RADemics

Explainable Artificial Intelligence Techniques for Transparent Decision Support Models

Dr. Renuka Viswanathan, C. Gethara Gowri, Dr. P .Usha

ST. JOSEPH'S COLLEGE OF ENGINEERING, RAJALAKSHMI INSTITUTE OF TECHNOLOGY, ST.PETER'S ENGINEERING COLLEGE

Explainable Artificial Intelligence Techniques for Transparent Decision Support Models

[1]Renuka Viswanathan, Professor, Department of Chemical Engineering, St. Joseph's College of Engineering, OMR, Chennai - 600 119. renukav@stjosephs.ac.in

[2]C. Gethara Gowri, Assistant Professor, Department of Computer Science and Engineering, Rajalakshmi Institute of Technology, Poonamalle, Chennai- 600124. gowri.smak@gmail.com

[3]P. Usha, Assistant professor, Department of Computer Science Engineering, St. Peter's Engineering College, Medchal, Hyderabad- 500100. papineni.usha@gmail.com.

Abstract

Explainable Artificial Intelligence (XAI) represents a critical advancement in developing transparent, accountable, and trustworthy decision support systems across complex application domains. This book chapter examines fundamental explainability principles, model-agnostic interpretation techniques, deep learning explanation methods, and their applications in classification and predictive decision-making. Particular attention focuses on permutation feature importance, Layer-Wise Relevance Propagation, misclassification analysis, and the interpretation of structured financial data. The discussion explores how SHAP, LIME, feature attribution, and counterfactual explanations reveal influential variables, clarify individual predictions, and support systematic model evaluation. Financial risk and credit assessment serve as important application areas, demonstrating the relevance of interpretable predictions for lending decisions, default risk estimation, fairness evaluation, and regulatory accountability. Critical challenges, including black-box complexity, explanation instability, correlated features, computational overhead, and potential algorithmic bias, receive focused consideration. The chapter also highlights the importance of evaluating explanation fidelity, robustness, consistency, and human comprehensibility alongside predictive performance. An integrated explainability framework combining technical interpretation, domain expertise, fairness assessment, and continuous validation provides a foundation for responsible artificial intelligence deployment. The chapter contributes practical insights for researchers, developers, and decision-makers seeking reliable predictive models that balance computational effectiveness with meaningful transparency.

Keywords: Explainable Artificial Intelligence, Model-Agnostic Explainability, Feature Importance, Layer-Wise Relevance Propagation, Financial Risk Assessment, Transparent Decision Support

Introduction

Artificial Intelligence (AI) has transformed contemporary decision-making through its ability to analyse large datasets, identify complex relationships, and generate predictive insights across diverse application domains [1]. Deep learning and machine learning models support critical activities such as financial risk assessment, credit approval, medical diagnosis, fraud detection, and industrial process optimization [2]. The increasing complexity of these models frequently

limits the interpretability of their predictions, creating difficulties for users seeking to understand the reasoning behind automated decisions [3,4]. Explainable Artificial Intelligence (XAI) addresses this challenge by developing methods that communicate model behaviour through understandable explanations. Its growing importance reflects the demand for reliable, transparent, and accountable intelligent decision support systems [5].

The development of explainable decision support requires a comprehensive understanding of how different interpretation techniques operate across machine learning architectures [6]. Model-agnostic approaches, including permutation feature importance, SHAP, and LIME, examine relationships between input variables and predictive outcomes without depending on a specific model structure [7]. Deep learning interpretation techniques, such as Layer-Wise Relevance Propagation, trace prediction relevance through neural network layers to identify influential input components [8]. These approaches provide complementary perspectives on global feature significance and individual prediction behaviour [9]. Selecting an appropriate technique requires consideration of model architecture, dataset characteristics, computational requirements, and intended decision-making objectives, ensuring that explanations remain meaningful within their operational context [10].

Reliable predictive decision-making also depends on systematic investigation of model errors and explanation quality [11]. Misclassified instances can expose weaknesses associated with incomplete training data, overlapping class characteristics, inappropriate decision thresholds, and misleading feature relationships [12]. Explainability methods support the examination of these errors by identifying influential variables and highlighting patterns associated with incorrect predictions [13]. Such investigations contribute to model refinement, improved data representation, and more effective validation procedures. Explanation quality also requires assessment through fidelity, stability, robustness, and human comprehensibility [14,15].