TATA CONSULTANCY SERVICES

Building on belief

The role of data efficacy and human intervention in explainable AI

Abstract

Artificial intelligence (AI) has already become a mainstream technology, growing at a compound annual rate of 39.7% during 2021-2026. Business entities are increasingly leveraging AI to drive their growth and transformation agenda in a purpose-centric way. The fact that AI augments humans and does not replace them is widely acknowledged. But as businesses adopt AI, it is imperative that the method by which the output of the AI solution is arrived at, is completely transparent and traceable to make it a trustful solution. Thus, AI solutions that work as a black-box are typically not acceptable.

In a highly regulated business environment, every business decision must be fully explained to meet compliance norms. Compliance violation can lead to severe penalties and loss of reputation and business, thereby demanding explainable AI solutions.

Decision making and AI

For AI to be explainable, we need to consider its decision-making capabilities. It can either be deterministic, based on specific and well-defined business rules, or probabilistic, driven by human ability to correlate with similar cases or case-based reasoning.

By its very nature, AI learns like we do before it gets ready to provide business solutions. With deterministic decision making, understanding and explaining the decision-making logic is straightforward since it is fully objective. Probabilistic decision making, on the other hand, is ambiguous due to its inherent subjectivity.

When AI processing is based on deep learning, it follows the black-box approach, lacking transparency, and therefore explaining the decision becomes challenging.

We explore the available options for explainable AI and delve into one such approach.

Three techniques of explainable AI

Businesses can manage AI explainability-related challenges through:

1. **Traditional techniques without deep learning**

   Techniques such as clustering and dimensionality reduction model can be leveraged, so that the performance and evaluation of the AI model can be assessed through measures such as ‘precision and recall’, ‘accuracy’, and ‘coefficient of determination’.

2. **Ante-hoc technique with partial deep learning**

   The explainability here is built into the AI model, wherein techniques such as RETAIN (reversed time attention) and BDL (Bayesian deep learning) can be deployed.

3. Post-hoc technique with deep learning

Here the explainability is built externally using a completely auditable proxy model through which the output of the black-box AI model is explained. Techniques such as LIME (local interpretable model agnostic explanation) and LRP (layer-wise relevancy propagation) can be leveraged here.

Enterprises must have thorough due diligence in place to select the most appropriate option for their specific business requirement to manage explainability mandates effectively.

Considerations around data

AI and data share a symbiotic relationship. While AI can glean valuable and unprecedented insights from enterprise data, it needs high-quality data for gainful training and learning for AI to do its job. Signal-to-noise ratio of data is another important consideration for enterprises to derive value from data and build a right and explainable AI model.

When the signal-to-noise ratio or the ratio of the predictable-to-unpredictable component of data is low, it results in the AI model:

(1) necessitating more data to validate the signal.

(2) demanding extensive involvement of humans (human-in-loop) for its validation.

The AI model built with such a low signal-to-noise ratio causes more challenges for explainability.

While human-in-loop is a manageable challenge for business entities, those linked to limited usable data sets are more complex. In such a scenario, enterprises typically look for more usable alternate data, which has many more clues associated with their business, to improve their signal-to-noise ratio.

Post-hoc method

In this method, the explainability of AI is built externally.

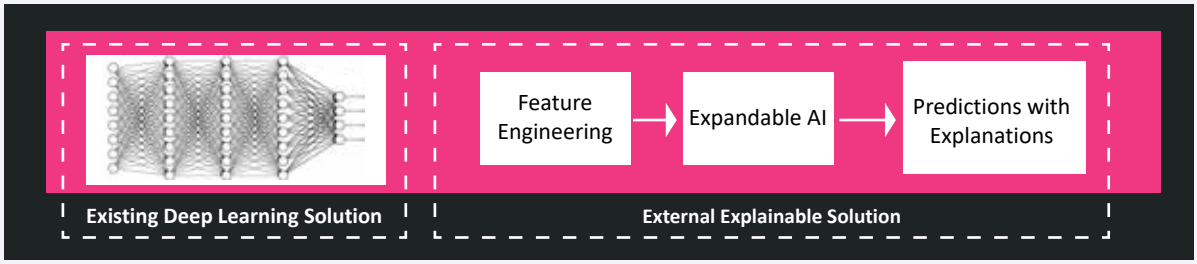

Figure 1: From black-box AI solutions to auditable proxy models

As illustrated in Figure 1, a deep learning-based AI solution or black-box AI solution already exists. To explain its output, we have to build a proxy model using completely auditable techniques.

The external explainable solution must be based on feature engineering, and should analyze the features, understand their intensity and demystify the working of the deep learning-based black-box solution.

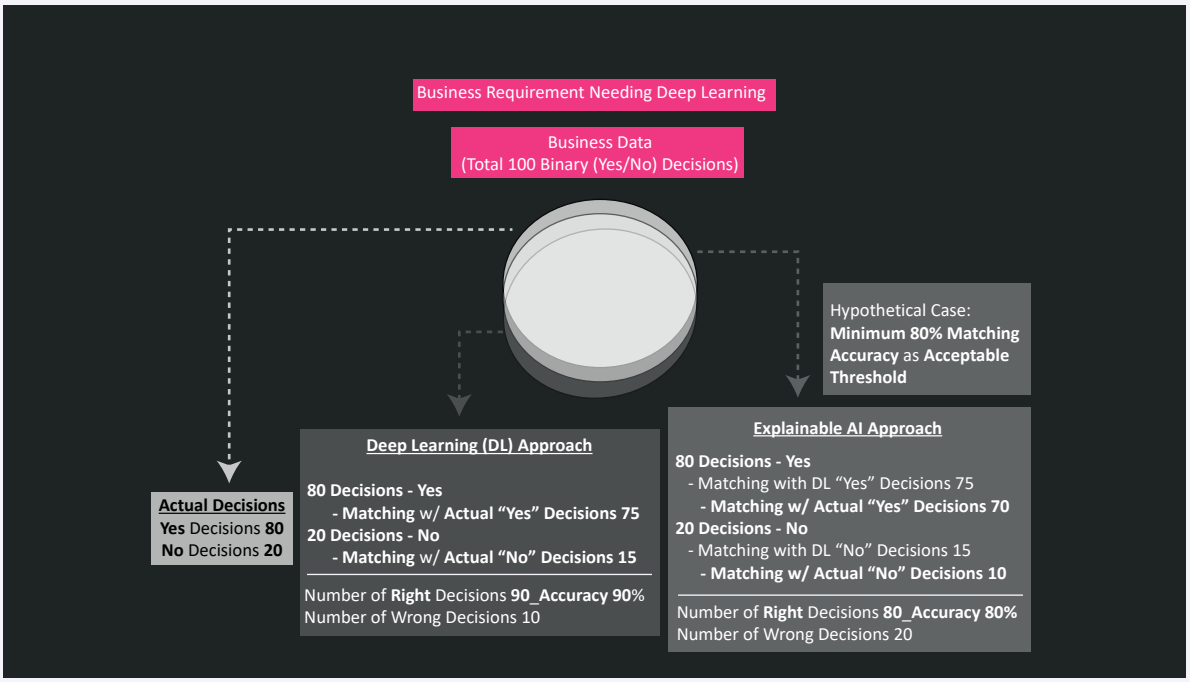

Figure 2: External explainability of AI

Consider the hypothetical case where the AI solution has to process 100 business cases, wherein for each case, the decision could be either "Yes" or "No" (see Figure 2). The actual decisions, manually taken outside the AI solution, are "Yes" for 80 cases and "No" for 20 cases.

When the same cases are processed by the AI solution, using the deep learning-based black-box approach, it too produced the decision of "Yes" for 80 cases and "No" for 20 cases.

- However, when these decisions were mapped with the actual manual decisions, only 75 of the total 80, which were "Yes", and only 15 of the total 20, which were "No", matched. Thus, the total number of matching decisions produced by the AI solution were 90 (75 + 15).
- The accuracy of the AI solution can therefore be understood as 90%.

But this AI solution, working in the black-box mode, does not explain how it arrived at the individual "Yes" and "No" decisions for each of the use cases. Therefore, this AI solution claiming to provide 90% accuracy is of very limited use to business entities.

The external explainable AI solution, where post-hoc based method was deployed, produced the decision of "Yes" for 80 cases and "No" for 20 cases.

- However, when these decisions were mapped with those produced by the black-box AI solution, only 75 of the total 80, which were "Yes", and only 15 of the total 20, which were "No", matched.
- Next, when the decisions taken by the external explainable AI solution were mapped with the actual manual decisions, only 70 of the total 80, which were "Yes", and only 10 of the total 20, which were "No", matched. Thus, the total number of matching decisions produced by the AI solution were 80 (70 + 10).

Therefore, the external explainable AI solution can be understood as 80% accurate.

Deciphering this scenario and arriving at its usability for a specific business requirement calls for human intervention to judge the efficacy, and therefore the usability, of this external explainability approach.

Explainabiity through the post-hoc method

The post-hoc approach is about improving the interpretability of the black-box AI model by having a similar AI model working transparently and defending how a particular decision was arrived at and why. Here, typically three approaches are adopted:

1. Related qualifying examples: Cases where for the same input data, the same decision was arrived at, and thus the decision arrived at is supported by similar related examples.

2. Related but non-qualifying examples: Cases where for the same input data, different decisions were arrived at. The different decisions are supported by stating that similar input values do not mean that the business entities would take the same decision as they need to take into account other factors while arriving at a decision.

3. Unrelated qualifying examples: Cases where a decision is supported by having unrelated examples, with different input data values, which qualify.

While post-hoc method is quite common in use, enterprises should carefully evaluate its efficacy at both the atomic and then the aggregate level. While it is important to vet each of the specific decisions regarding their accuracy and rationale, it is equally important to vet them at an aggregate or overall level. At an overall level, it must be validated if the post-hoc method provides the required confidence and faith for its acceptance by all stakeholders in the business entity value chain (the business/operations executives, the internal compliance executives, the end customers, and also the regulators).

Though post-hoc approach is quite common, enterprises should be mindful of the following limitations and factor them in their considerations:

1. It is an indirect explanation approach.

2. Acceptability of indirect, or derived explanation, is always low compared to a direct explanation.

3. At an aggregate level, this approach may still not be acceptable if the errors, where the decision could not be gainfully explained, are in significant numbers, that impact cannot be ignored.

4. Currently there are no benchmarks available to validate the efficacy of post-hoc approach, and establishing such a benchmark is not an easy task, given the above considerations.

Thus, while the explanation is provided by the post-hoc method, to decipher that explanation and arrive at its usability for the specific business requirement calls for human intervention.

Conclusion

Achieving explainable AI is not just an AI algorithm challenge. While AI algorithm is one of the key considerations in explainable AI, equally important considerations are around the data processed (such as the signal-to-noise ratio) and human interventions judging the soundness of the explainable AI. Business entities should carefully select the type of AI solution needed for a specific business requirement so that expectations around that AI solution that of being completely transparent and traceable to ensure it as a trustful solution, are gainfully met.

About the authors

Mahesh Kshirsagar
CTO – Analytics & Insights, TCS

Mahesh Kshirsagar is responsible for shaping innovative and purpose-led solutions that accelerate the growth and transformation agenda of enterprises. Mahesh is an engineering graduate, who started his career in TCS in 1990, and has IT expertise of 30+ years spanning technology domains, industry verticals and software processes. Over the years, he has incubated several high-impact, business-aligned IT solutions/services having high revenue potential, focusing on thought leadership and innovation to enable growth and transformation. His IT expertise stems from his employment experience in system integrator companies, end-user organizations, IT products, and BPO organizations.

Conceptualizing, architecting, and delivering state-of-the-art business IT solutions is his strength. He has applied for more than 30 patents for his solutions, of which around 10 have been granted. His solutions have also won several prestigious industry awards.

Sanjeev Manchanda
Consultant – Analytics & Insights, TCS

Sanjeev develops futuristic solutions for different industries. With more than 19 years' experience in working with large companies, he has created more than 200 solutions. Sanjeev has contributed to research in data mining, has written for 13 international journals, and has filed three patents. Sanjeev has done his MBA and MCA and also has a PhD in computer science.

Contact

Visit the Analytics and Insights page on https://www.tcs.com

Email: BusinessAndTechnologyServices.Marketing@TCS.com

About Tata Consultancy Services Ltd (TCS)

Tata Consultancy Services is a purpose-led transformation partner to many of the world's largest businesses. For more than 50 years, it has been collaborating with clients and communities to build a greater future through innovation and collective knowledge. TCS offers an integrated portfolio of cognitive powered business, technology, and engineering services and solutions. The company's 469,000 consultants in 46 countries help empower individuals, enterprises, and societies to build on belief.

Visit **www.tcs.com** and follow TCS news **@TCS_News.**

Corporate Marketing | Design Services | M | 06 | 21