Why Explain a decision made by AI? Explainable AI (XAI)

Rodrigo Almeida

Abstract—In this paper, I will review the literature relating to transparency for decisions made by artificial intelligence (AI). As AI advances, not only our dependence on AI models are increasing but also a need for understanding how such results and decisions were made by the model. Due to the complexity of AI, Explainable Artificial Intelligence (XAI) has become a great choice, bringing more clarity, where outputs can be explained reducing bias and bringing transparency to the decision-making process, as opposed to "Black Box" algorithms that we cannot explain why or how the AI arrived in such a conclusion/classification/decision, it outputs directly from data, and not even who developed such algorithms are able to explain how the model arrived in such result.

1 Introduction

Explainable AI is emerging, and with it, the concerns from scientists, the general public and policymakers are also on the rise. As artificial intelligence, machine learning and deep learning are becoming integrated into applications, decisions from such models can carry consequences for a person and society, changing the focus from accuracy to the necessity of being able to explain and clarify how these decisions were made.[1] As we are becoming more dependent and relying hugely on intelligent machines, whether it is email filters, AI helping with medical diagnoses or self-driving cars, machine learning is being used across almost all sectors. Such growth in these technologies only highlights the importance of understanding the interpretability of the outcome generated by it. This is the main motivation for explainable artificial intelligence.[2] The impact of the outcome will vary based on the field. A wrong prediction for computing and business might lead to misleading recommendations, potentially affecting the revenue of such businesses, however wrong predictions in critical sectors like healthcare could put human life at risk. Opening the "black box" is incredibly important, not just for social acceptance but also for regulatory purposes. Understanding how AI systems reach their conclusions is crucial in ensuring accountability, transparency, and ethical use of these technologies, making it essential for both public trust and regulatory compliance.[1]

2 Future Work

3 Conclusion

Appendix A
Proof of the First Zonklar Equation

Appendix B

References

[1] P. Angelov et al. "Explainable artificial intelligence: an analytical review". In: *WIREs Data Mining and Knowledge Discovery* 11 (5 2021). DOI: 10.1002/widm.1424.

[2] Finale Doshi-Velez and Been Kim. *Towards A Rigorous Science of Interpretable Machine Learning*. 2017. arXiv: 1702.08608 [stat.ML].