KPMG

To Explain or to Interpret

Explainability of AI: the approaches, the challenges

Artificial Intelligence and Mathematical Modelling

In many domains, it is mission critical to be able to explain the decisions taken by algorithms. Explanations might be necessary in order to build trust in algorithms and drive adoption. Explanations are often a legal requirement (e.g. the right to explanation in the EU General Data Protection Regulation [1]). While the need to explain AI is pressing, there is still no reliable methodological approach that tackles this problem.

1. When is a model a black-box?

There is no widely accepted definition of black-box. In general, black-box refers to proprietary models whose code is not accessible. However, the term black-box is also used to describe non-simulatable models, i.e. models too complex for a human to go through their computations. Non-simulatable models include Random Forests and Deep Neural Networks.

2. Why are black-box models used?

Black-box models have achieved strong performance in many domains, especially for image recognition tasks and natural language processing. Behind their success there is also their ability to identify patterns in data with limited feature engineering.

3. What is xAI?

Black-boxes have been the models of choice in many academic and commercial use cases. When an explanation is required, it is common practice to apply xAI methodologies. xAI is a set of methodologies that analyze models post-hoc (after the model has been trained) in order to understand their decisions. xAI seems to address the lack of transparency of black-boxes and help (fig.1)

- understand and validate the behavior of the model
- identify edge cases and anticipate potential model failures
- gain the trust of customers and internal stakeholders

4. One model, several explanations

AI projects have multiple stakeholders: business owners, developers, regulators, users and the individuals that are ultimately affected by the models.

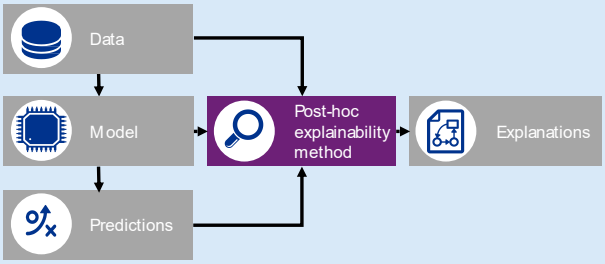

Fig. 1 – Post-hoc explainability methods

Local vs global explanations

Explanations can be local or global. Local explanations explain one single decision. Global explanations provide insights into the behavior of the model on the overall dataset. Developers and product owners are generally interested in the global behavior of the model. Users are generally interested in understanding the decision made on their specific case. A local explanation might differ from the global behavior of the model.

5. The danger of local surrogates

Local methodologies such as SHAP and LIME develop simpler models (surrogates) that offer explanations for a single data point that correspond to a specific decision. Local surrogates are inherently interpretable models: the output of surrogates provides insights that can be understood and interpreted by a human with an understanding of basic mathematics.

The adoption of local surrogates has been widespread. However, we believe that the use of local surrogates should

be avoided when the use case requires strict guarantees that explanations are faithful. A surrogate model might in fact use, exploit completely different features than the black-box model for the same predictions.

Furthermore, surrogate models might not be robust: the explanation provided by a surrogate model might be very different for very close data points [2]. Our xRAI solution tackles this challenge (see fig. 2).

6. Counterfactuals

Acknowledging the limitations of local surrogates, the use of counterfactuals has been advocated in a broad variety of regulatory contexts. Counterfactual explanations describe the smallest change to the world that can be made to obtain a desirable outcome [3].

While counterfactuals are a compelling alternative, there are significant risks associated to the use of counterfactuals, including:

- Changes may not always be actionable
- Changes in dependent features could impact the desired decision
- Feature normalization is dataset dependent and does not always align with the users' opportunity cost
- Counterfactuals could constitute an implicit recommendation in an industry where recommendations are legally prohibited

7. Alternatives to the use of black-boxes

As we have seen, explaining black-boxes is extremely challenging. Should we always resort to black-box models?

The accuracy-interpretability trade-off

It is often believed that it is necessary to sacrifice the interpretability of a model for accuracy and performance. This is a common myth in the AI industry: the existence of an accuracy-interpretability trade-off has not been proven [4]

Inherently interpretable models

The progress in inherently interpretable models has been steep: from Prototypic Neural Networks [4] to Explainable Boosting Machines [5], there are currently many inherently interpretable models that achieve state-of-the-art performance on a wide range of problems.

KPMG xRAI

KPMG xRAI evaluates the robustness of the explanations produced by a local surrogate and helps identify regions where the explanations for very close datapoints.

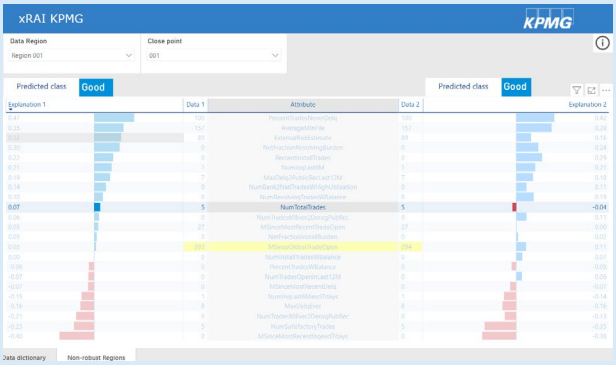

Fig. 2 – KPMG xRAI make it easy to identify regions where ex-post estimations are not robust

8. Our solution

KPMG has a long-standing experience in providing assurance on AI systems and helping clients

- Define explainability requirements
- Evaluate your current xAI approach
- Identify inherently interpretable approaches that are an alternative to the use of black-box models
- Identify the appropriate xAI approach
- Evaluate the regulatory framework around the use of AI and its explanation

References

[1] Automated individual decision-making and Profiling, GPDR, Recital 71

[2] Alvarez-Melis, David, and Tommi S. Jaakkola. "On the robustness of interpretability methods." arXiv preprint arXiv:1806.08049 (2018).

[3] Wachter, Sandra, Brent Mittelstadt, and Chris Russell. "Counterfactual explanations without opening the black box: Automated decisions and the GDPR." Harv. JL & Tech. 31 (2017): 841.

[4] Rudin, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." Nature Machine Intelligence 1.5 (2019): 206-215.

[5] Caruana, Rich, et al. "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission." Proceedings of the 21th ACM SIGKDD conference. 2015.

Contacts

KPMG AG
Räffelstrasse 28
P.O. Box
CH-8036 Zurich

kpmg.ch

Mark Meuldijk
Partner
Data & Analytics

+41 58 249 48 84
markmeuldijk@kpmg.com

Mattia Ferrini
Director
Artificial Intelligence

+41 58 249 30 51
mattiaferrini@kpmg.com