Universidad Politécnica de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

Data Science Seminars

Explainable Artificial Intelligence (XAI) and Model Interpretability

Authors:

Joseph Tartivel

Teacher:

MS Management Solution

Date:
September 4, 2025

Contents

1 Introduction to XAI

Explainable Artificial Intelligence (XAI) is a burgeoning field focused on making AI systems more transparent and understandable to humans. As AI becomes increasingly integrated into critical sectors such as healthcare, finance, and law, the need for models that can explain their decisions is paramount. XAI aims to bridge the gap between the complexity of AI models and the need for transparency and trust in their outputs. By providing clear explanations for AI-driven decisions, XAI enables stakeholders to make better-informed choices and ensures that AI is used responsibly and ethically12.

2 Key Techniques

LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations) are two widely used techniques for interpreting machine learning models. LIME works by perturbing input data and fitting simple, interpretable models to approximate complex model predictions, helping stakeholders understand which features influence decisions, such as income or credit score in loan approvals. SHAP, based on Shapley values from game theory, assigns each feature a contribution score, offering both local and global explanations, useful in fields like healthcare to understand diagnostic models. Other techniques include PDP (Partial Dependence Plots), which visualize the marginal effect of one or more features on model predictions; ICE (Individual Conditional Expectation), which extends PDP by showing how predictions change for individual instances; and Feature Importance methods like permutation importance, which measure how a model's performance degrades when a feature is shuffled. These techniques collectively enhance interpretability and trust in machine learning models.

3 Applications

XAI is crucial in various domains where understanding the reasoning behind AI decisions is essential. In medicine, interpretable models can help clinicians understand why a particular treatment is recommended, fostering trust and ensuring patient safety. For example, an XAI model might explain that a patient's genetic profile and medical history are key factors in recommending a specific treatment plan. In the legal sector, transparent decision-making processes are vital for upholding justice and accountability. For instance, an AI system used in sentencing could provide explanations for its recommendations, ensuring fairness and accountability. In finance, interpretable models aid in risk assessment and regulatory compliance, ensuring ethical and understandable decisions. For example, a bank might use XAI to explain why a particular loan application was denied, providing transparency to both the applicant and regulators.

4 Conclusion

The future of Explainable AI (XAI) lies in overcoming the challenge of balancing model complexity with interpretability, ensuring explanations remain both accurate and accessible to non-experts. While techniques like LIME and SHAP provide valuable insights, future research will likely focus on enhancing user-friendliness through intuitive visualizations and natural language explanations. Ethical considerations will also become integral to AI design, ensuring transparency and fairness. As AI continues to shape various domains, from finance to healthcare and autonomous vehicles, the ability to explain decisions will be crucial for building trust, improving accountability, and ensuring responsible AI deployment.