



Explainable AI

By
Pavan Aditya
Shijie Zhou
Ziyuan Xu
Tianqi Yuan
Abhishek Walia



What is Explainable AI?

- Explainable AI provides methods and techniques to produce explanations about the used AI and the decisions made by it, consequently, helps attaining human trust.
- It aims to address how black box decisions of AI systems are made and answers different questions like Why did the AI system make a specific prediction or decision? Why didn't the AI system do something else? When did the AI system succeed and when did it fail?
- One of the common way to gain explainability in AI is to use Machine Learning Algorithms that are explainable.

Explainable AI is the Future Why?

The goal of explainable AI

Today



Training
Data



Learning
Process



Learned
Function



Output



User with
a Task

Tomorrow



Training
Data



New Learning
Process



Explainable
Model



Explainable Interface



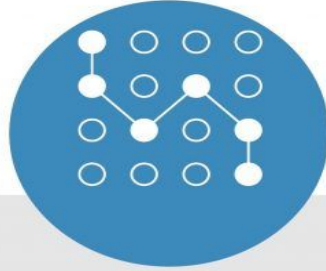
User with
a Task

Explainable AI - A Data Driven Approach



Explainable Data

What data was used to train the model and why?



Explainable Predictions

What features and weights were used for this particular prediction?



Explainable Algorithms

What are the individual layers and the thresholds used for a prediction?

Questions around AI explainability help us understand how data, predictions and algorithms influence decisions.



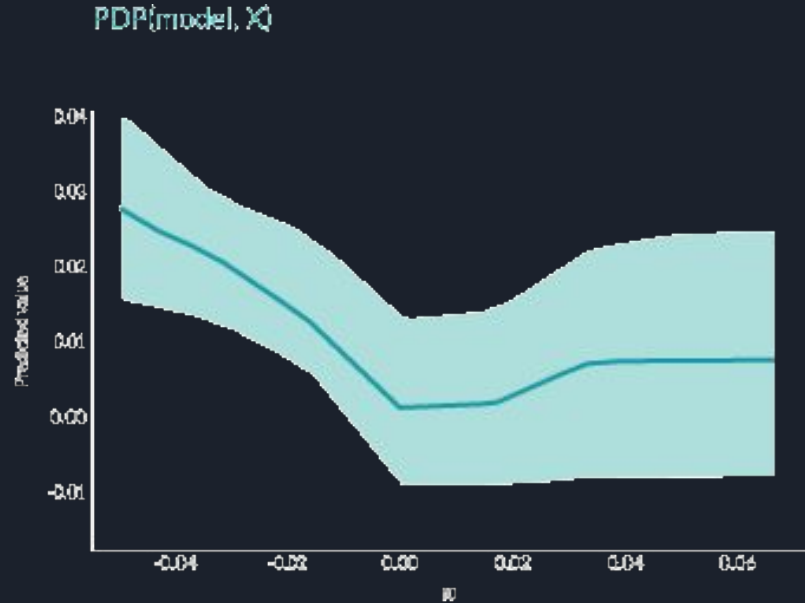
Interpretable Machine Learning Methods

- Interpretability is a rapidly growing area in machine learning.
- There have been several works in machine learning that have studied various aspects of interpretation (sometimes under the heading of explainable AI).
- What is commonly referred to as interpretation occurs primarily in the problem during the modeling and post-analysis phases. The problem, the data, and the audience provide the context needed to select an appropriate method

Interpretable Machine Learning Methods

- Partial Dependence Plot (PDP)

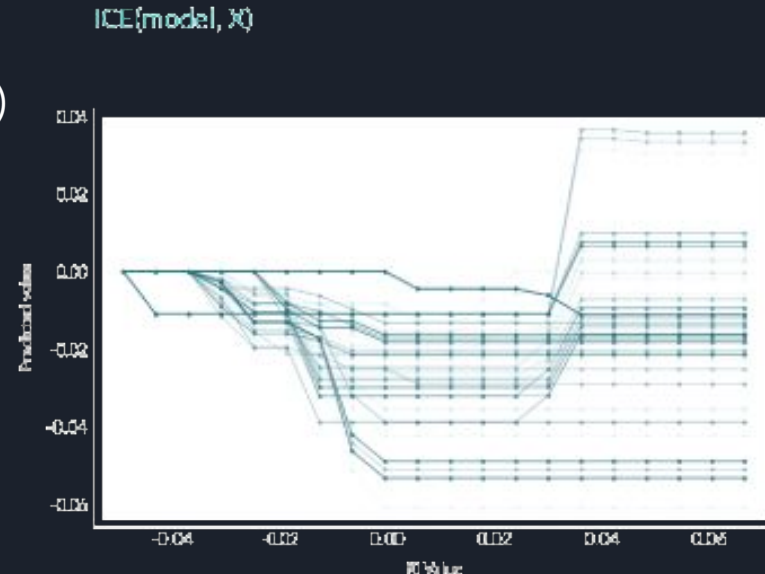
It helps researchers determine what happens to model predictions as various features are adjusted.



Interpretable Machine Learning Methods

- Individual Conditional Expectation (ICE)

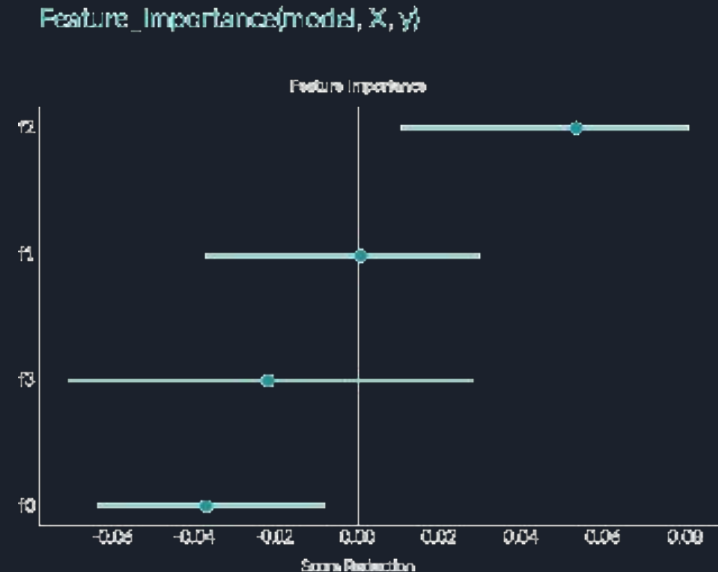
This method is more intuitive than PDP because each line represents the predictions for one instance if one varies the feature of interest.



Interpretable Machine Learning Methods

- Permuted Feature Importance

The importance of a feature is the increase in the model prediction error after the feature's values are shuffled.



Interpretable Method and Model-Agnostic Method

Is the model interpretable by design?

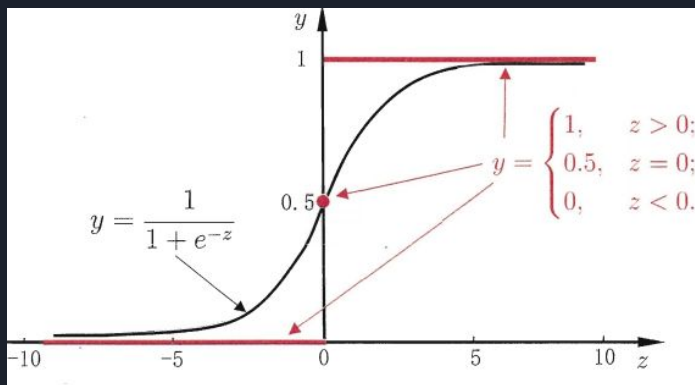
YES

Interpretable Models: KNN, Linear Models, Logistic Regression

e.g. Logistic Regression

$P \geq 0.5$, class = 1

$P < 0.5$, class = 0



They are complex for human to understand.

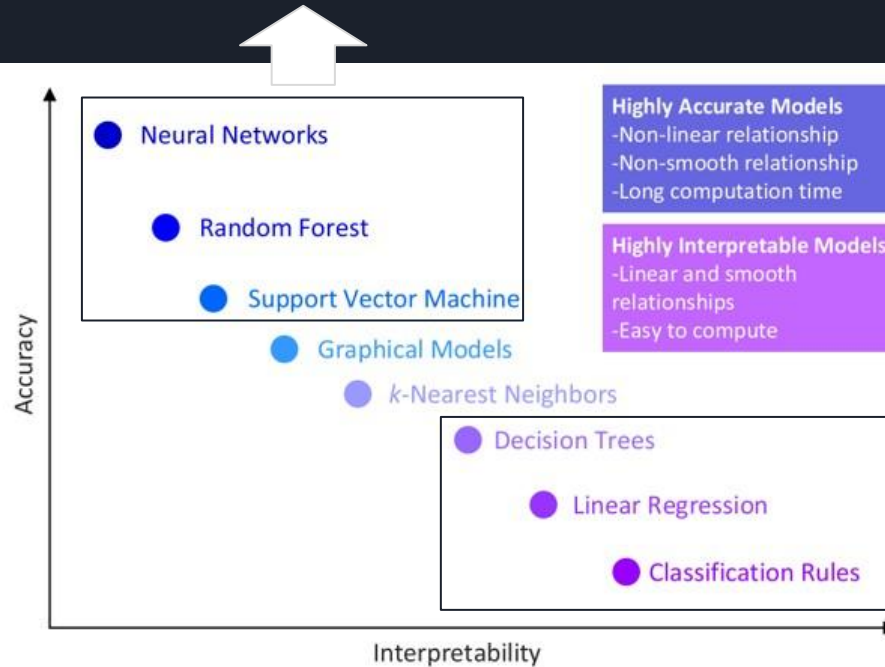



FIGURE 11. The trade-off between interpretability and accuracy of some relevant ML models. Highly interpretable algorithms such as classification rules, or linear regression, are often inaccurate. Very accurate DNNs are a classic example of black boxes.

These models are sometimes called glass box models as we can directly look into the model and understand what is going on.



Is the model
interpretable by
design?

NO


Need a method
that works on all
models?

Model-Agnostic
Methods: SHAP,
LIME

YES

NO







Model-Specific
Methods:
Integrated
Gradients

- 
- Model-Agnostic means the XAI algorithm can be applied on any kind of model.
 - Model-Specific means the method was designed for a specific type of ML model.
 - Two of the most popular Model-Agnostic methods: LIME(Local Interpretable Model-Agnostic Explanations) and SHAP(SHapley Additive exPlanations).
 - Desirable aspects of Model-Agnostic methods: **Model flexibility, Explanation flexibility, Representation flexibility.**

Source: <https://christophm.github.io/interpretable-ml-book/agnostic.html>

Integrated Gradient

Integrated Gradient(IG) is an interpretability or explainability technique for deep neural networks which visualizes its input feature importance that contributes to the model's prediction.

Original image	Top label and score	Integrated gradients	Gradients at image
	Top label: reflex camera Score: 0.993755		
	Top label: fireboat Score: 0.999961		



What is interpretability

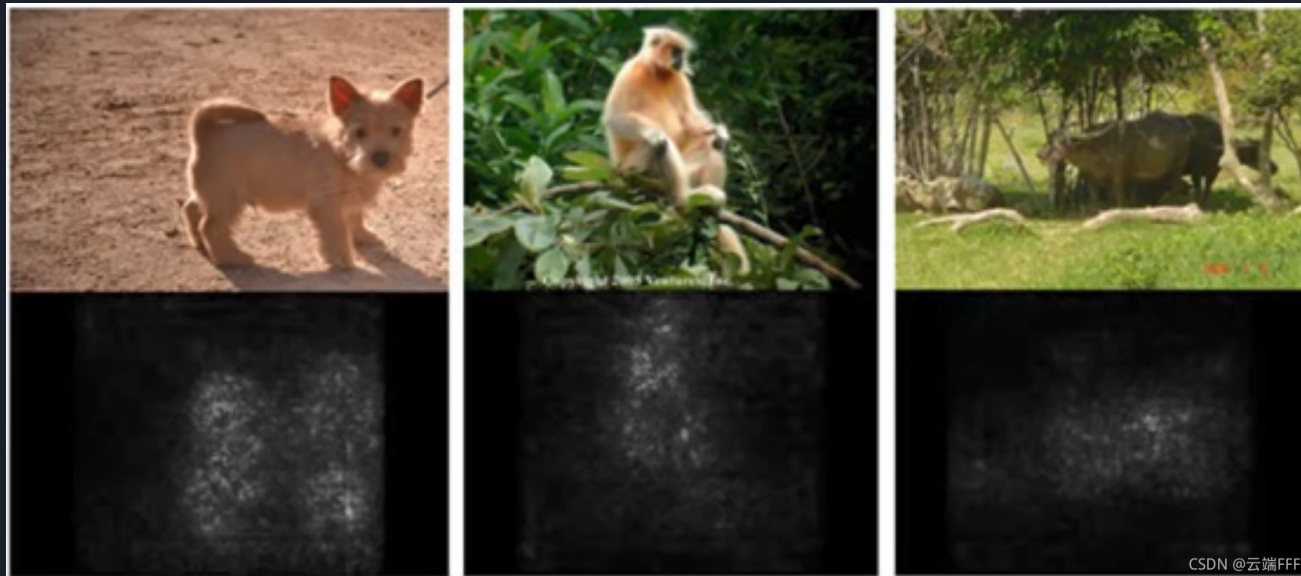
In general conceptual terms, interpretability is the reason why we want to know that the model gives a prediction $\hat{y} = f(x)$ for the unseen sample x .

Interpretability can be generally divided into "ex ante interpretability" and "ex post interpretability".

Among them, ex post interpretability is divided into "global interpretability" and "local interpretability".

Interpretability

Let's first explain the idea of "sensitivity interpretation". We know that each sample consists of multiple features, denoted as $x = \{x_1, x_2, \dots, x_n\}$, where the features are selected in a variety of ways.



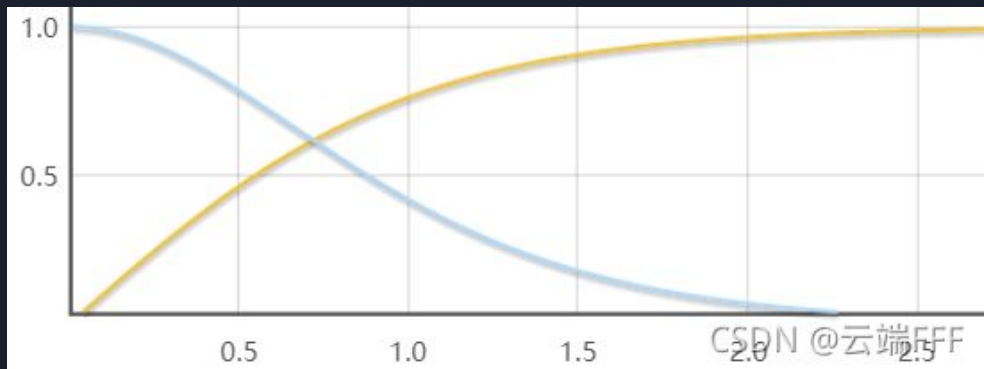


Why we need IG

Gradient saturation is an important issue in the gradient based approach described above. Suppose we want to read whether a certain picture is an elephant picture or not, obviously the trunk length is an important indicator. However, for an elephant that already has a long trunk, increasing or decreasing the trunk length a little is not very meaningful for predicting the result.

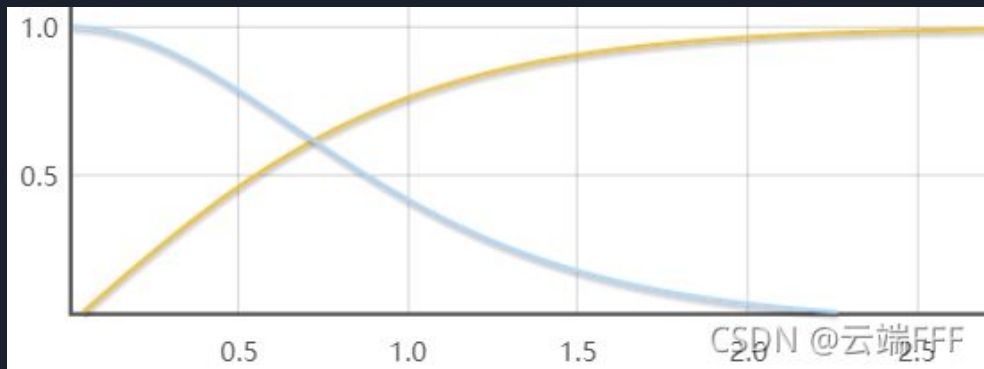
Why we need IG

The vertical axis is the probability of the model classifying the image as an elephant $f(x)$, and the horizontal axis is the elephant trunk length. The yellow line represents the increasing probability of classification as an elephant as the trunk length increases and converges to 1; the blue line is the gradient of the yellow line $\partial f(x) / \partial x_i$. As can be seen, when the yellow line is in the saturation zone, its corresponding gradient is close to 0. If we only look at the gradient, this case will lead to the wrong conclusion that the trunk length is irrelevant



Why we need IG

The idea of the integral gradient method is very simple, since the nose is too long when the gradient saturated, then I will start from the current length of the shortening, each shortening a little to find a gradient, until the shortening to a certain minimum value called baseline (to ensure that in the non-saturated area, here set to the nose length of 0), and finally add up all the gradient on it!





Integrated Gradient

IG can be applied to any differentiable model like image, text, or structured data.

Understanding feature importance by extracting rules from the network

Debugging deep learning models performance

Identifying data skew by understanding the important features contributing to the prediction



Local Interpretable Model agnostic explanations (LIME)

- It is a method in explainable AI that uses local linear models of machine learning outcomes to explain parts of complex machine-learned response functions.
- LIME is used around an area of interest. Therefore it explains parts of an ML models behaviour and is a local interpretability model rather than a global one.
- LIME is very powerful when it focuses on the regions of the model that exhibit linear behaviour, yet it can fail in areas of non-linearity.
- LIME is model agnostic i.e. it can be used to explain all classifiers, including complex models like deep neural networks and random forest.
- It can reveal the linear trends in AI Models and boosts AI 's acceptance, especially when explanations in a region match human domain knowledge and common sense.

Applications of Explainable AI

1. Healthcare



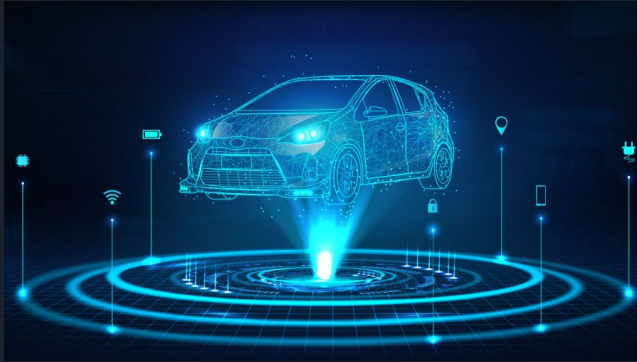
- The potential benefits of AI in the field of healthcare are high and the risk associated with an untrustworthy AI system is even higher.
- It goes without saying that the decisions made by AI models to assist doctors in classifying critical diseases, using structured parameters or unstructured data like medical imaging, have far-reaching consequences.

2. BFSI



- Explainable AI systems that are capable of providing superlative results along with comprehensible explanations, would build enough trust as well as satisfy the regulatory requirements which can lead to better adoption of AI solutions in the industry.

3. Automobiles



- Autonomous driving has been an evolving theme and is the future of the automobile industry. Driverless cars or self-driving cars are fascinating, as long as there is no wrong move made.
- One wrong move can cost one or more lives in this high-stake application of AI. Explainability is the key to understand the capabilities and limitations of the system before deployment.

4. Judicial system



- There is increasing adoption of AI systems in decision-making in the judicial process in western countries.
- The bias in AI applications, like granting parole based on the probability of repeat offense, has far-reaching consequences, and fairness in them is a must because it deals with the rights and liberties of an individual.



Conclusion

- To Conclude XAI is an important research area within the AI community. Explanations of AI models have the potential to make our AI systems more trustworthy, compliant, effective, fair, and robust, and that could drive adoption and business value.
- XAI aims at creating AI techniques resulting in more explainable models allowing humans to understand and trust the rise of Artificial Intelligence systems.



Thank You