Enhancing Scalability and Transparency in AI-Driven Credit Scoring: Optimizing Explainability for Large-Scale Financial Systems

Daniel Thomas
daniel.cromwel@gmail.com

***Abstract*—The growing adoption of artificial intelligence (AI) in credit scoring has significantly enhanced predictive accuracy, but it has also raised concerns regarding transparency, fairness, and trust. The "black box" nature of many machine learning models used in financial decision-making can hinder understanding and accountability, particularly in high-stakes scenarios such as loan approvals. To address these challenges, it is essential to develop methods that improve the explainability and scalability of AI-driven credit scoring systems. This study explores how the scalability of explainability techniques degrades with increasing data volume in tree-based ensemble models like XGBoost and investigates strategies to optimize performance, such as feature selection and model refinement. By applying these approaches to a dataset of 2.3 million loan applications from Lending Club, the research aims to provide insights into improving the efficiency and transparency of large-scale AI systems. The findings will contribute to more transparent, fair, and efficient credit scoring models, ensuring that AI-driven decisions are both interpretable and compliant with regulatory standards.**

I. Introduction

"Expanding access to affordable credit" was the motto of Upstart when it was first founded in 2012. Upstart is one of the pioneers in adopting artificial intelligence (AI) for credit scoring. Traditional credit scoring models, such as those based on FICO scores, can often be biased against those with thin or non-traditional credit files, which is common in underprivileged communities. With the rapid growth of machine learning (ML) in financial services, these modern AI-driven models can leverage big data to assess a broader range of variables, potentially providing more accurate credit assessments. However, the black-box nature of these predictive models has led to concerns over fairness, accountability, and transparency.

In addition, recent regulations, such as the 'right to explanation' introduced by the General Data Protection Regulation (GDPR), have driven the growth of Explainable AI (XAI), which focuses on providing human-understandable explanations for decisions made by these models. Applying XAI in credit scoring is crucial to ensuring that AI models can achieve stellar decision accuracy and comply with regulatory standards and ethical practices.

The primary research questions guiding this review are:

RQ1 What XAI techniques are adopted for the credit scoring model?

RQ2 How do different XAI methods compare regarding their ability to explain credit-scoring decisions?

The main objectives of this literature are:

1) To identify and compare different XAI techniques applied to credit scoring models.
2) To identify gaps in current research, especially in the practical application and user understanding of XAI explanations in credit scoring models.

This review will focus on XAI applications in credit scoring within customer lending, excluding broader financial applications such as fraud detection or insurance underwriting.

II. Literature Review

A. Terminologies

To provide a clearer understanding of the concepts discussed in this review, the following key terminologies are defined:

Classifiers: Algorithms that label input data by predicting the class or category it belongs to. Examples of classifiers include Logistic Regression, Support Vector Machine (SVM), Decision Trees, and Neural Networks.

Post-hoc methods: Techniques used to explain and interpret the predictions made by an ML model after it has been trained. LIME and SHAP are two examples of post hoc methods aiming to explain how a "black box" model arrived at its decisions.

Model agnostic methods: XAI techniques that apply to all ML model types regardless of their algorithms. They analyze the model's inputs and outputs to generate general explanations. Both LIME and SHAP are model-agnostic techniques.

Local explanation: Aims to explain individual predictions made by the model

Global explanation: This aims to provide an overall understanding of the entire model's behavior, for instance, how the model makes predictions across an entire dataset.

Feature: Features are the inputs of an ML model. The model uses these inputs to make predictions.

Feature Importance: The degree of influence of each feature (input) on the model's predictions.

Credit risk / Credit score: These terms are synonymous and are thus used interchangeably throughout this report.

B. Overview of Selected Papers

1) **Paper 1: A Study on Credit Scoring Modelling with Different Feature Selection and Machine Learning Approaches**
Research Focus:
Written by [1], this paper investigates three feature selection techniques (Information Gain, Gain Ratio, and Chi-Square) and five machine learning models (Bayesian, Naïve Bayes, Random Forest, Decision Tree (C5.0), and Support Vector Machine) to determine the most effective combination for an automated credit scoring model.
Methodology:
The author chooses the publicly available Statlog German Credit Dataset to develop the credit scoring model. The pre-processing step is omitted as the dataset is already clean and structured in nature. Next, each of the three feature selection techniques is applied to the dataset to prioritize the most relevant features before being used to train each of the five machine learning models. Hence, for each feature selection technique, there are five machine learning models trained, leading to a total of 15 credit scoring models. The author then evaluates each of these 15 combinations based on its Accuracy, F-measure, False Positive Rate (FP), False Negative Rate (FN) and Training Time.
Key Findings:
The combination of Random Forest and Chi-Square feature selection was identified as the best-performing model in terms of accuracy, F-measure, and low false positive and false negative rates. The Decision Tree (C5.0) and Chi-Square combination was the second-best performing model, showing results comparable to the first combination. However, the Random Forest has a longer training time (16.20s versus 8.22s) as compared to the Decision Tree.

2) **Paper 2: Explainable AI in Credit Risk Management**
Research Focus:
Written by [2], this paper focuses on applying explainable artificial intelligence (XAI) in credit risk management, specifically developing transparent and interpretable machine learning models for credit scoring. The authors used two post-hoc explainability techniques, LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (Shapley Additive exPlanations). The objective is to explain complex "black box" credit scoring models, from individual prediction to general decision-making.
Methodology:
The authors used a dataset from Lending Club, an American peer-to-peer lending platform that contained information on over 2.2 million loans. After preprocessing the dataset to handle missing values, reduce feature space, and encode categorical variables, it was used to train the four machine learning classifiers: Logistic Regression, Random Forest, XGBoost, Support Vector Machine (SVM), and one Neural network-based binary classifier. The primary objective of these models was to classify the loans as either 'default' or 'fully paid.'
The author then used LIME and SHAP to interpret and explain the predictions made by these models. LIME explained individual predictions by approximating complex models with simpler interpretable models, while SHAP provided both local and global feature explanations.
Key Findings:
The authors found that both LIME and SHAP effectively explained model predictions and identified the most critical features influencing credit risk. These features included total payment, loan amount, and recoveries. These explanations were in line with financial logic. LIME was beneficial for instance-level explanations, providing insights into why a particular loan was classified as high-risk or low-risk. SHAP offered broader insights into the entire model's decision-making process, showing the global impact of each feature and consistency across various samples. Despite these advances, the authors noted that both techniques had limitations as SHAP's computational complexity posed challenges for larger datasets, and LIME's scope was somewhat limited to probabilistic models.

3) **Paper 3: Explainable AI for Interpretable Credit Scoring**
Research Focus:
Written by [3], this paper proposes a credit scoring model that is both accurate and interpretable to different people in diverse situations. The authors choose the Extreme Gradient Boosting (XGBoost) model for classification as it performs strongly in credit scoring. In terms of interpretability, instead of only using a single XAI method, the model is enhanced with multiple post-hoc XAI methods to provide global, local instance-based, and local feature-based explanations to meet diverse needs.
Methodology:
The model proposed in this study uses two datasets: the Home Equity Line of Credit (HELOC) dataset and the Lending Club (LC) dataset. The data is first pre-processed by cleaning, feature selecting, test splitting, normalizing, cross-validating, and balancing. Next, a classifier function, XGBoost, is adopted to classify the data instances in this paper. Lastly, the classifier is extended by three post-hoc XAI methods, namely GIRP + SHAP, Anchors, and ProtoDash, to generate global, local feature-based, and local instance-based explanations, respectively.
The evaluations of the explanations generated are done through functionally grounded analysis, application-grounded analysis, and human-grounded analysis. The first analysis solely focuses on evaluating the explanations according to predetermined metrics. The second analysis involves domain experts (such as loan officers) to quantify the correctness and quality of the explanation. The final analysis focuses mainly on evaluating the interpretability of the explanations by lay humans rather than the cor-

rectness.

Key Findings:

Through the functionally grounded analysis, it is shown that all the types of explanations are simple, consistent, and complete. Six out of seven domain experts interviewed prefer the visual representation (a decision tree) rather than the IF-THEN representation of the provided global explanation. The multitude of explanations is deemed useful and has the potential to be implemented into banking systems. The local feature-based explanation also achieved an average understandability of 78% among 100 lay human participants.

4) **Paper 4: Artificial Intelligence and Bank Credit Analysis: A Review**

Research Focus:

Written by [4], this paper investigates and discusses the role of artificial intelligence (AI) in the economic sector, specifically in enhancing credit risk analysis by banks and FinTechs. The authors outline how AI techniques such as machine learning (ML) and neural networks are being used to process big data to improve credit risk prediction and enhance credit availability. Finally, the authors outline how using AI with big data, as opposed to traditional databases, poses ethical, legal, and regulatory challenges.

Methodology:

The authors conducted a literature review to investigate the role of AI in economic practices and, specifically, in credit analysis. It evaluates various AI models (such as random forests, classification trees, boosting, and bagging) and compares them to traditional credit risk models. The analysis includes discussing data pre-processing methods (e.g., handling missing or aberrant data, discretization of variables) and applying AI techniques that minimize the need for manual variable selection and preprocessing. The paper also analyses AI's influence on risk management tasks, such as credit scoring and fraud detection, and the implications of big data for enhancing credit assessments.

Key Findings:

The paper highlights the superiority of AI models, such as neural networks and machine learning algorithms, over traditional statistical models. Compared to the tedious data pre-processing to be done by a statistician to build a model, an AI model's ability to identify patterns and correlations in large datasets reduces the need for manual pre-processing and feature selection. Additionally, AI models also provide immense predictive gains and reduce human bias. For instance, classification tree algorithms automatically group variables and select the most predictive features, enhancing the efficiency and accuracy of credit scoring. Despite significant advantages, though, these AI models suffer due to their "black box" nature, making adopting these models challenging.

C. Critical Analysis of Each Paper

1) **Paper 1: A Study on Credit Scoring Modelling with Different Feature Selection and Machine Learning Approaches**

Strengths:

- **Comprehensive Methodology**
  It employs a systematic comparison of five machine learning classifiers in combination with three feature selection techniques. The author uses multiple evaluation metrics to provide a well-rounded assessment of each model's performance. These metrics capture not just the model's predictive accuracy but also the model's training time.
- **Comparison of Feature Selection Techniques**
  This author goes beyond just comparing classifiers by incorporating multiple feature selection techniques, which play a crucial role in improving model performance and reducing biases. The comparison between Information Gain, Gain Ratio, and Chi-Square gives insight into which feature selection technique works well with which machine learning model.

Weaknesses:

- **Limited Dataset**
  The author uses the Statlog German Credit Dataset, which is relatively small (1000 instances) and static (since 1994) when compared to Home Equity Line of Credit (HELOC) and Lending Club (LC) Datasets. While this may be sufficient for academic benchmarking, it might limit the generalisability of the findings to larger, more diverse, and modern datasets. The dataset's historical nature may also make the finding less relevant in today's financial landscape.
- **Lack of Exploration into The Trade-off Between Interpretability and Accuracy**
  The author does not explore the trade-off between model interpretability and accuracy. Even though Random Forest has exceptional accuracy, it is a black-box model, while Decision Tree (C5.0), which is inherently interpretable, performs only slightly worse in comparison. A deeper discussion on the importance of choosing interpretable machine learning models or using post-hoc XAI techniques to explain black-box models could strengthen the paper.

Relevance:

The paper provides valuable insights into the performance of five different commonly used machine learning models in credit scoring. The extension comparison of different combinations of credit scoring models helped to solidify the foundation to explore the strengths of each classification technique in terms of their interpretability and accuracy.

2) **Paper 2: Explainable AI in Credit Risk Management**

Strengths:

- **Robust Methodology**
  The authors trained multiple machine learning models to evaluate the most dominant features affecting credit risk thoroughly. They also utilized two of the most well-regarded XAI techniques, LIME and SHAP, providing both local and global explanations. This dual approach increases the paper's credibility by ensuring that the models are accurate, interpretable, and, most importantly, explainable.
- **Originality of Research Question**
  The paper uniquely explores LIME and SHAP as complementary tools for improving explainability. Most existing studies focus on one explainability method, making this paper's comparison and integration of multiple frameworks innovative.

Weaknesses:

- **Lack of Focus on Data-Driven Feature Engineering**
  Although the authors took steps to preprocess the data, they mention that feature engineering was minimal. In a real-world setting, domain-specific feature engineering is crucial for improving the performance and explainability of machine learning models. The paper could have been strengthened by incorporating more advanced feature engineering techniques tailored to credit scoring.
- **Generalisability**
  The study uses data from the Lending Club platform, which may not be generalized to other financial institutions or geographic regions. While the methodology is sound, the lack of diversity in the dataset (focused on peer-to-peer lending) could limit the generalisability of the findings to different credit-scoring contexts or larger financial institutions.

Relevance:
This work is a significant contribution to the field of XAI in credit scoring models, especially when considering the regulatory landscape and the need for explainability in high-stakes decisions like loan approvals. The extensive comparison of LIME and SHAP techniques offers practical insights into which method is more suitable under different conditions, aiding in the understanding of model explainability for credit risk.

3) **Paper 3: Explainable AI for Interpretable Credit Scoring**
Strengths:

- **Practical Application**
  The findings of the paper are highly relevant for the real-world integration of AI in credit risk management. By evaluating how multiple XAI techniques can complement an exceptional predictive model to provide both transparency and accuracy, the paper provides a practical insight that can be adopted by financial institutions.
- **Real-World User Testing**
  The authors employ real people, both domain experts and lay humans, to rate and evaluate the explanations generated by XAI techniques. By involving end-users in the evaluation process, the authors ensure that the explanations are not only meeting technical benchmarks, but also practically useful for real users. This enhances the paper's applicability in real-world financial settings.

Weaknesses:

- **Computational Complexity**
  Although the paper uses advanced XAI techniques such as SHAP, all of them are at the higher end of computational cost [cite Shayan's paper], especially when they have to deal with high numbers of features. Not to mention, the authors intend to employ all of them at the same time to generate a multitude of explanations. This computational burden is not addressed in the paper, which could be a practical concern for institutions seeking to implement these methods in real-time decision-making environments.
- **Limited Focus on Bias Mitigation**
  While the paper addresses how XAI can aid in interpretability, it does not delve deeply into the issues of bias and fairness, which are also significant concerns in AI-driven credit scoring models. Given that credit scoring decisions can have profound impacts on individuals' financial lives, more attention could be given to how XAI techniques can help identify and mitigate biases.

Relevance:
This paper is highly relevant to our research as it directly contributes to the ongoing discussion about making AI credit scoring models more transparent and accountable. By employing multiple XAI techniques to generate 360-degree explanations, the paper bridges the gap between technical model development and practical, regulatory, and ethical considerations.

4) **Paper 4: Artificial Intelligence and Bank Credit Analysis: A Review**
Strengths:

- **Comprehensive Coverage of AI in Credit Scoring**
  The paper provides an extensive overview of how AI is applied in credit scoring. It discusses multiple AI techniques, traditional ML models, and more advanced ones like deep learning. Integrating big data into AI-driven credit scoring models is well explored, highlighting how non-traditional data sources (e.g., social media and digital fingerprints) contribute to enhanced credit risk assessment.

- **Significance of Findings on Financial Inclusion**
  This paper contributes significantly to the discussion on how AI models can improve financial inclusion, especially for the underserved population with limited or no credit access. This is paramount, provided that the opacity of these models can be overcome. The field of XAI aims to do exactly that: make complex AI models explainable.

Weaknesses:

- **Limited Empirical Evidence**
  The paper heavily relies on case studies and examples from FinTech companies. However, it lacks empirical evidence or statistical data to back up some of the claims, particularly regarding the effectiveness of AI in directly reducing discrimination or bias in credit scoring. The conclusion seems speculative without concrete data.
- **Over-reliance on Theoretical Discussions**
  Most of the claims about the advantages and drawbacks of AI in credit risk analysis remain theoretical. The paper could benefit from a quantitative analysis to validate the assertions, particularly in the field of XAI.

Relevance:

The paper is highly relevant to the research topic of XAI in credit scoring. It addresses the role of AI in credit risk analysis, focusing on the use of machine learning models to predict creditworthiness. While the paper doesn't focus solely on XAI, it provides valuable insights into the need for explainable AI (XAI) in financial models due to the complex nature of AI-driven credit decisions and the potential ethical and legal challenges.

D. Comparative Analysis and Synthesis

Themes and Patterns:

The commonalities that can be found across the papers reviewed are:

- **Importance of Model Interpretability in Credit Scoring**
  Across the papers, the need for interpretability in AI-driven credit scoring models is repeatedly stressed. This is important for meeting regulatory requirements and ensuring fair and transparent credit scoring for all. The literature consistently points out the importance of having understandable models, either by design (such as Decision Trees) or through post-hoc explanations (SHAP, LIME, etc.).
- **Fairness and Bias Concerns**
  A recurring theme across the reviewed papers concerns bias and overfitting in AI-driven credit scoring models. Several papers address the potential for AI to exacerbate existing inequalities, especially for marginalized groups. Even if regulated characteristics like gender, race, or religious affiliation are omitted, the intersection of certain variables, such as applicants who shop online at Website X and communicate with Messenger App Y, may serve as a proxy for a particular race, possibly resulting in indirect biases [5].

Gaps in the Literature:

The questions that are inadequately addressed in the papers reviewed are:

- **Trade-off Between Accuracy and Interpretability**
  While all papers mentioned the need for XAI to make prediction models interpretable, they fail to compare the trade-off between accuracy and interpretability across different classifiers or models. For instance, shallow classifiers like Decision Trees are intrinsically interpretable and do not require additional post-hoc XAI techniques to explain their decision. However, such classifiers may or may not fall short regarding their predictive performance or fair better in terms of their training time.
- **Bias Mitigation through XAI**
  Although there is considerable discussion around identifying bias through XAI, none of the studies focuses on how XAI techniques can be used to mitigate bias during the decision-making process actively. Instead, they are more focused on producing consistent and understandable explanations.

Trends:

The emerging trends shown across the papers reviewed are:

- **Use of Real-Time and Evolving Data**
  The use of datasets from real and currently standing credit lending institutions such as Lending Club and Home Equity Line of Credit is common across the papers. This trend aids in ensuring the models are up to date with new patterns in borrower behavior and regulatory changes, which is particularly important in volatile financial markets.
- **Use of Complementary XAI Techniques**
  It is common to see multiple XAI techniques, including feature selections, intrinsically interpretable models like Decision Trees, and post-hoc model-agnostic methods like SHAP, employed to complement one another and further improve the interpretability of credit scoring models. The emerging trend towards holistic explainability that is useful for diverse people also aided this trend.

Synthesis:

In summary, the common themes across these papers highlight the importance of interpretable credit scoring models as financial institutions increasingly adopt AI-driven models. They also shed light on the dual challenge of maintaining accuracy and transparency. With recent regulatory changes, XAI techniques such as SHAP, LIME, and Anchors are seen as essential tools to ensure that these models and their decisions remain interpretable and compliant. Additionally, the user-centric evaluations proposed by [cite Jun's paper] and the integration of alternative data such as education [cite Upstart] reflect a growing understanding that XAI needs to not only tick off technical checkboxes but also be practical and accessible to end users.

III. Discussion

Evaluation of the Literature:

The reviewed literature collectively presents a strong foundation for understanding the integration of AI and XAI into credit scoring models. Papers 2 and 3 provide a broad overview of AI's impact on credit risk management, outlining the advantages AI brings in terms of predictive accuracy and efficiency over traditional statistical models. They also touch on the ethical, legal, and regulatory concerns, which are crucial considerations when implementing AI in sensitive areas like credit scoring. However, these papers primarily focus on AI's potential and challenges rather than offering a deep dive into XAI methods specifically.

Papers 1 and 4 align more with the research's core focus: XAI in credit scoring. These papers contribute significantly to answering RQ1 by discussing specific XAI techniques like LIME, SHAP, GIRP+SHAP, Anchors, and ProtoDash. Paper 3 particularly highlights how LIME and SHAP can provide local and global explanations of machine learning models, addressing the issue of AI's "black box" nature. Paper 4 expands on this by proposing a robust model that incorporates multiple XAI methods, offering explanations for various stakeholders in different formats, which enhances both transparency and usability.

Together, these papers ultimately contribute to RQ2, comparing the effectiveness of different XAI techniques in explaining credit scoring decisions. Evaluating these methods' limitations (e.g., SHAP's computational complexity and LIME's scope) is especially valuable, as it highlights the practical challenges of implementing XAI in real-world settings.

Gaps in the literature:

- **Scalability of XAI Methods**
  The papers underline the complexity of techniques like SHAP, particularly for larger datasets, but do not provide sufficient research into scalable XAI solutions. Credit scoring models are bound to involve increasingly vast amounts of data, and thus, understanding how to scale XAI methods efficiently is essential.
- **Lack of Real-World Case Studies on XAI in Credit Scoring**
  The literature relies heavily on theoretical discussions and dataset-driven models. However, it doesn't dive deeply into how XAI is currently being implemented by banks and financial institutions in real-world credit scoring systems on a large scale. More empirical studies that involve real-world applications of XAI in banks would help validate the efficacy of these methods.

Implications for Future Research:

- **Comprehensive Evaluation and Comparison of XAI Techniques**
  Future research should focus on a more extensive comparison of various XAI techniques across different credit scoring models. Studies could involve side-by-side performance evaluations of multiple XAI methods, including newer or less explored techniques, to determine which methods provide the most useful explanations in different contexts (e.g., large datasets and diverse user groups).
- **Scalability and Efficiency of XAI Techniques**
  Research is needed to make XAI methods more scalable and computationally efficient for larger datasets. This could mean optimizing existing methods or developing new ones that balance explainability and efficiency.

IV. Conclusion

Summary of Key Findings:

The literature review highlights the increasing role of AI in credit scoring, specifically how AI models enhance prediction accuracy and efficiency compared to traditional statistical methods. AI-based models like random forests, neural networks, and XGBoost significantly improve predictive power. However, the "black box" or "opaque" nature of these models raises concerns about transparency, fairness, and ethical considerations in financial decision-making.

The review of XAI techniques, such as SHAP, LIME, GIRP+SHAP, Anchors, and ProtoDash, shows how XAI can address the lack of interpretability in AI-driven credit scoring models. XAI methods provide both local and global explanations that help stakeholders and the general public better understand and trust AI decisions. Although these techniques present opportunities for enhanced transparency, their computational complexity, and limited scalability pose practical challenges for large datasets.

Reinforcing the Importance of the Research Topic:

The growing application of XAI in credit scoring is a critical area in computer science. It merges technical advances with crucial societal and financial concerns. As we continue integrating AI into the financial sector, ensuring these models are transparent, fair, and explainable is of utmost importance. This will build trust among consumers and regulators, which is essential for adopting these AI models. The research into XAI techniques not only contributes to the technical evolution of AI-based systems but also addresses legal and regulatory issues. These will have broader implications for responsible AI deployment.

In the context of computer science, XAI in credit scoring is at the intersection of ML, data science, ethics, and human-computer interaction, making it a multidisciplinary research field that is significant for both academia and industry.

Suggestions for Future Research:

- **Real-world implementation and Case Studies**
  Empirical research needs to be conducted on real-world case studies of XAI implementation in banks and financial institutions. Studies should focus on the practical impacts of XAI on credit scoring processes. These include but are not limited to improving customer trust, reducing bias, and ensuring regulatory compliance.
- **User-Centric XAI Design**
  Future research is needed to develop XAI techniques that

cater to different stakeholder needs, including loan officers, regulators, and consumers. Human-centered design approaches should focus on making XAI explanations even more intuitive, actionable, and accessible to non-technical users.

REFERENCES

[1] S. K. Trivedi, "A study on credit scoring modeling with different feature selection and machine learning approaches," *Technology in Society*, vol. 63, p. 101413, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0160791X17302324

[2] B. H. Misheva, J. Osterrieder, A. Hirsa, O. Kulkarni, and S. F. Lin, "Explainable ai in credit risk management," *arXiv preprint arXiv:2103.00949*, 2021.

[3] L. M. Demajo, V. Vella, and A. Dingli, "Explainable ai for interpretable credit scoring," *arXiv preprint arXiv:2012.03749*, 2020.

[4] H. Sadok, F. Sakka, and M. E. H. El Maknouzi, "Artificial intelligence and bank credit analysis: A review," *Cogent Economics & Finance*, vol. 10, no. 1, p. 2023262, 2022.

[5] K. Langenbucher and P. Corcoran, "Responsible ai credit scoring–a lesson from upstart. com," *Digital Finance in Europe: Law, Regulation, and Governance. De Gruyter*, 2022.