

Abstract

Machine learning-based systems became widely used both in industrial and everyday life situations. As machine learning model complexity followed the rise of problem complexity in different safety- and security-critical areas, it is more of a challenge to explain the decision-making process of an AI system. As a result, new explanation techniques appeared in both industrial and academic communities. This proposal aims to conduct experiments on the most state-of-the-art XAI technologies.

Introduction

In order to call an AI system human-centered, it has to have some degree of **explainability**, **interpretability**, and **transparency**. An example of explaining a model can be seen in Fig.1.

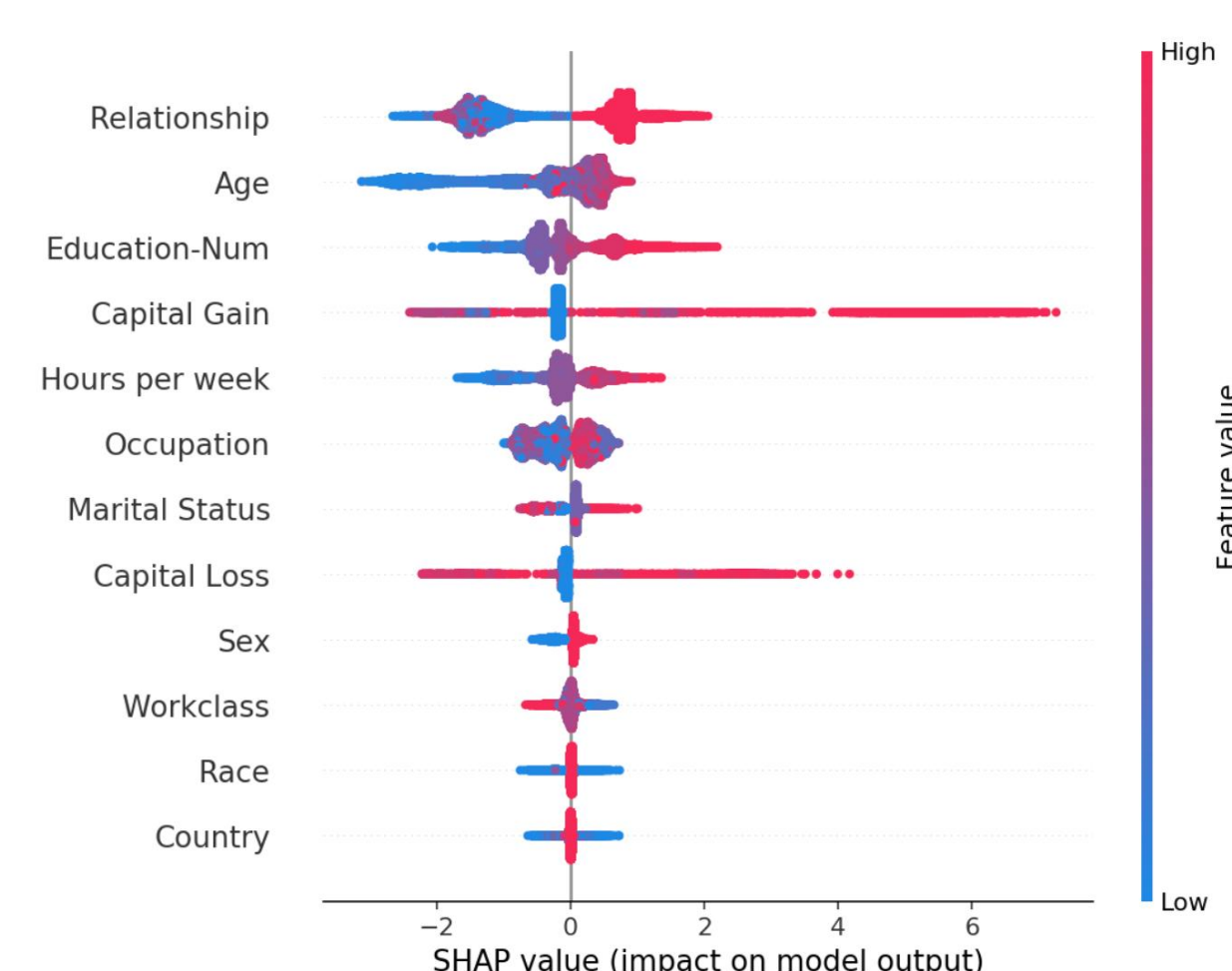


Fig.1. Example output of the SHAP python tool

Research methods

Experiments on XAI tool while considering aspects like:

- **Scope** (Local / Global)
- **Methodology** (BackProp / Perturbation)
- **Usage** (Intrinsic / Post-Hoc)

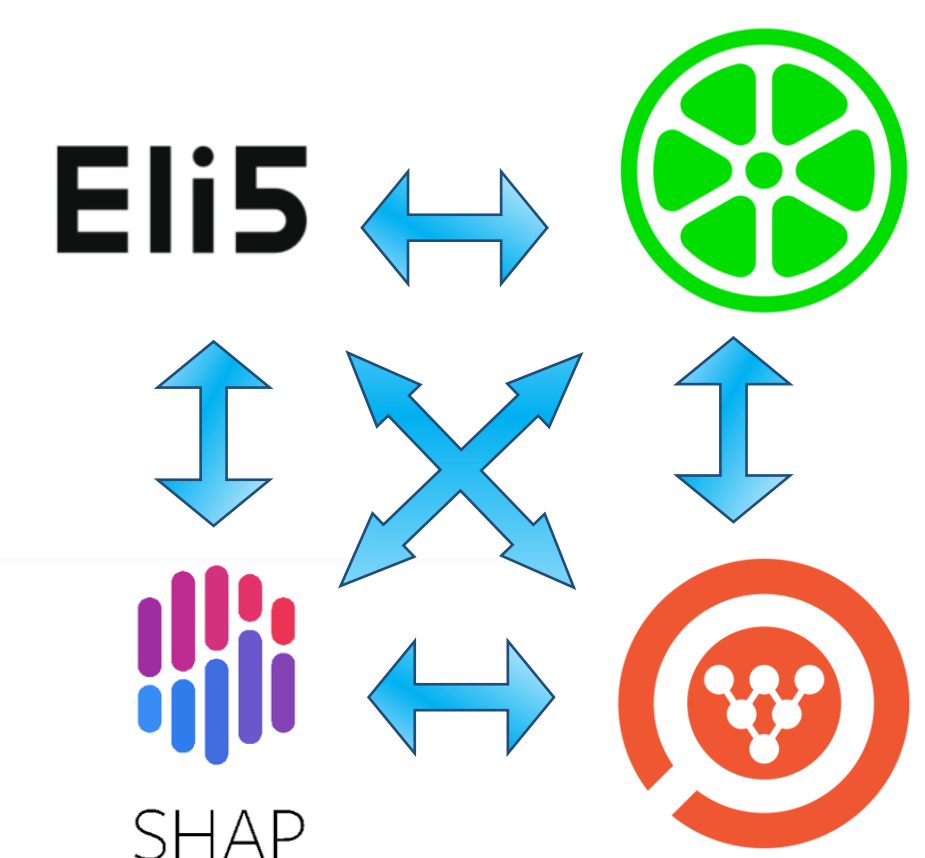


Fig.2. Comparing state-of-the-art techniques like Eli5, Lime, SHAP, Captum

Explainability

How to take an ML model and explain the behavior in human terms.

Interpretability

The degree to which an observer can understand the cause of a decision.

Transparency

A model is considered transparent if it is expressive enough to be human understandable.

Conclusion

As a first iteration of my research on the topic, I conducted a literature review on the state-of-the-art techniques and tools for explaining machine learning models. I gathered information to be able to make a proposal on my future study which is a throughout literature overview on how can we help humans to interpret more complex AI models, and comparing the most important tools in the field based on different aspects like methodology or human usability.

Usability perspective

Besides feature importance and the methodology, the research focuses on finding XAI tools that are easily applicable in practice. An important aspect of this is the usability of the XAI. Can we use it independently from the model architecture or it has to be integrated in the model? Later on the research, I will focus on particularly Post-Hoc XAI techniques.

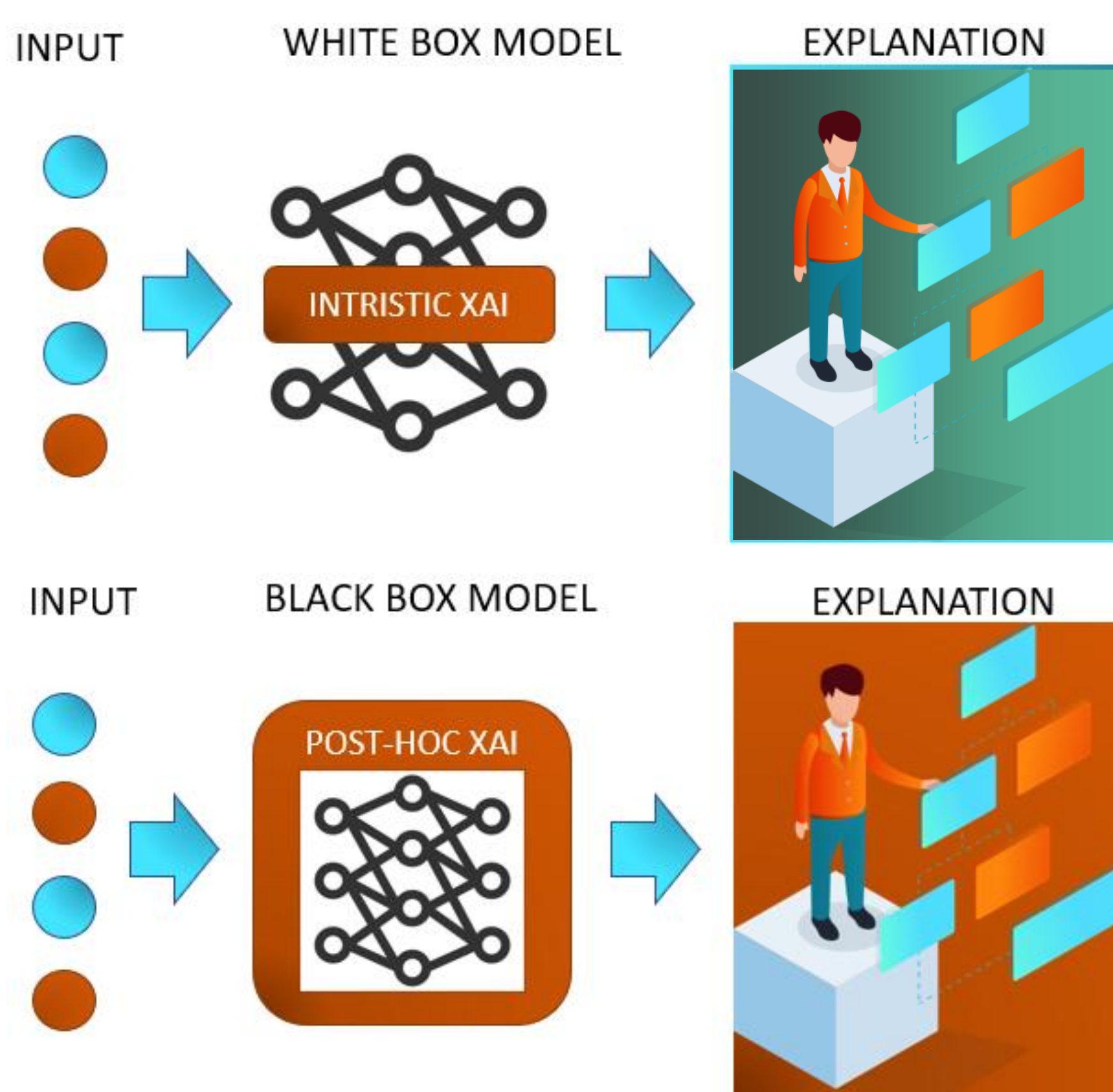


Fig.3. Intrinsic vs Post-Hoc solutions regarding the usage of Explainable AI

I propose to implement a method that uses existing Post-Hoc methods and allows end users to easily bind any of their models with tools that fit the best to explain them.

References

- AlejandroBarredo Arrieta et Al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI.
- Das, A., & Rad, P. (2020). Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. ArXiv.
- Idit Cohen. (2021) Explainable AI (XAI) with SHAP -Multi-Class Classification Problem.