A Scalability Study of SHAP and TreeSHAP in Credit Scoring Applications

Group Members:

Chan Jun Kit (1231302583)

Khan Shayan (1231301827)

Muhammad Ameer Rafiqi Bin Mohamad Shahizam (1211106255)

Marcus Chin Wei Hern (1211107284)

Course: Research Methodologies for Computer Science

Assignment 2 - Research Proposal

Date of Submission: September 25, 2024

Executive Summary

The rise of AI in credit scoring has significantly improved predictive accuracy but introduced challenges related to transparency and fairness, largely due to the "black box" nature of many AI models. Explainable AI (XAI) techniques, such as SHapley Additive exPlanations (SHAP), aim to address these challenges by providing insights into model decision-making processes. However, SHAP's high computational complexity and scalability limitations make its application in large-scale financial systems difficult.

This research investigates SHAP's scalability in tree-based ensemble models, specifically Extreme Gradient Boosting (XGBoost), and explores methods to enhance its efficiency. The study focuses on two primary objectives: quantifying how SHAP's scalability degrades with increasing prediction instances and identifying strategies, such as feature selection and TreeSHAP, that may improve scalability.

The methodology includes training two XGBoost models—one with full features and one with Chi-Square-selected features—using a large Lending Club dataset of 2.3 million loan applications. SHAP and TreeSHAP will be applied to both models, and the time complexity and explanation generation rates will be compared to assess scalability.

The outcomes will offer quantitative insights into SHAP's efficiency with large datasets and identify strategies to improve its scalability, which will be beneficial for enhancing transparency in AI-driven credit scoring.

Keywords: Explainable AI (XAI), SHapley Additive exPlanations (SHAP), Scalability, Credit Scoring

Contents

1 Introduction

The rise of artificial intelligence (AI) has revolutionised the financial industry, particularly in credit scoring, where machine learning models have significantly improved the accuracy of credit risk assessments. However, the "black box" nature of many AI models poses challenges related to transparency, fairness, and trust, especially in high-stakes decision-making processes such as loan approvals. As financial institutions increasingly adopt these advanced models, the need for explainable AI (XAI) techniques has become critical to ensure that credit-scoring decisions are understandable, transparent, and compliant with regulatory standards.

Among the various XAI methods, SHapley Additive exPlanations (SHAP) have gained attention for their ability to provide detailed, instance-level explanations of model predictions. Despite its effectiveness, SHAP often faces scalability issues, particularly when applied to tree-based ensemble models with large datasets. This research proposal aims to explore how the scalability of SHAP degrades with increasing data volume and investigates potential strategies, such as feature selection methods and optimised versions of SHAP (TreeSHAP), to improve its efficiency in credit scoring applications.

2 Problem Statement

In the quest to make robust but opaque credit-scoring models explainable and compliant, Explainable AI (XAI) techniques are applied to them. However, these techniques face several challenges:

1. **Computational Complexity:** Many well-known XAI techniques, such as SHAP and LIME, are flagged for their computational complexity. SHAP, for instance, has an exponential time complexity (Misheva et al., 2021). Such a characteristic makes the XAI techniques challenging to scale as the volume of data grows and would not be viable in real-world settings where decisions and explanations are expected to be made swiftly.

2. **Perturbation Process:** Perturbation is a critical process in many XAI techniques, whereby the XAI generates multiple slightly altered versions of a particular input

to observe how the prediction changes and provide actionable explanations. In addition to the XAI techniques' inherent computational complexity, this further inhibits their scalability.

Although many papers acknowledge the difficulty of scaling XAI techniques (Marques-Silva & Ignatiev, 2022), few focus on how to tackle the issue. There is also a lack of quantitative data on how specific XAI techniques are more time-efficient than others. The scope of this research will be limited by only considering the scalability of SHAP on a tree-based, ensemble machine-learning model, specifically Extreme Gradient Boosting (XGBoost). This paper will explore strategies such as reducing the number of features (also known as dimensionality), employing the optimised version (TreeSHAP), or both to improve SHAP's scalability with a growing volume of requests. Equally important, this paper aims to quantify the effectiveness of the strategies in aiding scalability.

3 Research Questions, Hypotheses and Objectives

Research Questions:

1. How does the scalability of SHapley Additive exPlanations (SHAP) degrade as the volume of incoming prediction instances increases in a tree-based, ensemble machine learning model?

2. How well can a feature-selection method or the optimised version of SHAP or both improve the scalability of SHAP in a tree-based, ensemble machine learning model?

Hypothesis:

1. The scalability of SHAP degrades exponentially as the volume of the incoming prediction instances increases in a tree-based, ensemble machine learning model.

2. With a feature-selection method or the optimised version of SHAP or both, SHAP can handle more prediction instances before it reaches its practical capacity.

Objectives:

1. To quantify how the scalability of SHAP degrades exponentially as the volume of the incoming prediction instances increases in a tree-based, ensemble machine learning model.

2. To explore and quantify how a feature-selection method, or the optimised version of SHAP or both, can improve the scalability of SHAP in a tree-based, ensemble machine learning model.

4 Literature Review

The literature review highlights the increasing role of AI in credit scoring, showing how AI models like random forests, neural networks, and XGBoost significantly enhance predictive accuracy compared to traditional statistical methods. However, these models' "black box" nature poses challenges related to transparency, fairness, and ethical considerations in financial decision-making.

To address these challenges, researchers have explored various XAI techniques, such as SHAP, LIME, GIRP + SHAP, Anchors, and ProtoDash, aiming to improve the interpretability of AI-driven credit scoring models (Demajo et al., 2020). These techniques provide both local and global explanations, helping stakeholders understand and trust AI-based decisions in credit scoring. While XAI methods offer opportunities for enhanced transparency, they face computational complexity and scalability issues, particularly when dealing with large datasets Demajo et al. (2020). SHAP, for instance, is particularly limited by exponential time complexity with large datasets (Misheva et al., 2021).

Gaps Identified:

1. **Scalability of XAI Techniques:** The existing literature rarely examines the scalability of XAI methods like SHAP in handling large, real-world datasets, particularly in ensemble machine learning models.

2. **Bias Mitigation and Real-World Application**: Researchers discuss how XAI can identify bias but often overlook how to use XAI to mitigate biases actively

during decision-making processes. Additionally, the literature lacks real-world case studies demonstrating XAI's implementation in financial institutions.

Trends:

1. Use of Real-Time and Evolving Data: Researchers increasingly use dynamic, real-world datasets from institutions like Lending Club, showing the importance of adapting models to current borrower behaviour and regulatory changes (Misheva et al., 2021) (Demajo et al., 2020).

2. Complementary Use of Multiple XAI Techniques: We observe a trend where researchers combine multiple XAI techniques to enhance both interpretability and accuracy in credit scoring models, addressing the diverse needs of stakeholders (Demajo et al., 2020).

Implications for Future Research:

1. Real-World Implementation and Case Studies: Future studies should focus on empirical case studies of XAI implementation in financial institutions, examining practical impacts on customer trust, bias reduction, and regulatory compliance (Sadok et al., 2022).

2. Scalability and Efficiency of XAI Techniques: Researchers should explore how feature-selection methods or optimised versions of SHAP can enhance scalability in large datasets, particularly in ensemble models (Trivedi, 2020).

5 Research Methodology

This project consists of the following main components:

A. Data gathering and preprocessing

B. Model training

C. Testing XAI methods

D. Data collection and analysis

E. Data visualisation

PART A – DATA COLLECTION AND PREPROCESSING

We will be using the Lending Club (LC) Dataset, which includes around 2.3 million loan applications with 145 features of different types (Demajo et al., 2020). Firstly, the data is cleaned by handling any special values, encoding categorical variables into numerical formats and eliminating noisy data and outliers. Next, an analysis of the correlation matrix will be performed on the dataset to remove redundant features. Finally, the data will be split with an 80:20 ratio using stratification into training and testing sets.

PART B – MODEL TRAINING

Two different machine learning models will be trained using Extreme Gradient Boosting (XGBoost) as the base mode:

1. Full Feature XGBoost: This model will use all the features from the dataset without applying any feature selection techniques.

2. Chi-Square Feature Selection XGBoost: We chose the Chi-Square Test as our feature selection technique due to its performance in (Trivedi, 2020). The test ranks features based on relevance, and only the top-ranked features are selected for model training. This model will be referred to as the pruned XGBoost.

Both models will be trained on the pre-processed LC Dataset to ensure comparability using the same hyperparameters. Cross-validation will ensure robustness in model performance.

PART C – TESTING THE SCALABILITY OF SHAP

SHAP and TreeSHAP will generate an explanation for the decisions made by the models. TreeSHAP is an optimised version of SHAP designed specifically for tree-based models like XGBoost (Muschalik et al., 2024). The goal is to evaluate the time taken for SHAP/TreeSHAP to generate explanations under varying conditions. Four scenarios will be tested:

1. SHAP + Full Feature XGBoost

2. SHAP + Pruned XGBoost

3. TreeSHAP + Full Feature XGBoost

4. TreeSHAP + Pruned XGBoost

In each scenario, the selected model will be instructed to generate predictions at an increasing rate of requests per minute. The time taken to generate explanations will be recorded for each minute.

PART D – DATA ANALYSIS

The primary metrics that this paper will analyse are:

1. **Time complexity:** The time taken to generate explanations at each rate of incoming predictions. This will be measured for each scenario.

2. **Scalability:** The system's ability to keep up with explanation generation will be evaluated. The point at which explanation time exceeds acceptable limits will indicate the scalability threshold.

3. **Feature Selection Impact:** The difference in time between models with and without feature selection will be analysed to determine the impact of feature selection on scalability.

PART E – DATA VISUALISATION

All collected data will be visualised to illustrate the time complexity and scalability of SHAP and TreeSHAP. The visualisation will include:

1. **Explanation Time vs. Request Rate:** The plot shows how explanation time changes as the number of incoming prediction instances per minute increases.

2. **Comparison of Methods:** Side-by-side comparisons of the four different combinations of models to highlight the effects of feature selection and optimisation.

3. **Scalability Threshold:** Identifying the point at which each combination becomes computationally expensive or impractical for real-time applications.

6 Research Activities and Milestones

Activity Flowchart:

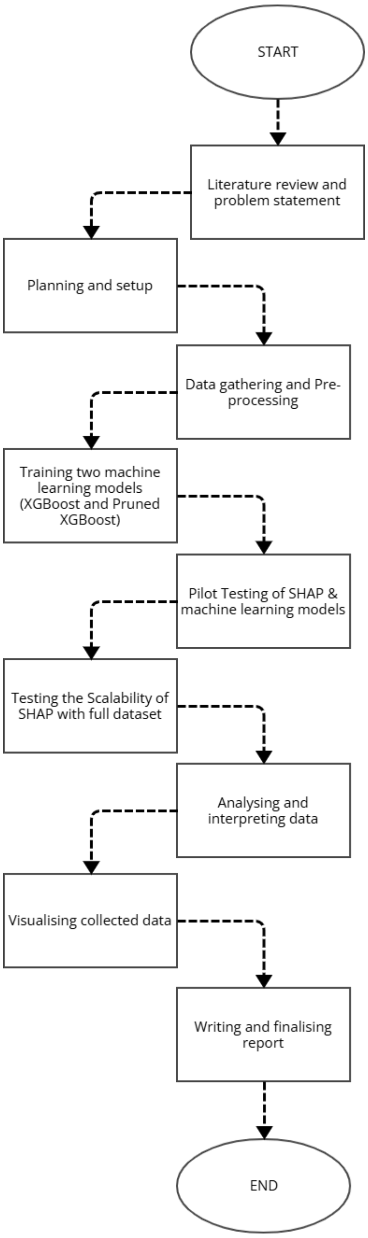
START
Literature review and problem statement
Planning and setup
Data gathering and Pre-processing
Training two machine learning models (XGBoost and Pruned XGBoost)
Pilot Testing of SHAP & machine learning models
Testing the Scalability of SHAP with full dataset
Analysing and interpreting data
Visualising collected data
Writing and finalising report
END

Figure 1: Activity Flowchart

Research Activities:

Activity	**Start Date**	**End Date**
Planning and Setup	01/01/2025	28/02/2025
Data Gathering and Pre-processing	01/03/2025	30/03/2025
Model Training	01/04/2025	16/05/2025
Pilot Testing	17/05/2025	16/06/2025
Testing SHAP with Full Dataset	17/06/2025	17/08/2025
Data Analysis and Integration	18/08/2025	30/09/2025
Report Writing and Publication	01/10/2025	31/12/2025

Table 1: Research Activities

Milestone and Dates:

Activity	**Date**	**Cumulative Completion Percentage (%)**
Completion of Planning and Setup	28/02/2025	10
Completion of data gathering and pre-processing	30/03/2025	20
Completion of model training	16/05/2025	35
Completion of pilot testing	16/06/2025	45
Completion of the full experiment	17/08/2025	60
Completion of data analysis and visualisations	30/09/2025	75
Completion of report and publication	31/12/2025	100

Table 2: Research Activities

Gantt Chart with Research Activities and Milestones:

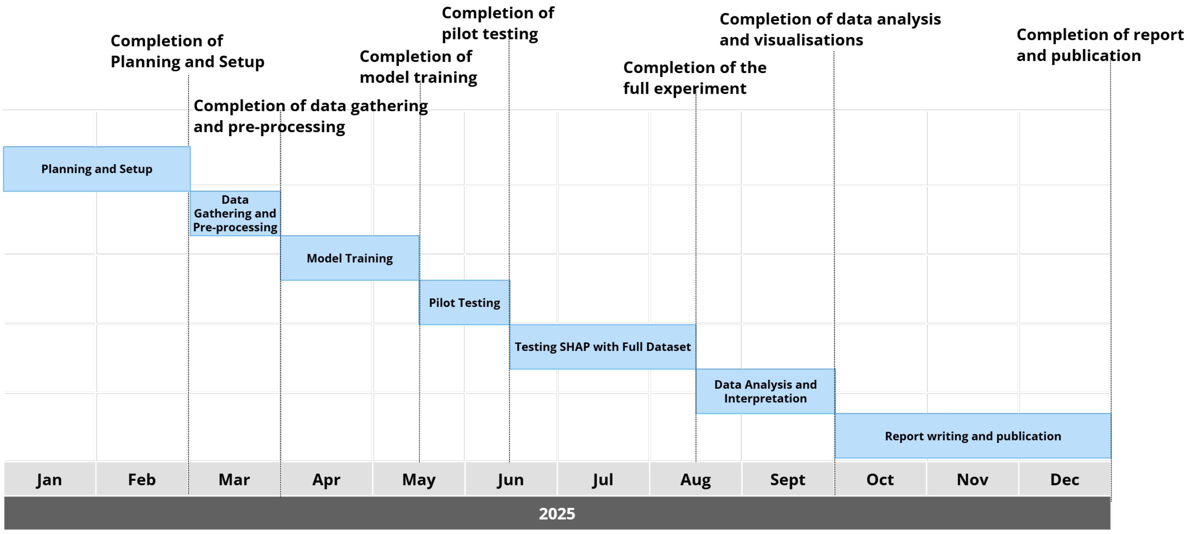

Figure 2: Gantt Chart with Research Activities and Milestones

7 Expected Results and Impact

1. **Novel Theories/Findings:**

   - Provide quantitative insights into SHAP's scalability in tree-based ensemble models, filling a gap in current literature.
   - Identify effective strategies (feature selection, TreeSHAP) to improve SHAP's efficiency in handling high-volume datasets.
   - Offer solid, quantified data to guide future researchers in XAI scalability studies.

2. **Impact on Society, Nation, and Economy:**

   - Enhance the explainability of AI-driven credit decisions, building trust among consumers and regulators.
   - Support financial institutions in meeting transparency requirements, ensuring fair and ethical AI deployment.
   - Improve credit decision accuracy and speed, lowering operational costs and promoting wider access to credit, aiding underserved communities.

- Contribute scalable XAI methods applicable beyond finance, benefiting sectors like healthcare, insurance, and legal systems.

References

Demajo, L. M., Vella, V., & Dingli, A. (2020). Explainable ai for interpretable credit scoring. *arXiv preprint arXiv:2012.03749*.

Marques-Silva, J., & Ignatiev, A. (2022). Delivering trustworthy ai through formal xai. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 36, pp. 12342–12350).

Misheva, B. H., Osterrieder, J., Hirsa, A., Kulkarni, O., & Lin, S. F. (2021). Explainable ai in credit risk management. *arXiv preprint arXiv:2103.00949*.

Muschalik, M., Fumagalli, F., Hammer, B., & Hüllermeier, E. (2024). Beyond treeshap: Efficient computation of any-order shapley interactions for tree ensembles. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 38, pp. 14388–14396).

Sadok, H., Sakka, F., & El Maknouzi, M. E. H. (2022). Artificial intelligence and bank credit analysis: A review. *Cogent Economics & Finance*, *10*(1), 2023262.

Trivedi, S. K. (2020). A study on credit scoring modeling with different feature selection and machine learning approaches. *Technology in Society*, *63*, 101413. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0160791X17302324` doi: https://doi.org/10.1016/j.techsoc.2020.101413

A Appendix A: Member Contributions

Name	**Student ID**	**Contribution (%)**
Chan Jun Kit	1231302583	25
Khan Shayan	1231301827	25
Muhammad Ameer Rafiqi Bin Mohamad Shahizam	1211106255	25
Marcus Chin Wei Hern	1211107284	25
	Total	100