Explainable AI – The Need to Know ‘Why’

Point of View

Abstract

With rapid increase in their adoption, AI systems are increasingly being entrusted with making critical decisions. Many of these decisions might have a considerable impact on businesses and even our lives. Machine Learning (ML) is at the core of these decision systems. The evolution of deep learning has resulted in a tremendous increase in the accuracy of these decisions, but the machine learning models that these AI systems are based on are mostly “black boxes”. The human mind, however, is not comfortable at trusting a system that makes a decision without letting us into the logical reasoning behind it. And where trust is deficient, acceptance is difficult.

One of the major challenges in deploying ML models in production is the reluctance on the part of the business stakeholders to accept the decisions made by these black box machine learning models without being able to understand why those decisions were made. There is an ever-increasing need for businesses to be accountable for the consequences of the decisions made by ML powered models. There are regulatory and legal ramifications to business decisions and in the absence of an explanation, it is practically impossible to defend these decisions.

Let us consider the financial services sector where AI systems are being deployed to transact on financial instruments, assess insurance claims, assign credit scores, and optimize investment portfolios. Let us take a specific example of an AI-based credit scoring system that rates individuals for their creditworthiness. It is very likely that since many of these models are trained on large datasets, they can make good decisions. But the decisions are only as good as the data that they feed on. It is possible that the models create or reinforce bias in the decision that could be seen as discriminatory against a particular group of people, subjecting the business to risks from litigation to loss of reputation.

Explainable AI (XAI) refers to a set of tools and techniques that help us humans interpret and trust the decisions made by ML models. There are two aspects to this trust:

1) "Do I trust this specific decision, and can I go ahead performing an action based on this decision". This would help model stakeholders understand, accept and act on these decisions.
2) "Do I trust the model as a whole enough to deploy this in production". The insights provided by these tools into the model functioning can help model developers debug the model and improve its accuracy. All this results in the model being trusted and accepted when it is deployed in production.

There are several approaches to explainability, based on when in the ML model lifecycle is it required, its dependency on the type of model, and whether it explains the specific prediction or the entire model.

Explainability methods can be broadly classified based on:

- **Model Lifecycle Stage:** Pre-model, In-Model, Post-model
- **Model Dependency:** Model-specific, Model-agnostic
- **Model Predictions:** Global, Local

Pre-model versus In-Model versus Post-model

Pre-model methods give us a better understanding of the data that goes into model development. Their importance stems from the fact that the behavior of the model is largely influenced by the data used to train the model. These methods can be broadly categorized under exploratory data analysis (EDA), explainable feature engineering, and dataset summarization. One popular method in this category is principal component analysis (PCA) that simplifies model features into fewer components to help visualize patterns in your data.

Methods that help build inherently explainable models fall under the in-model explainability category. It does seem common sense that the best way to avoid black-box models is to build a model that is explainable by design. There are several methods for the in-model category, ranging from choosing from an explainable model family, incorporating explanation along with prediction to making architectural changes in deep networks.

Post-model methods provide explanations for pre-developed models. The bulk of recent research done in XAI falls under this category as methods are being explored for explainability of black-box models. Some of the common post-model methods are based on perturbation mechanisms.

Model-specific versus Model-agnostic

Model-specific methods have direct access to the internal model weights and parameters in use and are therefore based on the insights derived from them. These are mostly used to explain deep neural networks as they are increasingly in use and are more difficult to understand. For example, the Gradient-weighted Class Activation Mapping (Grad-CAM) approach is used to produce visual explanations specifically for convolutional neural networks (CNN). Model-agnostic methods are not constrained by the model architecture and are mostly used in post-model explanations. For example, LIME (Local Interpretable Model-agnostic Explanations) can be applied to any model provided we can create perturbations on the input and observe the corresponding output. In the case of object detection in machine vision using deep learning, this would mean hiding sections of the image, observing the predictions, computing the weights and fitting a linear model that is explainable.

Global vs Local

Global methods deal with the overall understanding of the models, their training, the data used for training and in general the behavior of the models. This would be useful to assess and improve the performance of models, debugging models and to gain better insights into the functioning of the models. Local methods deal with the interpretation of a specific outcome of a model. They help to explain a specific prediction or decision and which specific features and characteristics contributed towards it. Consider an example of loan approval based on an applicant's details that include income, age, number of dependents and so on.

A global method would explain the overall attribution of these features on the outcome while a local method would help explain a specific applicant's loan approval decision.

Gartner has placed Explainable AI at the peak of the Gartner's Hype Cycle for Emerging Technologies 2020. Gartner predicts that "By 2023, over 75% of large organizations will hire artificial intelligence specialists in behavior forensic, privacy and customer trust to reduce brand and reputation risk." This would mean that with the increasing adoption of AI, Explainable AI would be a necessity for businesses to maintain their brand value and reputation.

How does AI Explainability help?

Improved visibility: Model developers benefit by the visibility that explainability offers to help understand the functioning of their models and to be able to debug poor performance.

Better acceptance: With increased trust in the decisions made by AI solutions, businesses would be more willing to adopt them

Reduced risks: Business risks due to biased or poor predictions are minimized with the visibility into these decisions and correction. Also, there is reduced risk associated with legal and regulatory authorities.

References

1. The How of Explainable AI

2. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI

3. "Why Should I Trust You?": Explaining the Predictions of Any Classifier

4. Machine Learning Interpretability - MDPI - *https://www.mdpi.com/2079-9292/8/8/832/pdf*

About The Author

Jayashree Arunkumar

Jayashree is a Technical Architect working with the Incubation unit, Research and Innovation, TCS. Her focus areas are in identifying and developing new offerings leveraging automation and machine learning in delivering IT services. She holds a graduate degree in Electronics and Communications Engineering and has 23+ years of experience in IT Service Management, Network Management, Operations Automation, DevOps and more recently, MLOps.

Contact

Visit: TCS Incubation, https://www.tcs.com/tcs-incubation

Blogs: Research and Innovation, https://www.tcs.com/blogs/research-and-innovation

Email: tcs.incubation@tcs.com

About Tata Consultancy Services Ltd (TCS)

Tata Consultancy Services is an IT services, consulting and business solutions organization that has been partnering with many of the world's largest businesses in their transformation journeys for over 50 years. TCS offers a consulting-led, cognitive powered, integrated portfolio of business, technology and engineering services and solutions. This is delivered through its unique Location Independent Agile™ delivery model, recognized as a benchmark of excellence in software development.

A part of the Tata group, India's largest multinational business group, TCS has over 453,000 of the world's best-trained consultants in 46 countries. The company generated consolidated revenues of US $22 billion in the fiscal year ended March 31, 2020, and is listed on the BSE (formerly Bombay Stock Exchange) and the NSE (National Stock Exchange) in India. TCS' proactive stance on climate change and award-winning work with communities across the world have earned it a place in leading sustainability indices such as the Dow Jones Sustainability Index (DJSI), MSCI Global Sustainability Index and the FTSE4Good Emerging Index.

For more information, visit us at **www.tcs.com**

All content / information present here is the exclusive property of Tata Consultancy Services Limited (TCS). The content / information contained here is correct at the time of publishing. No material from here may be copied, modified, reproduced, republished, uploaded, transmitted, posted or distributed in any form without prior written permission from TCS. Unauthorized use of the content / information appearing here may violate copyright, trademark and other applicable laws, and could result in criminal or civil penalties.
Copyright © 2021 Tata Consultancy Services Limited