Explainable AI (XAI)

What is it? How do we implement it?

Daria Elena Vlăduțu (236578)

What is XAI?

XAI, which stands for Explainable Artificial Intelligence, refers to methods and tools that make the outputs of AI models understandable by humans.

- **Purpose**

The main purpose of XAI is to explain in understandable terms how an AI model comes to a specific decision or prediction. This is particularly crucial for deep learning models, like neural networks, which are often described as ‘black boxes' due to their complex structures and lack of transparency in decision-making.

Accuracy vs. Interpretability

It is difficult to (mathematically) define interpretability. A (non-mathematical) definition of interpretability by Miller (2017) is: Interpretability is the degree to which a human can understand the cause of a decision. The higher the interpretability of a machine learning model, the easier it is for someone to comprehend why certain decisions or predictions have been made. A model is better interpretable than another model if its decisions are easier for a human to comprehend than decisions from the other model.

Good to know:

01. By default, machine learning models pick up biases from the training data. This can turn your machine learning models into racists that discriminate against underrepresented groups. Interpretability is a useful debugging tool for detecting bias in machine learning models.

02. The process of integrating machines and algorithms into our daily lives requires interpretability to increase social acceptance. People attribute beliefs, desires, intentions and so on to objects.

03. Accuracy is prioritised if the model has no significant impact on societal beliefs; when the problem is already well studied, interpretability is also negligible. Additionally, accuracy is preferred in situations where interpretability could lead to abusing the model/ system itself.

Use-case scenarios in media

- *Prioritizing accuracy:*

Movie Recommendation Engine on Streaming Platform

High accuracy in suggesting movies that users will enjoy would maximize engagement and platform usage. The exact internal workings of the model are less crucial because the model has no significant impact on users besides recommendations.

- *Prioritizing interpretability:*

Content Moderation on Social Media Platforms

The goal would be to accurately identify and remove harmful content like hate speech or violent threats while minimizing the removal of legitimate content. If a post is flagged for removal, the platform needs to be able to explain why, allowing users to understand the reasoning behind the decision, potentially appeal it if it's a mistake, and help maintain transparency and trust on the platform. Additionally, a simpler model can be more easily audited to identify and address potential biases in its training data.

Use-case scenarios in logistics

- Prioritizing accuracy:

Demand Forecasting in E-commerce Logistics

Predicting future demand for products with high accuracy would optimize inventory levels and minimize stockouts or excess inventory. While understanding the general trends influencing demand might be interesting, the exact internal workings of the model are less crucial. As long as the forecasts are accurate and prevent stockouts or excessive inventory, the specific factors driving the prediction might not be as important.

- Prioritizing interpretability:

Fraud Detection in Logistics Network

The model would accurately identify fraudulent transactions or activities within the logistics network, such as fake orders or shipment rerouting attempts. If a transaction is flagged as fraudulent, the logistics company needs to be able to explain why in order to investigate the flagged activity, potentially contact the customer for clarification, and ultimately make informed decisions.

Interpretability methods

Taxonomy

Intristic

Is interpretability achieved by restricting the complexity of the machine learning model? (machine learning models that are considered interpretable due to their simple structure)

Post hoc

Is interpretability achieved by applying methods that analyze the model after training ? (application of interpretation methods after model training)

Feature summary statistic:
Many interpretation methods provide summary statistics for each feature. Some methods return a single number per feature or a more complex result, such as the pairwise feature interaction strengths, which consist of a number for each feature pair.

Feature summary visualization:
Most of the feature summary statistics can also be visualized.

Model internals:
The interpretation of intrinsically interpretable models falls into this category (e.g., weights in linear models or the learned tree structure of decision trees). The lines are blurred between model internals and feature summary statistics because the weights are both model internals and summary statistics for the features at the same time. Another method that outputs model internals is the visualization of feature detectors learned in convolutional neural networks.

Interpretability methods that output model internals are by definition model-specific.

Data point:
This category includes all methods that return data points (already existent or newly created) to make a model interpretable. To be useful, interpretation methods that output new data points require that the data points themselves can be interpreted. This works well for images and texts, but is less useful for tabular data with hundreds of features.

Intrinsically interpretable model:
One solution to interpreting black box models is to approximate them (either globally or locally) with an interpretable model. The interpretable model itself is interpreted by looking at internal model parameters or feature summary statistics.

Classified by various criteria

Classified by results

Interpretation tools

Model-specific

Model-specific interpretation tools are limited to specific model classes. The interpretation of regression weights in a linear model is a model-specific interpretation, since – by definition – the interpretation of intrinsically interpretable models is always model-specific. Tools that only work for the interpretation of e.g. neural networks are model-specific.

Model-agnostic

Model-agnostic tools can be used on any machine learning model and are applied after the model has been trained (post hoc). These agnostic methods usually work by analyzing feature input and output pairs. By definition, these methods cannot have access to model internals such as weights or structural information.

Scope of interpretability

Local

Single prediction - You can zoom in on a single instance and examine what the model predicts for this input, and explain why.

Group of predictions - Model predictions for multiple instances can be explained either with global model interpretation methods or with explanations of individual instances.

Global

Holistic level - This level of interpretability is about understanding how the model makes decisions, based on a holistic view of its features and each of the learned components such as weights, other parameters, and structures.

Modular level - While global model interpretability is usually out of reach, there is a good chance of understanding at least some models on a modular level. Not all models are interpretable at a parameter level.

What about the use-cases?

Type of Method: Post-hoc, Feature Summary Statistic
How it Works: Feature importance techniques analyze the impact of individual features on the model's predictions. Imagine a model predicting movie recommendations – feature importance would tell us which features, like a user's watch history or genre preferences, contribute most to the final recommendation.

Media and Logistics Use Cases with Interpretability Priority:

- *Content Moderation (Media):* Here, interpretability is crucial. By using feature importance, we can understand which aspects of a post (e.g., specific words, hateful sentiment) most influenced the model's decision to flag it for removal. This allows the platform to explain its reasoning, potentially avoid wrongful removals, and improve user trust.
- *Fraud Detection (Logistics):* Similar to content moderation, interpretability is key. Feature importance can reveal which factors (e.g., unusual order size, billing address inconsistencies) were most critical in the model flagging a transaction as fraudulent. This allows logistics companies to investigate flagged activities more effectively, potentially contact customers for clarification, and ultimately make informed decisions.

Final reflections and future steps

Explainable AI (XAI) is essential for building trust in AI systems. Complex models are opaque, and XAI methods that rely on existing data might not translate well to unseen situations.
XAI techniques can be computationally expensive.

Ethically, XAI needs to avoid perpetuating biases from the underlying model, and legally, it raises questions about accountability and the need for clear regulations.

Despite these challenges, the future of XAI is bright. Research is exploring models inherently easier to understand and techniques like counterfactual explanations that provide a more nuanced view of model decisions. Human expertise combined with XAI can offer a more complete understanding. Standardized benchmarks will allow for better evaluation of XAI methods.

References:

01

Miller, Tim. “Explanation in artificial intelligence: Insights from the social sciences.” arXiv Preprint arXiv:1706.07269. (2017).

02

Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.). christophm.github.io/interpretable-ml-book/

03

Kelly, D. (2021, July 15). Introduction to explainable AI (ML Tech talks). YouTube. https://www.youtube.com/watch?v=6xePkn3-LME&ab_channel=TensorFlow

04

Ruiz, C., & Quaresma, M. (1970, January 1). Explainable AI for entertainment: Issues on video on Demand Platforms. SpringerLink. https://link.springer.com/chapter/10.1007/978-3-030-74614-8_87

Thank you very much!