Explainable Artificial Intelligence

CONTENTS

By,

Ankush Mulkar

AI RESEARCHER

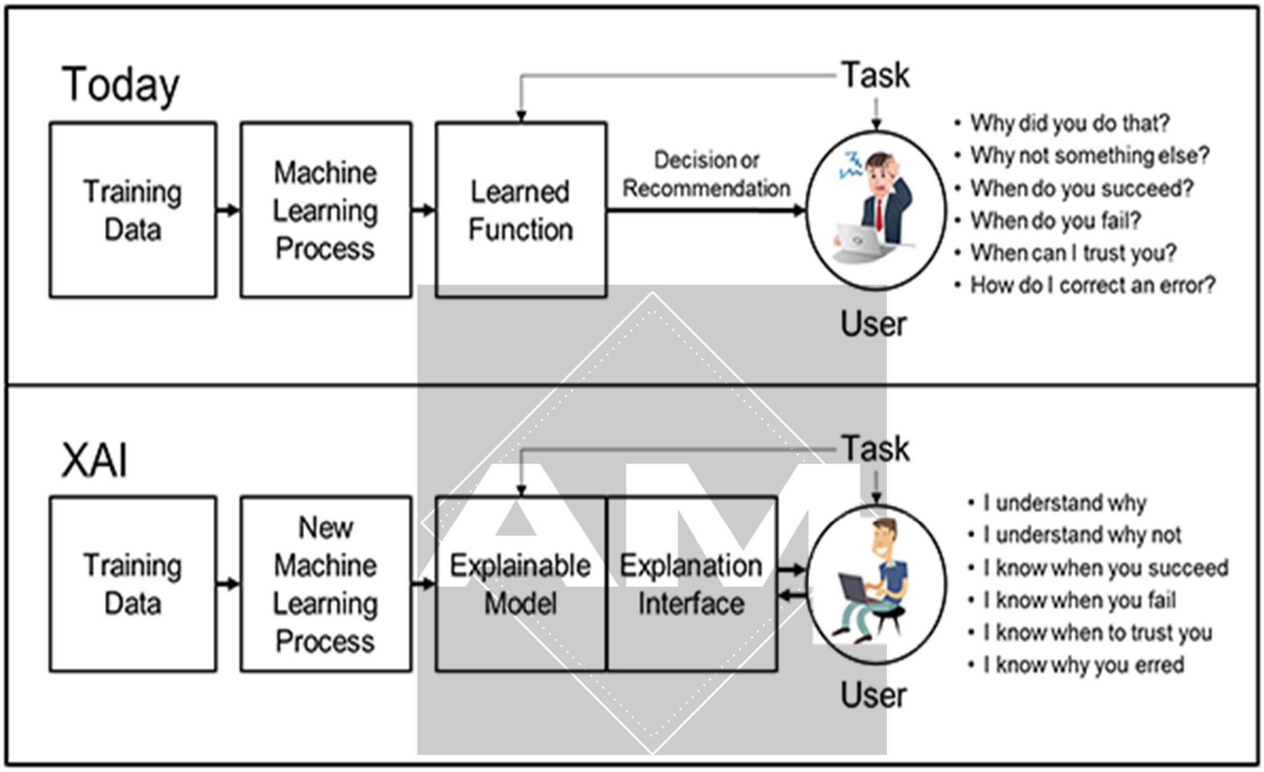
GitHub Portfolio- https://ankushmulkar.github.io/Portfolio/

LinkedIn- https://www.linkedin.com/in/ankushmulkar/

Medium- https://medium.com/@ankushmulkar

Explainable Artificial Intelligence

Explainable AI, or Interpretable AI, or Explainable Machine Learning, is artificial intelligence in which humans can understand the decisions or predictions made by the AI. It contrasts with the "black box" concept in machine learning where even its designers cannot explain why an AI arrived at a specific decision.



Types of Explainable Artificial Intelligence

Interpretable AI

Interpretable AI is one of the most important types of XAI because it explains how the AI made its decision. Many definitions are floating around for “Interpretable.” But my favourite is “A system where a user cannot only see, but also study and understand how inputs are mathematically mapped to outputs.” Following this definition, it’s trivial to see why it’s so important. If an AI fits this definition of interpretable, we can very easily build XAI to support the AI because the AI isn’t a black box. Unfortunately, this definition is not always feasible because there will be performance trade-offs. Interpretable means lower complexity models, but this is the opposite trajectory of our models. It’s easy to say that GPT-4 isn’t interpretable. However, if an Interpretable AI is sufficient for your application, I’d highly recommend it.

Transparent AI

Transparent AI is essential for building trust between people and algorithms. Transparency is another buzzword that doesn’t have an agreed-upon definition, but in practice, it refers to any methods to help answer questions of what, why, or how an AI is working. Not only do transparent methods analyse an AI, but also the data that the AI is processing.

A few typical outputs of transparent AI are summaries, visualizations, or even numerical descriptions. Each of these outputs is an attempt to translate the details of an AI to a human. By better seeing how an AI operates, we can build trust in using it. By seeing how AI processes data, users can ensure that the algorithm behaves as expected, which can help build confidence in the machine’s decisions. Whereas interpretability may not be feasible, I believe all AI should be Transparent at some level. I could be convinced otherwise, but I haven’t found a reasonable counterargument yet.

Interactable AI

Interactable AI is a type of XAI that allows users to interact with the machine learning model to understand why it made a specific decision. This type of XAI is beneficial for explaining complex models that are difficult to interpret. By seeing and interacting with the data and the machine learning model, users can better understand how the machine makes its decisions.

Interactable AI allows humans and algorithms to work together by providing feedback to each other to achieve a common goal. Humans play a vital role when working with machines because we can give context to an algorithm. Interactable AI may lead to new accomplishments previously out of reach for humans alone. Interactable AI is tricky because it is likely going to be domain-specific. For example, if you’re using an AI to generate a logo, you will have to guide the AI to create something acceptable. The type of interface you’ll need is drastically different than if you are a doctor trying to diagnose a patient. In each case, humans need to be able to insert their preference or knowledge into the decision-making of AI.

Top 5 techniques for Explainable AI

I will demonstrate explainable AI using Logistic Regression machine learning model techniques

●Explaining with Data Visualisation

●Explaining with Logistic Regression machine learning model

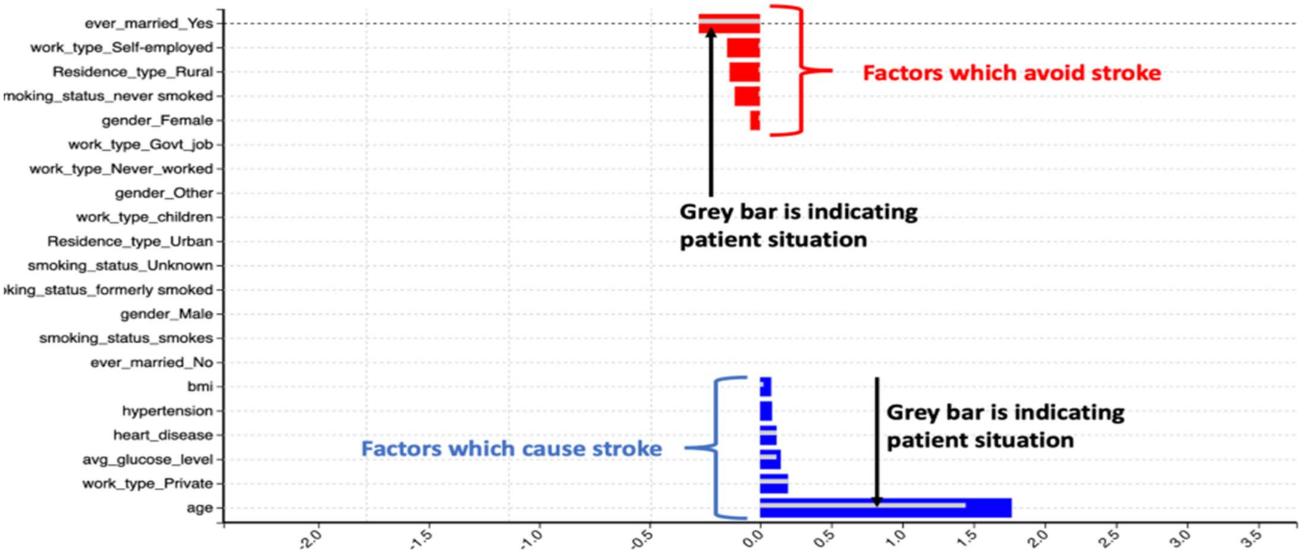
●Explaining with Decision Tree machine learning model

●Explaining with Neural Network machine learning model

●Explaining with SHAP

Logistic Regression Model

The result of a logistic regression model is shown here. On the Y-axis, you have different factors and on X-axis you see the importance of the factor for a stroke condition.



Explaining using Logistic Regression-

The blue bar indicates the factors which lead to a stroke and red shows the factors which help in avoiding a stroke. For example, age, working for a private company, high glucose levels, heart disease, and hypertension can result in a stroke. Being married, being self-employed, living in rural areas can reduce the probability of having a stroke.

The patient's situation is shown as a grey bar. We see that patient's age, his private work, glucose level, and his heart condition are factors that explain why he is at risk of stroke. His BMI (body-mass index) is low and not a reason for a potential stroke.

This approach is a more accurate way compared to the earlier radar plot technique. We are now able to precise which are the top factors leading to a potential stroke.

Logistic regression models are an excellent way to explain any predictions, as it helps us to identify which factors are more important than others. For example, here we know that patients' glucose level is more important than the patient's body-mass index

Benefits of Explainable AI

Building interpretability into AI systems can bring significant business benefits such as addressing regulation and ethical practices, and promoting innovation and competitiveness. It can also increase confidence in the AI, allowing for faster and more widespread deployment

Reducing Cost of Mistakes

Decision-sensitive fields such as Medicine, Finance, Legal, etc., are highly affected in the event of wrong predictions. Oversight over the results reduces the impact of erroneous results & identifying the root cause leading to improving the underlying model. As a result things such as AI writers become more realistic to use and trust over time.

Reducing Impact of Model biasing

AI models have shown significant evidence of bias. Examples include gender Bias for Apple Cards, Racial Bias by Autonomous Vehicles, Gender, and Racial bias by Amazon Rekognition. An explainable system can reduce the impact of such biased predictions cause by explaining decision-making criteria.

Errors can be minimized

AI models always have some extent of error with their predictions, and enabling a person who can be responsible and accountable for those errors can make the overall system more efficient

Code Confidence and Compliance

Every inference, along with its explanation, tends to increase the system's confidence. Some user-critical systems, such as Autonomous vehicles, Medical Diagnosis, the Finance sector, etc., demands high code confidence from the user for more optimal utilization.

For compliance, increasing pressure from the regulatory bodies means that companies have to adapt and implement XAI to comply with the authorities quickly.

Model performance

One of the keys to maximising performance is understanding the potential weaknesses. The better the understanding of what the models are doing and why they sometimes fail, the easier it is to improve them. Explainability is a powerful tool for detecting flaws in the model and biases in the data which builds trust for all users. It can help verifying predictions, for improving models, and for gaining new insights into the problem at hand. Detecting biases in the model or the dataset is easier when you understand what the model is doing and why it arrives at its predictions.

Informed decision making

The primary use of machine learning applications in business is automated decision making. However, often we want to use models primarily for analytical insights. For example, you could train a model to predict store sales across a large retail chain using data on location, opening hours, weather, time of year, products carried, outlet size etc. The model would allow you to predict sales across my stores on any given day of the year in a variety of weather conditions. However, by building an explainable model, it’s possible to see what the main drivers of sales are and use this information to boost revenues.