
- 6 – Interpretability AUTHORS Audrey Huang, Jeffrey Li and Naveen Shankar AFFILIATIONS MLD, CMU PUBLISHED August 31, 2020 **MLD CMU Link**
- Explainable ML versus Interpretable ML Posted on October 30, 2018 2:49 PM by Keith O'Rourke **statmodeling official link**

Towards Interpretable and Explainable AI

References

- Definitions, methods, and applications in interpretable machine learning W. James Murdoch, Chandan Singh, Karl Kumbier, +1, and Bin Yu binyu@berkeley.eduAuthors Info Affiliations Contributed by Bin Yu, July 1, 2019 (sent for review January 16, 2019; reviewed by Rich Caruana and Giles Hooker) October 16, 2019 116 (44) 22071-22080 **PNAS Link**
- The basics of ML model interpretability Albert Einstein once famously said, "If you can't explain it simply, you don't understand it well enough." Alon Lev Alon Lev Co-Founder CEO at Qwak June 13, 2022 **QWAK LINK**
- ML Model Interpretation Tools: What, Why, and How to Interpret Author image Abhishek Jha 11 min 18th August, 2023 **Neptune LInk**
- Interpretability in Machine Learning: An Overview 21.NOV.2020 . 17 MIN READ **The gradient official link**

Thank You

3

Thank You
For Your Attention

Agency Proposal Presentation Template