

Explainable Artificial Intelligence for Loan Credit Approval: A Literature Review of Transparency, Interpretability, and Fairness

Ashish Uttam Kakade, Prof. Bhosale Sachin Bajirao,
Dr. Khatri A. A, Prof. Bangar Abhishek Popat

Department of Computer Engineering Jaihind College of Engineering, Kuran Pune, India

Abstract- Artificial Intelligence (AI) is widely used in loan credit approval to assess applicants' creditworthiness. However, complex AI models often lack transparency. Explainable Artificial Intelligence (XAI) helps explain loan approval and rejection decisions using techniques such as SHAP and LIME. This paper reviews existing research on XAI in credit assessment, focusing on transparency, interpretability, and fairness. It also identifies key challenges and future research directions for developing reliable and understandable AI-based loan approval systems.

Keywords— Explainable Artificial Intelligence, Loan Credit Approval, Credit Risk Assessment, SHAP, LIME.

I. INTRODUCTION

Artificial Intelligence (AI) is increasingly being used in the financial sector for loan credit approval and credit risk assessment. Financial institutions evaluate factors such as income, credit history, existing loans, repayment capacity, and employment details to determine whether an applicant is eligible for a loan. AI-based systems can analyse these factors and support faster, data-driven decisions.

However, some AI models are complex and operate as black boxes, meaning that their decisions are difficult to understand. When a loan application is rejected, applicants may want to know the reasons behind the decision. Similarly, financial institutions need to understand the factors influencing predictions to evaluate reliability and potential bias.

Explainable Artificial Intelligence (XAI) aims to make AI predictions easier to understand. Techniques such as Local Interpretable Model-agnostic Explanations (LIME) and Shapley Additive explanations (SHAP) can identify features that influence individual predictions. This paper reviews relevant literature on explainability and examines its application to loan credit approval, focusing on transparency, fairness, interpretability, and research challenges.

II. REVIEW METHODOLOGY AND LITERATURE SELECTION CRITERIA

This paper follows a literature-review approach to examine research related to explainable AI and its application to credit assessment. The review focuses on foundational publications describing model-agnostic explanations, feature attribution, and the interpretation of AI predictions. These concepts provide a basis for understanding how loan approval and rejection decisions can be explained.

The literature selection criteria include relevance to explainable AI, clarity of the proposed explanation method, and usefulness for financial decision-making. Publications introducing LIME and SHAP are included because they provide widely recognised approaches to explaining model predictions. Research discussing fairness, transparency, and accountability is also relevant to evaluating the suitability of AI-based credit approval systems.

The review is conceptual and based on selected publications rather than a systematic review of every available study. Therefore, the findings should be understood as a focused discussion of important methods and challenges, not as an exhaustive assessment of all published research.

III. LITERATURE REVIEW

1. Local Interpretable Model-Agnostic Explanations (LIME)

Ribeiro, Singh, and Guestrin introduced LIME as a method for explaining the predictions of machine-learning classifiers [1]. LIME approximates the behaviour of a complex model around a particular prediction using a simpler, interpretable model.

In loan credit approval, LIME can help identify which applicant details contributed to a particular prediction. For example, an explanation might indicate that a high existing debt level contributed to a predicted rejection, while stable income contributed positively to the prediction. These are illustrative examples; actual explanations depend on the trained model and applicant data.

LIME is useful because it can be applied to different types of predictive models. However, its explanations are local and may vary depending on the sampled data and settings used to generate them. Therefore, explanations should be evaluated for consistency and reliability before being used in financial decision-making.

2. Shapley Additive explanations (SHAP)

Lundberg and Lee introduced SHAP as a unified approach to interpreting model predictions [2]. SHAP uses Shapley values from cooperative game theory to estimate how individual features contribute to a prediction.

In credit assessment, SHAP can help show whether factors such as income, credit history, outstanding debt, or repayment records increase or decrease a model's predicted creditworthiness. This can provide applicants, loan officers, and developers with a clearer understanding of individual predictions.

SHAP offers a consistent framework for feature attribution, but the explanations still require careful interpretation. A feature's contribution to a model prediction does not automatically establish a causal relationship or prove that a lending decision is fair. The method also depends on the model being

explained, with computational and implementation considerations.

3. Explainability in Loan Credit Approval

Explainability is important in credit approval because lending decisions can significantly affect individuals' financial opportunities. A useful explanation should identify the factors influencing a prediction in language that relevant users can understand.

For example, a system might explain that an application received a lower predicted creditworthiness score because of a high debt-to-income ratio and limited repayment history. Such an explanation can help loan officers review the prediction and help applicants understand which factors influenced the result. However, a real system must ensure that its explanations accurately reflect the model and do not expose confidential information.

Explainability should also be considered alongside model performance, data quality, fairness, privacy, and human oversight. An understandable prediction is not necessarily correct or unbiased. Consequently, XAI should support, rather than replace, responsible lending policies and appropriate human review.

IV. COMPARATIVE ANALYSIS OF EXISTING STUDIES

Existing studies on Explainable Artificial Intelligence (XAI) have introduced different methods to make AI predictions more understandable. Ribeiro et al. [1] proposed LIME, which explains individual predictions using a simpler interpretable model. Lundberg and Lee [2] introduced SHAP, which estimates the contribution of individual features to a prediction. Both methods can help explain the factors influencing loan credit approval and rejection.

Guidotti et al. [5] reviewed different methods for explaining black-box models and discussed their importance in understanding complex AI systems. Adadi and Berrada [6] presented an overview of XAI concepts, opportunities, and challenges. Wachter et al. [7] explored counterfactual explanations, which describe how changes to input features could lead

to a different automated decision. These approaches provide different ways to improve transparency and help users understand AI-based decisions.

Research on credit risk management highlights the importance of explainability in financial decision-making. Bussmann et al. [3] discussed the application of explainable machine learning in credit risk management, while Wang et al. [9] studied explainable credit risk scoring using small-business banking data. These studies are particularly relevant to loan credit approval because they connect explanation methods with credit assessment.

Table 1: Comparison of LIME and SHAP

Feature	LIME	SHAP
Main approach	Local approximation using a simpler model	Shapley-value-based feature attribution
Explanation focus	Behaviour near an individual prediction	Contribution of features to a prediction
Model compatibility	Model-agnostic approach	General framework with model-specific and model-agnostic implementations
Credit approval use	Explaining an individual approval or rejection	Estimating how applicant features influence a prediction
Main limitation	Explanations may vary with sampling and settings	Computational cost and interpretation can be challenging

Both techniques can improve the interpretability of complex AI systems. LIME generates local explanations, while SHAP provides a structured way to attribute predictions to input features. Neither technique independently guarantees fairness, correctness, or compliance with lending requirements. Their suitability should be assessed according to the model, data, explanation quality, and intended users.

V. PROPOSED CONCEPTUAL ARCHITECTURE

The proposed architecture for Explainable Artificial Intelligence (XAI) in loan credit approval consists of several stages. First, applicant data, including income, credit history, and existing debt, is collected and preprocessed through data cleaning, encoding, and bias auditing. An AI-based prediction model then analyzes the processed data to estimate the applicant's creditworthiness. The explanation layer uses techniques such as Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) to identify the important factors influencing the model's prediction.

The generated explanations are evaluated for stability, faithfulness, and fairness to improve their reliability. The system then converts the technical results into simple, understandable reasons that explain the loan assessment. A human loan officer reviews uncertain or complex cases before the final decision is made. Finally, the loan application is approved or rejected based on the assessment and appropriate review, with clear reasons provided for the decision. This architecture aims to improve transparency, interpretability, and trust in AI-assisted loan credit approval.

VI. RESEARCH CHALLENGES AND IDENTIFIED GAPS

Despite progress in explainable AI, several challenges remain when applying these techniques to loan credit approval.

- **Model Complexity:** Complex models may achieve useful predictive performance but remain difficult to interpret. Explanations must simplify model behaviour without misleading users about how decisions are produced.
- **Fairness and Bias:** Historical lending data may contain patterns associated with unfair treatment. XAI can reveal influential features, but explanation alone cannot remove bias or establish that a model treats applicants fairly.

- **Explanation Reliability:** Different explanation methods or settings may produce different results for the same prediction. Unstable explanations can reduce user trust and make decisions harder to review.
- **Privacy and Security:** Credit applications contain sensitive personal and financial information. Explanation systems must protect this information while providing meaningful reasons for decisions.
- **Understandability:** Technical explanations containing feature values or attribution scores may be difficult for ordinary applicants to understand. Explanations should be presented in clear and accessible language.
- **Evaluation of Explanations:** Model accuracy alone is insufficient to measure explanation quality. Further evaluation is needed to assess faithfulness, stability, clarity, usefulness, and fairness in practical lending situations.

These challenges indicate the need for credit approval systems that combine reliable prediction, understandable explanations, fairness assessment, and human oversight.

VII. PROPOSED RESEARCH DIRECTIONS

Future research can explore an explainable loan credit approval framework that combines predictive modelling with post-hoc explanation techniques. Such a framework could use applicant information to generate a creditworthiness prediction and then apply LIME or SHAP to identify the factors influencing that prediction.

First, explanations should be evaluated for stability and consistency. Similar applicant profiles should not receive confusingly different explanations without a meaningful reason. Second, fairness checks should be conducted to identify potentially discriminatory patterns in data and model outputs. Third, explanations should be translated into simple language so that applicants and loan officers can understand the main factors behind a decision.

Future systems should also support human review of high-impact or uncertain decisions. The explanation

interface should clearly distinguish a model's prediction from the final lending decision. Privacy protection, secure data handling, and appropriate governance should be included in the design.

These directions represent a proposed research approach based on the reviewed concepts. They are not claims of a developed or experimentally validated system. Further empirical research would be required to compare methods using suitable datasets and defined evaluation measures

VII. CONCLUSION AND FUTURE SCOPE

Explainable Artificial Intelligence (XAI) can improve transparency in AI-based loan credit approval by helping users understand the factors influencing approval and rejection decisions. This review discussed LIME and SHAP as important techniques for explaining AI predictions. These methods can help financial institutions understand model outputs and provide clearer explanations to applicants.

However, explainability alone does not guarantee accurate or fair decisions. Challenges such as data bias, explanation reliability, privacy, and model complexity must be addressed. Combining explainable AI with fairness evaluation and human oversight can support more responsible credit assessment.

In the future, research can focus on developing reliable XAI-based loan approval systems that provide simple explanations, detect potential bias, and protect applicants' financial information. Testing these approaches using suitable credit datasets will help evaluate their effectiveness and practical usefulness.

REFERENCES

1. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

2. S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
3. N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, "Explainable Machine Learning in Credit Risk Management," *Computational Economics*, vol. 57, pp. 203–216, 2021.
4. A. Gramegna and P. Giudici, "SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk," *Frontiers in Artificial Intelligence*, vol. 4, art. 752558, 2021.
5. R. Guidotti et al., "A Survey of Methods for Explaining Black Box Models," *ACM Computing Surveys*, vol. 51, no. 5, art. 93, 2018.
6. A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.
7. S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR," *Harvard Journal of Law & Technology*, vol. 31, no. 2, pp. 841–887, 2018.
8. B. Hadji Misheva, J. Osterrieder, A. Hirs, O. Kulkarni, and S. F. Lin, "Explainable AI in Credit Risk Management," *arXiv:2103.00949*, 2021.
9. W. Wang, C. Lesner, A. Ran, M. Rukonic, J. Xue, and E. Shiu, "Using Small Business Banking Data for Explainable Credit Risk Scoring," *Proc. AAAI Conf. Artificial Intelligence*, vol. 34, no. 8, pp. 13396–13401, 2020.
10. R. K. Mothilal, A. Sharma, and C. Tan, "Explaining Machine Learning Classifiers Through Diverse Counterfactual Explanations," *Proc. FAT**, pp. 607–617, 2020.
11. U. Bhatt et al., "Explainable Machine Learning in Deployment," *Proc. FAT**, 2020.
12. M. Bucker, G. Szepannek, A. Gosiewska, and P. Biecek, "Transparency, Auditability, and Explainability of Machine Learning Models in Credit Scoring," *Journal of the Operational Research Society*, vol. 73, no. 1, pp. 70–90, 2022.
13. J. Dastile, T. Celik, and M. Potsane, "Statistical and Machine Learning Models in Credit Scoring: A Systematic Literature Survey," *Applied Soft Computing*, vol. 91, art. 106263, 2020.
14. X. Dastile, T. Celik, and H. Vandierendonck, "Model-Agnostic Counterfactual Explanations in Credit Scoring," *IEEE Access*, vol. 10, pp. 69543–69554, 2022.