An Analytical Review: Explainable AI for decision making in finance using Machine Learning

Tanvi Shah
Faculty of engineering &technology
Parul institute of engineering & technology
Vadodara, India
2203032010018@paruluniversity.ac.in
Kishori Shekokar

Faculty of engineering &technology
Parul institute of engineering & technology
Vadodara, India
kishori.shekokar20174@paruluniversity.ac.in

Amit Barve
Faculty of engineering &technology
Parul institute of engineering & technology
Vadodara, India
amit.barve17535@paruluniversity.ac.in

Pramod Khandare
Public university in Liverpool John Moores University
Liverpool, United Kingdom
Pramod.khandare09@gmail.com

2024 Parul International Conference on Engineering and Technology (PICET) | 979-8-3503-6974-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/PICET60765.2024.10716075

***Abstract*—Explainable AI is a type of artificial intelligence which enables the explanation of learning models and states why the system arrived at a particular decision, exploring its logical paradigms, contrary to the inherent black box nature of artificial intelligence. Similarly, machine learning interpretability allows users to comprehend the results of the learning models by providing reasoning for the decisions that it has arrived at. Because of the enormous, continued success in machine learning, including statistically learning from large data, the finance world is becoming increasingly interested in this topic with more transparency in decisions to reduce risk. Machine learning algorithms with AI can shift through vast amounts of historical data to identify patterns and trends, enabling businesses to make accurate predictions about future outcomes. This paper reviews brief historical introduction about XAI, its approaches and taxonomy as well as links to their programming implementations with specific directions for finance.**

Keywords— Explainable AI, XAI, Machine Learning, Finance, Decision making, White BoxModels

I. Introduction

The use of cutting-edge technologies has become essential for making well-informed and effective judgments in the constantly changing financial scene. A serious issue with the intrinsic opacity of many machine learning models arises when financial institutions depend more and more on ML algorithms and datasets to improve their decision-making processes. Deep neural networks in particular are typical examples of traditional machine learning models that are viewed as "black boxes", making it challenging to comprehend their decision-making process. Through investigating Explainable AI, and its application of "whiteboxes" in the financial sector with subject matter experts, this review seeks to close this gap.

A. Explainable AI

The idea of Explainable AI, or XAI, has become a crucial topic of study in an era where artificial intelligence and machine learning are being incorporated into almost every part of our lives. Fundamentally, XAI aims to solve the mysterious black box, which is a major obstacle in the field of artificial intelligence. The term 'black box' describes the innate intricacy of numerous machine learning models, which frequently function as enigmatic entities, rendering judgments grounded in complicated patterns and relationships found across extensive datasets. Even if these models are incredibly good at a variety of tasks, from financial forecasting to image recognition, their lack of transparency raises serious concerns, particularly for high-stakes applications. The goal of XAI is to demystify these "black boxes" and give stakeholders who are not in the AI field insight into how these systems make decisions.

Making AI understandable, interpretable, and most importantly better in accountability is the main goal. This demand for explainability is not specific to any one industry; it is present in a wide range of fields, including healthcare, banking, autonomous cars, and criminal justice. But this review states the significant effects of XAI in the context of finance and decision-making—an area with significant risks, intricate data, and far-reaching consequence.

B. Decision-making in Finance

Financial decisions cover a broad spectrum of options, from government fiscal policies and regulatory measures to asset management, business financial planning, and personal investment plans. A complicated web of factors, such as market conditions, legal limits, economic forecasts, and risk assessment, impact these decisions. The characteristic of financial acumen is the capacity to make wise and practical financial decisions. This is a skill that is highly valued and continuously developed in both personal and professional settings.

Although decision in finance varies for vital roles of quantitative analysis, economic forecasting, and risk assessment, the evolution of banking sectors is very broad. Within its ever-changing industry, decision-making takes on greater importance, especially when it comes to crucial tasks like credit assessment, loan approvals, and general financial transaction administration. The preciseness of decisions made in these areas are crucial in determining the safety and soundness of banks' finances, highlighting the complex link between successful decision-making and the prosperity of an organization. These decisions have an impact on social equality as well as economic stability, going beyond personal portfolios and business profits.

C. Machine Learning Algorithms

In artificial intelligence, machine learning is essentially a revolutionary paradigm. It's the technology that allows computers to draw conclusions, identify patterns in data, and learn from them without the need for explicit programming. Their versatility and adaptability allow them to tackle a wide range of tasks, including natural language processing, financial forecasting and medical diagnosis.

As long as the world keeps producing vast volumes of data,

979-8-3503-6974-8/24/$31.00 ©2024 IEEE

machine learning algorithms will become even more important in helping us make sense of this data-rich environment. They serve as the pivot in our efforts to clean insights, mechanize the process of making decisions, and dissect the intricate structure of the digital era.

II. Literature Review

A comprehensive review of the literature on "Explainable AI for Decision Making in Finance using Machine Learning" reveals a growth in academic research as well as practical applications in this subject. They have highlighted the need of openness and interpretability in AI models for financial decision-making.

Explainability is the ability to demonstrate and describe a model's behavior in terms that are understandable to humans, whereas interpretability is the ability to understand and witness a model's mechanism or expected behavior in response to input process. Gaining intrinsic interpretability involves creating prediction models, like all of the White-Box models, which are interpretable by definition [4].

In general, white box models are less complicated than their more complicated black box equivalents, like deep neural networks. These models are easier to comprehend and evaluate due to their simplicity. A more direct relationship between input feature and output prediction can be seen in transparent models, which frequently contain fewer parameters.

Although they are not totally opaque, black box models frequently lie in the middle of the transparency and opacity spectrums. These models provide difficulties in comprehending their decision-making processes because of their intricate internal mechanics and complexity. The absence of interpretability creates issues, especially in high-stakes fields like banking, healthcare, and autonomous systems, even though they might offer correct predictions or classifications. To tackle this difficulty, a growing number of scholars and professionals are investigating ways to improve the explainability of black box models. Grey-box models blend data-driven techniques with physics-based modeling, falling in between black-box and white-box models [19]. Approaches like interpretability frameworks, post-hoc analysis, and model-agnostic explanations provide ways to illuminate these models' internal mechanisms and give stakeholders a better understanding of the outcomes of their predictions.

A. Visualization-based techniques

In order to convey information about the model's choice, visual explainable approaches generate plots or visuals. To provide an explanation of a model's choice using a saliency map, most approaches generate values that represent the significance and input component contribution to the decision [17].

1) Model Agnostic outputs: The fact that one does not need to understand the internal model structure makes them easy to understand. This encompasses several approaches like attribute- based, attention-based, perturbation-based, Class Activation Mapping- based, counterfactual interpretation, and more [15].

2) Model specific explanations: The information concealed within the model can be revealed via these designs. It covers dimensionality reduction, concept attribution and explanation of graph neural networks [16].

B. Symbolic-based techniques

In Symbolic AI algorithms operate through the processing of symbols, which are representations of real- world objects or concept and their relationships. Logic- based programs use axioms and rules to draw conclusions, and are the primary method used in symbolic artificial intelligence.

1) Coding Symbolic: Given their strong ties to actual facts, these explanations are easier to accept and more objective, knowledge graph being a part of it.

2) Qualifying Symbolic: Past information may be incorporated into these explanations. It comprises semantic web outputs and rule-extraction methods [15].

C. XAI techniques in Finance

Making decisions is the foundation of practically every facet of the vast and multifaceted world of finance. In addition to many other areas, it covers banking, real estate finance, corporate finance, risk management, personal finance, and investment management.

1) Techniques for Interpretability to Describe Any Black- Box Model: The interpretability strategies covered in this section are applicable to all black/grey box model approaches [18].

a) SHAP: Finding the relevance values for every characteristic to individual predictions is the aim of the game theory-inspired Shapley Additive explanations approach, which aims to improve interpretability.

b) LIME: The local comprehensible model-agnostic explanation technique is a popular interpretability paradigm for black-box models. With a straightforward but effective methodology, it can interpret single prediction values generated by anyclassifier.

c) DLIME: A deterministic variant of the earlier approach, called method, is proposed to overcome various interpretations that can lead to deployment issues. This version groups the data by random perturbation rather than utilizing k-nearest neighbors to determine which cluster the instance is thought to belong to.

d) Anchors: This technique expresses adequate circumstances locally for prediction using high-precision objects known as anchors.

e) Protodash: It was subsequently enhanced by giving prototypes non-negative weightings based on their contributions, resulting in a coherent framework that accounts for all outliers. Prototypes were utilized to demonstrate its work in detail.

2) Techniques for Interpretability to Describe Any White-Box Model: The focus of this section is interpretability methods, which are applicable to any white-box model approach. [18]. The rule-based, decision- tree, and linear models are among the more complex and sophisticated models that fall under this category. These models are transparent as well and show promise in the interpretability space.

a) SLIM: A highly interpretable type of predictive system, super-parse linear integer models only allow the addition, subtraction, and multiplication of input feature values in order to generate predictions.

b) BRCG: Utilize Boolean rule Conjunctive Normal Form or Disjunctive Normal Form of Boolean rules are used in Column Generation as a way to construct predictive models.

c) GLM: Generated from rules, Generalized Linear Models are linear combinations of features that are often called rule ensembles. These models, although rather complicated and adaptable, have the benefit of being intuitively interpretable because rules may describe nonlinear interactions and dependencies sequentially.

Global and local interpretation approaches were used in the sensitivity analysis of Machine Learning Model Predictions in the past, depending on whether the output alteration was examined in relation to a single example or all cases in the dataset. These techniques involve an examination whereby models are evaluated in terms of the stability of their learned functions and the degree to which small but deliberate modifications in the pertinent inputs can impact their anticipated outputs. Among other things, it also suggests using conventional adversarial example- based sensitivity analysis to get better results even with the imbalanced datasets.

III WHITE-BOX MODELS

In order for an algorithm to function, which is composed of smaller problems that are part of algorithms, must be resolved. A structure can be constructed by joining reusable components through sub-problems to show white- box algorithm design, which can generate decision tree models with higher accuracy [5]. Its performance is enhanced by the employment of an ensemble of weak learners in gradient boosting. Typically, decision trees are the weakest learners but their combined output yields more advanced models. AdaBoost builds the model sequence using a different method than XGBoost, which is an optimized version of Gradient Boosting with various extensions and improvements.

A. Gradient Boosting

Gradient boosting machines uses a learning process that fits new models one after the other to produce a response variable estimate that is more accurate. The concept underlying this method is to construct new base-learners having a significant relationship with the negative gradient of a loss function in the aggregate [6].

First, a single-node tree is constructed that predicts the total value of Y in regression problems or the log of Y in classification issues. Next, trees with more depth are developed on the residuals of the preceding classifier. In GBM, each tree typically grows with 8–32 terminal nodes. To help the model take baby steps in the correct direction to capture the variance and train the classifier on it, learning rates are provided as constants for each tree.

It mainly consists of 3 elements including Loss Function, Weak Learners and Additive model. The used loss functions are not limited; however, to provide a clearer understanding, the learning process would lead to successive error-fitting if the error function was the traditional squared loss. Weak learners are used to developing powerful prediction model designs for boosting algorithms by learning from prior mistakes. The basic idea of the additive model is the addition of trees in a single step.

B. XGBoost

XGBoost stands for Extreme Gradient Boost. XGBoost is an inclination boosting system-based decision tree-based machine learning calculation [7]. It offers parallel tree as a boosting mechanism which shows a fairly good interpretive capacity regarding the operation of the learning rate/shrinkage. In order to fully utilize the CPU core of the computer, a multi-threaded technique is used, which increases speed and performance [8].

The effectiveness of machine learning models is heavily influenced by the hyperparameters that regulate the learning process. The criterion, maximum depth, and a few estimators are all hyperparameters in the XGBoost model. The goal of hyperparameter optimization is to identify the best set of hyperparameter values that will yield the greatest efficiency on the data in an adequate amount of time [9].

C. AdaBoost

AdaBoost stands for Adaptive Boost. An AdaBoost classifier fits a copy of its original classifier on an updated dataset after first fitting it on the original dataset. This allows the classifiers to focus on cases that cause more inaccuracy by balancing out the inaccurate and error points [7]. The principle behind optimizing algorithms is to construct a model on the training dataset first, then a second model to fix the errors in the first model. This process is repeated until the mistakes are as small as possible and the dataset can be anticipated accurately.

The process is known as Adaptive Boosting since each instance has its weights re-assigned. Under these circumstances, the alpha parameter's value will be inversely correlated with the weak learner's error.

A brief summary of several approaches and the accompanying predictive modeling accuracies is given in this table. The Local Interpretable Model-Agnostic Explanations method has an impressive accuracy of 99%, which sets it apart from other methods like XGBoost with TreeSHAP (73%), SHAP algorithm (75%), Shapley values combined with an XGBoost model (81-93%), and Logistic Regression with Decision Tree (81%) that show varying degrees of success.

Though LIME is a global surrogate method for an interpretable model that can explain predictions for a black-box approach, boosting models can be enhanced with hyperparameter tuning to get better results.

TABLE I. AI MODEL REVIEW IN FINANCE

Title	Methodology	Accuracy (%)
"Explainable Machine Learning in Credit Risk Management"[10]	XG Boost, Tree SHAP	73
"The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to- peer lending"[11]	SHAP Algorithm	75

"Explainable AI in Fintech Risk Management" [12]	Shapley values and XGBoost model	81 - 93
"Employing Explainable AI to Optimize the Return Target Function of a Loan Portfolio" [13]	Logistic Regression, Decision Tree	81
"Extending machine learning prediction capabilities by explainable AI in financial time series prediction" [14]	Local Interpretable Model- Agnostic Explanations (LIME)	99

IV Future Scope

Data security and data loading are the main problems that need to be resolved while connecting a multi-agent environment, even though more current models have produced improved results. Secure data becomes critical in a multi-agent context when several independent entities communicate with one another. This entails making certain that the information shared amongst agents is essential, private, and accessible to those who are permitted. The preparation, aggregation, and distribution of data to agents for analysis or decision-making are all included in data loading.

In order to provide more visceral explanations, future research should look into alternatives to word depiction of things, including annotations in a virtual environment. It is essential to create such environments where agents can interact with annotated objects or simulations to better comprehend abstract concepts or complex scenarios. Employing data visualization techniques such as graphs, charts, or 3D models to represent data and relationships visually, facilitating interpretation with interactive interfaces that allows to manipulate data representations.

The study's findings necessitate application scenarios where one must deal with target functions whose assessments are noticeably asymmetrical and imbalanced datasets. In real terms, this refers to circumstances in which there is a significant degree of skewness in the distribution of data points among several classes or categories, making it difficult to attempt to make precise judgments or forecasts. Future research should include a better understanding of the predictions through clustering of the Shapley values using correlation network models. The hybrid explanation should be applied by concerning fusing heterogeneous knowledge from different sources, managing time-sensitive data, inconsistency and uncertainty.

For the better results, "traditional logic-based methods" and "Statistical machine learning" must work together flawlessly. There should be ways to strengthen the integration of domain-relevant data into the AI system and integrate domain expertise with AI outcomes. To evaluate the frameworks using data with many labels and classes and improve the accuracy of the models, more datasets are also required. Enterprises can offer tailored experiences and suggestions across the whole consumer journey by collecting and evaluating data from several touchpoints. Finally, when evaluating XAI for finance, local laws and regulations need to be considered.

Further, XAI methods can be used to improve Risk Management, Regulatory Compliance, Algorithmic Trading, Credit Scoring and Lending, Fraud Detection, Customer Service and Personalization with Ethical and Responsible AI.

V. Conclusion

Considering this experiment, XGBoost predicts better results than AdaBoost with tuned parameters whereas AdaBoost can be more efficient when predicted with the mounted data respectively. Furthermore, improved collaboration between AI systems and human experts is made possible by the ability to comprehend machine learning models in the banking sector. Financial professionals can utilize XAI to validate model outputs, understand the key variables impacting projections, and refine their strategies based on lucid insights. This cooperation improves decision-making by combining the domain knowledge of human specialists with the analytical capacity of AI. By enhancing openness, accountability, and collaboration, XAI contributes to the growth of a more robust and dependable financial ecosystem.

REFERENCES

[1] G. Niklas Bussmann, Paolo Giudici, Dimitri Marinelli & Jochen Papenbrock, "Explainable Machine Learning in Credit Risk Management", Springer, Computational Economics 57, 203–216 (2021).

[2] J Plamen P. Angelov, Eduardo A. Soares, Richard Jiang, Nicholas I. Arnold, Peter M. Atkinson, "Explainable artificial intelligence: an analytical review", WIREs Data Mining and Knowledge Discovery, 72,2021.

[3] Murat Dikmen, Catherine Burns, "The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending", ScienceDirect, International Journal of Human-Computer Studies, Volume 162, June 2022, 102792.

[4] Emmanuel Pintelas, Ioannis E. Livieris, Panagiotis Pintelas, "A Grey- Box Ensemble Model Exploiting Black-Box Accuracy and White-Box Intrinsic Interpretability", MDPI, Department of Electrical & Computer Engineering, University of Patras, Jan 05, 2020.

[5] Boris Delibas, Milan Vukic Evic, Milos Jovanovic, "White-Box Decision Tree Algorithms: A Pilot Study on Perceived Usefulness, Perceived Ease of Use, and Perceived Understanding", International Journal of Engineering Education Vol. 29, No. 3, pp. 674–687, 2013.

[6] Alexey Natekin, Alois Knoll, "Gradient boosting machines, a tutorial", Frontiers in Neurorobotics, Volume 7 – 2013.

[7] Syeda Sarah Azmi, Shwetha Baliga, "An Overview of Boosting Decision Tree Algorithms utilizing AdaBoost and XGBoost Boosting strategies", International Research Journal of Engineering and Technology, Volume: 07 Issue: 05 | May 2020.

[8] Ramraj S, Nishant Uzirb Sunil R, Shatadeep Banerjee, "Experimenting XGBoost Algorithm for Prediction and Classification of Different Datasets", Researchgate, International Journal of Control Theory and Applications, Volume 9 • Number 40, 2016.

[9] Surjeet Dalal, Bijeta Seth ,Magdalena Radulescu, Carmen Secara, Claudia Tolea, "Predicting Fraud in Financial Payment Services through Optimized Hyper-Parameter-Tuned XGBoost Model", MDPI Special Issue, Multivariate Data Analysis and Machine-Learning Models in Financial Analysis, 9 December 2022.

[10] Niklas Bussmann, Paolo Giudici, Dimitri Marinelli & Jochen Papenbrock, "Explainable Machine Learning in Credit Risk Management", Springer, Computational Economics 57, 203–216 (2021).

[11] Murat Dikmen, Catherine Burns, "The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending", ScienceDirect, International Journal of Human-Computer Studies, Volume 162, June 2022, 102792.

[12] Niklas Bussmann, Paolo Giudici, Dimitri Marinelli and Jochen Papenbrock, "Explainable AI in Fintech Risk Management", Frontiers, Sec. Artificial Intelligence in Finance, Volume 3 – 2020.

[13] Thomas Gramespacher, Jan-Alexander Posth, "Employing Explainable AI to Optimize the Return Target Function of a Loan Portfolio", Frontiers Artificial Intelligence, 15 June 2021, Sec. Artificial Intelligence in Finance, Volume 4 – 2021.

[14] Taha Buğra Çelik, Özgür İcan, Elif Bulut, "Extending machine learning prediction capabilities by explainable AI in financial time

series prediction", ScienceDirect, Applied Soft Computing, Volume 132,January2023, 109876.

[15] Pascal Hamm, Michael Klesel, Patricia Coberger & H. Felix Wittmann, "Explanation matters: An experimental study on explainable AI", Springer, Electronic Markets Article 33, Article number: 17 (2023).

[16] Wenli Yang, Yuchen Wei, Hanyu Wei, Yanyu Chen, Guan Huang, Xiang Li, Renjie Li, Naimeng Yao, Xinyi Wang, Xiaotong Gu, Muhammad Bilal Amin & Byeong Kang, "Survey on Explainable AI: From Approaches, Limitations and Applications Aspects", Springer, Human-Centric Intelligent Systems 3, pages161–188 (2023).

[17] Ilias Papastratis, "Explainable AI (XAI): A survey of recent methods, applications and frameworks", AI Summer, 2021.

[18] Pantelis Linardatos, Vasilis Papastefanopoulos and Sotiris Kotsiantis, "Explainable AI: A Review of Machine Learning Interpretability Methods", Department of Mathematics, University of Patras, 26504 Patras, Greece, 2021

[19] Christy Green, Srinivas Garimella, "Residential microgrid optimization using grey-box and black-box modeling methods", Energy and Buildings, Volume 235, 15 March 2021, 110705.