Credit Scorecards in Retail Banking: Enhancing Interpretability through Shapley Values and Evaluating the Effectiveness of Alterna tive Data for Improved Accuracy

Thesis presented for the degree of:

Doctor of Philosophy

in the Department of:

The Graduate School of Business,

University of Cape Town

01 December 2024

By:

Rivalani Willie Hlongwane

Supervisor: Prof Kutlwano Ramaboa

Co-Supervisor: Dr Wilson Mongwe

Declarations

I Rivalani Willie Hlongwane hereby declare that the work on which this thesis is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university. I authorise the University to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

I confirm that I have been granted permission by the University of Cape Town's Doctoral Degrees Board to include the following publications in my thesis, and where co-authorships are involved, my co-authors have agreed that I may include the publications. Furthermore, please note these manuscripts have been published by the PLOS ONE journal:

1) Hlongwane, R., Ramaboa, K. K. K. M., & Mongwe, W. (2024). Enhancing credit scoring accuracy with a comprehensive evaluation of alternative data. *PLoS ONE*, *19*(5), e0303566. https://doi.org/10.1371/journal.pone.0303566 *(Published)*
2) Hlongwane, R., Ramaboa, K. K. K. M., & Mongwe, W. (2024). A novel framework for enhancing transparency in credit scoring: Leveraging Shapley values for interpretable credit scorecards. *PLoS ONE*, *19*(8), 1–20. https://doi.org/10.1371/journal.pone.0308718 *(Published)*

Firstly, in Chapter 4, two sections from the manuscript titled "*A novel framework for enhancing transparency in credit scoring: Leveraging Shapley values for interpretable credit scorecards*" are integrated:

These sections include:

i. "Proposed framework for calculating credit scores": this illustrates a systematic approach using Shapley values to enhance credit scoring models.
ii. "Results and Analysis": this presents the outcomes of credit scoring models and assesses their performance, including the interpretability framework.

Additionally, a section that covers predictor variables used in the study is included, which is not derived from the manuscript.

Lastly, Chapter 5 incorporates a section from the manuscript "*Enhancing Credit Scoring Accuracy with a Comprehensive Evaluation of Alternative Data*", focusing on:

i. "Results and Analysis." This section explores outcomes of credit scoring models with and without alternative data.

Signature of student	RW Hlongwane
Date	01 December 2024

Dedication

This thesis is dedicated to my late father - Hlengani John Hlongwane, whose wisdom and passion for education ignited my own curiosity. Though you are no longer with us, your influence continues to guide me. To my dear mother - Tirani Nellie Hlongwane, your unwavering love and sacrifices have been my source of strength. Thank you for inspiring me to reach for the highest standards. This accomplishment is as much yours as it is mine.

Acknowledgements

I would like to express my heartfelt gratitude to my supervisors, Prof Kutlwano Ramaboa and Dr Wilson Mongwe, for their invaluable time and guidance throughout this journey. Without their mentorship and the methodologies they introduced to me, the success of this journey would not have been possible.

To my beloved wife, Tebogo Hlongwane, your unwavering support has been my rock throughout this journey. I am profoundly grateful to have you by my side.

I extend my deepest appreciation to my family and friends for their encouragement and unwavering faith in me as I pursued my PhD. Your belief in me has been a constant source of motivation.

To Xinyata, Nkhensile, Nthlari, Ntokoto, and all those who have played significant roles in my life, I am grateful for your presence and influence.

Moreover, to my son, Nthlari Hlongwane, your presence and the content you shared have been a source of rejuvenation during the breaks in my research sessions.

Lastly, in loving memory of my siblings who have passed, their impact on my life will always be remembered. Though I may not have had the opportunity to meet or interact with some, their presence is felt through the stories and memories shared within our family: Sipho (1978–1979), Wendy (1980–2014), Abednigo (1984–1984), and *Brother* (1986–1986).

Abstract

This research addresses the dual challenges of improving credit scorecard accuracy and maintaining interpretability. While machine learning algorithms like random forest and eXtreme gradient boosting outperform traditional logistic regression in accuracy, their complex predictor variable representation hinders interpretability. To reconcile this, the study discretizes numerical variables, applies one-hot encoding, and employs Shapley values to derive interpretable credit scores for random forest, eXtreme gradient boosting, light gradient boosting machine, and categorical boosting models. This approach produces credit scorecards that align with industry standards. Additionally, the investigation into the role of alternative data in credit scoring reveals its impact on model accuracy. By analysing unique predictor variables such as an applicant's social circle default status, regional ratings, and local population size, the significance of alternative data is demonstrated. Leveraging the model-X knockoffs framework for predictor variable selection contributes to superior model performance, achieving the highest area under the curve on the Kaggle home credit data.

Table of contents

List of tables

List of figures

List of Abbreviations and Acronyms

AUC - Area Under Curve

CART - Classification and Regression Tree

CatBoost - Categorical Boosting

CDR - Call-Detail Record

DPD - Days Past Due

FCRA - Fair Credit Reporting Act

FICO - Fair, Isaac and Company

GBT - Gradient Boosted Trees

ICT - Information and Communication Technologies

ID3 - Iterative Dichotomies 3

IMF - International Monetary Fund

IoT - Internet of Things

IV - Information Value

LDA - Linear Discriminant Analysis

LightGBM - Light Gradient-Boosting Machine

LR - Logistic Regression

NCA - National Credit Act

NN - Neural Network

OOB - Out of Bag

P2P - Peer-to-Peer

RSA - Republic of South Africa

RF - Random Forest

ROC - Receiver Operating Characteristics

SHAP - Shapley Additive exPlanations

SMS - Short Message Service

SVM - Support Vector Machine

USA - United States of America

WOE - Weight of Evidence

XGBoost - eXtreme Gradient Boosting

Chapter 1 Introduction

1.1 Background

Through products like loans and credit cards, banks provide essential financial services (Cierniak-Emerych et al., 2021). According to the International Monetary Fund (IMF), global household debt reached USD 62 trillion in 2022, encompassing mortgages, credit cards, vehicle finance, and personal loans (International Monetary Fund, 2023). The report underscores the growing necessity of accessible credit as household debt continues to rise.

Banks evaluate credit applications using scoring models designed to minimize losses and maximize profits (Crook et al., 2007; Hand & Henley, 1997). These models leverage predictor variables to assess individuals' creditworthiness, aiding in informed decision-making.

Customers with a high likelihood of default lead to losses, while those with a lower likelihood are referred to as good credit customers. Defaulters, in contrast, are categorized as bad credit risk customers (Hand & Henley, 1997). Credit scoring models are specifically designed to predict and differentiate between potential clients with good credit history and those with less favourable credit histories (Siddiqi, 2016).

The accuracy of these predictions depends on factors such as the type of model or algorithm employed (Lessmann et al., 2015), the quality and relevance of the data used (Roa et al., 2021), and the effectiveness of predictor variables (Yu et al., 2021). Highly accurate credit scoring models consider a comprehensive set of predictor variables, including financial history, income stability,

and debt-to-income ratio, providing a nuanced understanding of an individual's creditworthiness (Roa et al., 2021). This accuracy is paramount for banks in managing risk effectively and making sound credit decisions (Siddiqi, 2016).

In instances where credit scoring models recommend denying credit to an applicant, clear explanations for the rejection are expected by credit regulators (Kelly-Louw, 2007; McCorkell & Smith, 2009). For example, in the Republic of South Africa (RSA), the National Credit Act (NCA) of 2005 mandates that a credit provider must furnish the consumer with reasons for rejecting their loan application (Kelly-Louw, 2007). Similarly, in the United States of America (USA), the Fair Credit Reporting Act (FCRA) of 1970 establishes similar expectations (McCorkell & Smith, 2009). Banks, therefore, tend to prefer interpretable credit scoring models that allow for a transparent understanding of the credit decision-making process. Failure to provide reasons for declining a loan application may result in consequences such as the loss of a trading license or the imposition of fines (McCorkell & Smith, 2009). This underscores the pivotal role played by the interpretability of credit scoring models in ensuring accountability and regulatory compliance within the banking sector.

In this context, banks commonly favour traditional credit models like logistic regression (LR) and linear discriminant analysis (LDA) to develop credit scorecards. These models facilitate the generation of credit scores for each predictor variable, elucidating their significance within the model (Tsagkarakis et al., 2021). The additive nature of LR and LDA models enhances interpreta bility, a vital attribute for their applicability in regulatory frameworks (Tsagkarakis et al., 2021).

Subsequently, banks use the credit scores obtained from LR and LDA models to discern the reasons for rejecting a credit application, with each predictor variable's credit score offering insights into the credit decision-making process (Siddiqi, 2016).

However, there is a growing interest in the literature aimed at enhancing accuracy through advanced machine learning algorithms. One of the comprehensive benchmarking studies conducted by Baesens et al. (2003) sought to understand the impact of these algorithms. The study found that support vector machines (SVM) and neural networks (NN) outperformed LR and LDA in various credit scoring datasets. Lessmann et al. (2015) extended this benchmarking study by identifying emerging trends in credit scoring. The authors compared 41 models across eight real-world credit scoring datasets and found that Random Forest (RF), a tree-based algorithm introduced by Breiman (2001), demonstrated superior predictive power compared to SVM, NN, LR, and LDA.

The research conducted by Chen and Guestrin (2016) introduced an advanced machine learning algorithm based on boosted trees known as eXtreme gradient boosting (XGBoost). Since its inception, XGBoost has consistently demonstrated high accuracy across various credit datasets. For instance, in a study by Munkhdalai et al. (2019), where credit scoring models like LR, NN, SVM, RF, and XGBoost were benchmarked on credit data, XGBoost outperformed other algorithms, exhibiting the highest accuracy. Additionally, other advanced boosted tree algorithms, including light gradient boosting machine (LightGBM) introduced by Ke et al. (2017) and categorical boosting (CatBoost) by Prokhorenkova et al. (2018), have been identified for their superior accuracy in credit scoring applications (Al Daoud, 2019).

Other efforts in literature to enhance the accuracy of credit scoring models involve the use of big data which introduces an extensive array of previously unexplored data sources (Lessmann et al., 2015). Data from these unexplored data sources is referred to as alternative data (Djeundje et al., 2021). While credit bureau remains the predominant data source in credit scoring (Roa et al., 2021), alternative data presents an opportunity to supplement this information and further improve prediction accuracy (Óskarsdóttir et al., 2019). Pedro et al. (2015) compared call detail records (CDRs) and credit bureau data, finding that CDRs were more predictive than credit bureau data in credit scoring.

Óskarsdóttir et al. (2019) conducted research to establish the effectiveness of CDRs in credit scoring, concluding that CDRs can either supplement credit bureau data or be used solely to build credit scoring models. Djeundje et al. (2021) conducted their research to gain an understanding of whether psychometric and email predictor variables are useful in discriminating between defaulting and non-defaulting customers, finding that this type of data is predictive of credit risk behaviour.

Despite the interest and the potential that alternative data has in improving credit scoring models, the use of this data type in research is rare, primarily due to privacy concerns that hinder the availability of such data for research purposes (Óskarsdóttir et al., 2019). Furthermore, it is crucial to note that privacy concerns are not limited to alternative data; they also extend to credit bureau data (Hiller & Jones, 2022). This broader privacy challenge hinders the conduct of extensive studies assessing the improvement in credit scoring when incorporating alternative data alongside conventional credit bureau information to enhance the accuracy of scoring models.

Additionally, the advent of big data introduces the challenge of high dimensionality, implying a large number of predictor variables in the data (Yu et al., 2021). Effectively addressing high dimensionality by selecting relevant predictor variables is crucial to avoid suboptimal real-world model performance (Yu et al., 2021).

In the credit data from the study by Al Daoud (2019), variables with missing values were iteratively removed at different proportions, and the remaining predictor variables were ranked using the gain-based feature importance. The gain-based feature importance is commonly used in credit scoring to identify predictive variables in the data (Shi et al., 2019). Yu et al. (2021) proposed a framework for identifying and removing redundant predictor variables in credit scoring data and successfully used this framework to reduce the dimensionality of the data.

Despite various efforts to improve accuracy in credit scoring, specifically by introducing advanced machine learning algorithms, the implementation of these algorithms remains uncommon due to concerns about meeting regulatory interpretability requirements (Alonso-Robisco & Carbó, 2022). In response to this challenge, researchers have explored the Shapley Additive exPlanations (SHAP) framework (for example, Bracke et al., 2019; Bueff et al., 2022; Bussmann et al., 2020). SHAP offers a way to explain these complex models in a human-understandable way. While these studies demonstrate the effectiveness of the SHAP framework, they have not fully aligned their explanations with the specific formats and conventions credit practitioners are accustomed to, such as those outlined in Siddiqi (2016).

This study contributes to the field in several meaningful ways. Firstly, it addresses the complexities of credit scoring by introducing a framework that enhances interpretability for advanced machine

learning algorithms. This research highlights the importance of aligning interpretability frameworks with the practices familiar to credit practitioners at different stages of credit assessment. Additionally, the study challenges the current disconnect between interpretability frameworks proposed in prior research and those commonly used by credit professionals. Secondly, the research contends with the prevailing belief that advanced machine learning algorithms, despite their superior accuracy, face challenges in adoption due to non-compliance with interpretability regulations (Alonso-Robisco & Carbó, 2022).

By introducing a framework that uses Shapley values to represent predictor variables as credit scores, this study bridges the gap between accuracy and interpretability, thus fostering the adoption of advanced machine learning algorithms in credit scoring. Lastly, the research acknowledges the pivotal role of alternative data in refining credit scoring models' accuracy (Óskarsdóttir et al., 2019). However, it recognizes the constraints imposed by privacy concerns, emphasizing the need for a balanced approach to data utilization in the context of credit assessment. In essence, this study challenges existing norms and contributes to advancing the understanding and application of advanced machine learning algorithms in credit scoring.

1.2 Problem statement

When individuals seek credit, they initiate the process by submitting a credit application to a bank, triggering the bank's use of credit scoring models to evaluate creditworthiness and make lending decisions (Siddiqi, 2016).

The objective of these credit scoring models is to be highly accurate, ensuring that banks effectively manage risks by minimizing losses and maximizing profits from the credit extended to consumers (Hand & Henley, 1997; Lessmann et al., 2015).

Credit scoring accuracy depends on the model choice (Lessmann et al., 2015), data quality (Óskarsdóttir et al., 2019), and predictor variable selection (Yu et al., 2021). Moreover, credit regulators expect a transparent and clear explanation for the rejection of a credit application (Kelly-Louw, 2007; McCorkell & Smith, 2009). Banks favour traditional models like LR and LDA for their ability to provide clear credit decision reasons (Tsagkarakis et al., 2021).

This regulatory focus on transparency, however, presents a challenge of the widespread adoption of advanced machine learning algorithms by banks (Alonso-Robisco & Carbó, 2022), despite their well-documented superior accuracy compared to traditional algorithms (Chen & Guestrin, 2016; Lessmann et al., 2015; Munkhdalai et al., 2019). A key factor contributing to this reluctance is the perceived lack of interpretability in these models, which falls short of meeting regulatory expectations (Tsagkarakis et al., 2021). Studies like Bracke et al. (2019) and Bueff et al. (2022), have explored making advanced credit scoring more interpretable. However, these frameworks for interpretability differ from the practical approaches outlined in Siddiqi (2016).

Additionally, the choice of data also plays a crucial role in the accuracy of credit scoring models. Previous research, such as, Djeundje et al. (2021) and Óskarsdóttir et al. (2019), has demonstrated that alternative data has the capacity to enhance model accuracy. Nevertheless, studies highlighting the significance of alternative data are scarce, primarily because the data is often unavailable, given the privacy concerns associated with it (Hiller & Jones, 2022).

Firstly, there is a noticeable gap in the literature regarding the provision of interpretability for advanced machine learning algorithms in a format aligned with credit practice. To address this gap, this study aims to introduce a Shapley values-based framework for RF, XGBoost, LightGBM, and CatBoost models to represent predictor variables as credit scores. The selection of these algorithms is guided by prior studies (for example, Al Daoud, 2019; Ma et al., 2018; Yao et al., 2022), which consistently demonstrated their superior predictive accuracy in credit scoring.

Closing this gap is significant, as the study aims to achieve interpretability of advanced machine learning algorithms. Therefore, this will empower banks to adopt advanced machine learning algorithms with superior accuracy and interpretability, ensuring compliance with existing credit regulations.

Secondly, this study aims to contribute to the existing literature by highlighting the potential of alternative data to enhance credit scoring models.

This comparison includes traditional data, such as credit bureau data (Roa et al., 2021), alongside factors like city rating, employment location, number of children, and insights from the customer's social circle regarding credit account defaults – a dimension not previously explored. To achieve this, the study utilizes the Kaggle home credit dataset provided by the Home Credit Group. This dataset includes application data featuring essential customer details such as age and income (Al Daoud, 2019). The significance of conducting this comprehensive comparison lies in its contribution to advancing the understanding of credit scoring models.

Lastly, the Kaggle Home Credit dataset is made up of a high number of predictor variables. To perform predictor variable selection, this study employs the model-X knockoffs framework. The

model-X knockoffs framework is an approach for predictor variable selection in high-dimensional data (Candès et al., 2018). Although this framework has demonstrated effectiveness in previous studies, such as those conducted by Fu et al. (2022), He et al. (2021), and Shen et al. (2019), where it was applied to data with high dimensionality, its utilization in credit scoring remains unexplored. The selection of predictor variables is crucial in this research due to the reliance on high-dimensional Kaggle Home Credit data, as highlighted by Yu et al. (2021).

1.3 Research objectives

This study aims to enhance both the accuracy and interpretability of credit scoring models by addressing the limitations of traditional methods, such as reduced predictive power and limited transparency. Key objectives include the following:

1. To develop a comprehensive framework that enhances the interpretability of advanced machine learning algorithms, with a specific focus on tree-based models such as RF, XGBoost, LightGBM, and CatBoost.

2. To use the Shapley Additive exPlanations (SHAP) framework to derive credit scores, following the methodology outlined by Siddiqi (2016), with the goal of providing interpretability for tree-based credit scorecards.

3. To apply the model-X knockoffs framework to high-dimensional credit scoring data and demonstrate its effectiveness in predictor variable selection.

4. To investigate the impact of including alternative data on the accuracy of credit scorecards, employing various statistical methods to assess the additive value of alternative data.

1.4 Research questions

This research aims to answer the following research question:

1. How does the accuracy of tree-based algorithms compare to that of logistic regression in credit scoring?
2. In what ways can Shapley values be effectively leveraged to represent predictor variables in the context of credit scoring, ensuring a transparent and understandable decision-making process?
3. How does the model-X knockoffs framework improve predictor variable selection in high-dimensional credit scoring data?
4. What impact does incorporating alternative data have on the accuracy of credit scoring models?

1.5 Structure of the thesis

This thesis is structured as follows:

Chapter 1 will introduce the research topic, providing background information and outlining the research objective, problem statement, as well as the research questions and hypotheses. This chapter lays the foundation for the study.

Chapter 2 the literature review will delve into credit scoring models, examining their accuracy and the interpretability. This section aims to provide a comprehensive understanding of existing

research and knowledge on credit scoring within the context of model accuracy and interpretability.

Chapter 3 will detail the research methodology employed, encompassing the approach, design, and methods for data analysis. This chapter provides an overview of the research process, including data preprocessing, feature engineering, model development, and evaluation methods, detailing the steps taken to achieve the research objectives.

Chapter 4 will delve into the first manuscript published in the PLOS ONE journal. The focus will be on the proposed interpretability framework and the results and analysis related to the interpretability of credit scoring models as presented in the manuscript.

Chapter 5 will examine the second manuscript published in the PLOS ONE journal, specifically focusing on presenting the results and analysis of the contributions of alternative data in credit scoring.

Chapter 6 will provide the discussion of the results, a comparison with literature, addressing research questions and hypothesis, and limitations of the study.

Chapter 7 will present the overall research conclusions, the theoretical and practical contributions of the study, and the recommendations for future research.

Chapter 2 Literature Review

2.1 Introduction

In this chapter, a review of the literature is presented, focusing on the dual objectives of accuracy and interpretability in credit scoring models. This exploration examines research methodologies aimed at balancing accuracy and interpretability. The chapter also explores how integrating alternative data sources can enhance credit scoring accuracy. This chapter evaluates credit scoring literature to outline advancements, challenges, and opportunities for improving models.

2.2 Historical development of classic credit scoring models

Research in credit scoring gained traction in the 1960s, with studies like Altman (1968) demonstrating the use of linear discriminant analysis (LDA). In later decades, research shifted focus to logistic regression (LR) models, emphasizing their superior accuracy (Fischer & Moore 1986). The interpretability and predictive power of LR have since made it widely adopted in banking. This continued focus on accuracy drives contemporary studies exploring more advanced techniques (Barboza et al., 2017; Guo et al., 2019; Gurný & Gurný, 2013; Jones et al., 2017; Tang et al., 2019; Van Gool et al., 2012; Wei et al., 2019).

Due to this shift away from LDA in both practical applications and recent scholarship, this study primarily focuses its attention on LR concerning classic scoring models.

2.3 Advantages and limitations of classic credit scoring

Up to now, existing studies consistently highlight the LR algorithm as a benchmark for assessing the performance of novel credit scoring models (Baesens et al., 2003; Lessmann et al., 2015). Its enduring popularity in both practical applications and research is attributed to its capacity to generate accurate and interpretable predictions (Wei et al., 2019). In real-world scenarios, LR remains the preferred choice over other algorithms due to its simplicity, interpretability, transparency, and seamless compliance with regulatory frameworks (Alonso-Robisco & Carbó, 2022).

Unlike advanced algorithms, LR provides transparency by clearly showing how predictor variables influence outcomes (Hertza, 2018). This transparency is crucial in the financial sector, where regulatory compliance and the ability to explain model decisions to stakeholders are paramount. LR's straightforward coefficients help practitioners understand variable impacts, fostering trust in decisions (Alonso-Robisco & Carbó, 2022).

An earlier benchmarking study by Baesens et al. (2003) demonstrated the performance of the LR algorithm compared to advanced machine learning algorithms like neural networks (NN) and support vector machines (SVM). While Finlay (2010) suggested close similarity in performance between LR and NN/SVM, subsequent studies argued for the potential of greater accuracy gains. Lessmann et al. (2015), updating the findings of Baesens et al. (2003), highlighted the need for the credit scoring literature to better incorporate advancements in predictive learning, particularly ensemble methods. Their study introduced RF, an ensemble model, which outperformed NN,

SVM, and LR. This finding, emphasizing the superiority of ensemble models in credit scoring applications, was affirmed in later research by Ala'raj and Abbod (2016) and Zhang et al. (2019).

For example, studies by Lessmann et al. (2015) and Couronné et al. (2018) have spurred interest in machine learning algorithms (Shi et al., 2022), prompting credit practitioners to show a keen interest in advanced algorithms. Furthermore, research indicates that adopting these algorithms empowers credit lenders to utilize previously untapped data sources (Lessmann et al., 2015; Siddiqi, 2016).

This aligns with the era of big data, where research suggests lenders can gain a deeper understanding of their customers (Lessmann et al., 2015; Siddiqi, 2016). However, big data brings complex properties to the data such as multicollinearity, non-linear relationships, and high-dimensional data (Zhang et al., 2016). Previous studies demonstrate that LR models struggle to produce accurate predictions when faced with data exhibiting these properties (Coussement et al., 2010; Jagric et al., 2011; Zhang et al., 2016).

Firstly, high-dimensional data can help credit lenders gain a deeper understanding of customer risk (Moscato et al., 2021). However, previous research has shown that high-dimensional data can reduce the accuracy of LR credit scoring models (Couronné et al., 2018).

Secondly, the non-linear relationship between target (the variable this study aims to predict) and predictor variables makes it difficult for LR to produce accurate models (Coussement et al., 2010; Jagric et al., 2011). While increasing the order of predictor variables can address this issue (Bishop, 2006; Zaidi et al., 2016), there are drawbacks. Such models become more complex and lose interpretability (Zaidi et al., 2016). Furthermore, increasing the order might unexpectedly decrease

accuracy, especially with smaller datasets (Zaidi et al., 2016). Simply discarding non-linear predictor variables, while an option, leads to a loss of potentially valuable information (Zaidi et al., 2016).

Lastly, convergent validity (or multicollinearity) presents a challenge for LR in both research and practical applications (De Jongh et al., 2015). While a significant advantage of LR for banking applications is its ability to provide clear explanations of its predictions, multicollinearity can negatively affect the interpretability of these predictions (De Jongh et al., 2015). In experiments conducted on behalf of a South African bank, de Jongh et al. (2015) investigated the effect of multicollinearity on different sample sizes and their analysis revealed that building credit scorecards on smaller datasets negatively affects the coefficients of LR models.

Previous research (for example, Aidoo et al., 2021; Barboza et al., 2017; Jones et al., 2017) arrived at a similar finding. To address this challenge, variables that have a high correlation with each other are discarded (De Jongh et al., 2015). However, it is important to note that removing predictor variables can result in information loss (Zaidi et al., 2016).

Despite these challenges, LR remains the default algorithm for developing credit scoring models (Siddiqi, 2016). However, the benefits of using both big data and advanced machine learning algorithms cannot be ignored. Especially considering that research findings (for example, Coussement et al., 2010; Couronné et al., 2018; Jagric et al., 2011; Lessmann et al., 2015; Zhang et al., 2016) have shown that advanced machine learning algorithms are more accurate than LR and cope with the highlighted challenges.

2.4 Overview of advanced credit scoring models

2.4.1 Bayesian methods in credit scoring

Bayesian methods are widely recognized for their ability to incorporate prior information and handle uncertainty in predictive modelling (Polanska et al., 2023). Giudici (2001) highlights the effectiveness of Bayesian data mining in credit scoring and benchmarking, emphasizing its dynamic nature in updating predictions with new data. This makes Bayesian models particularly suited for environments where data evolves over time.

2.4.2 Tree-based models

Research in credit risk modelling focuses significantly on improving model accuracy (Lessmann et al., 2015). Research has shown that advanced machine learning algorithms such as random forest (RF), eXtreme gradient boosting (XGBoost), categorical boosting (CatBoost) and light gradient-boosting machine (LightGBM) often outperform classic algorithms such as LR (Jabeur et al., 2021; Lessmann et al., 2015; Son et al., 2019).

RF, XGBoost, CatBoost, and LightGBM are all tree-based models (Jabeur et al., 2021). They utilize numerous non-linear decision trees built by randomly selecting a subset of data to build one tree at a time until there are enough trees to make predictions (Jabeur et al., 2021).

The primary difference between RF and gradient boosting lies in their tree-construction processes. RF employs a method called bagging to construct the trees, while gradient boosting algorithms use boosting. In contrast, gradient boosting sequentially builds decision trees to enhance the

performance of the earlier trees; however, bagging does not consider the previously constructed trees (Jabeur et al., 2021).

Tree-based algorithms use majority voting to make predictions from the individual trees (Tsai et al., 2014; Xia et al., 2018). This method involves considering the predictions of each decision tree, with the final predictions based on the highest number of trees that agree with each other (Tsai et al., 2014; Xia et al., 2018). This approach makes these algorithms superior in making predictions as demonstrated by different studies (for example, Tsai et al., 2014; Wei et al., 2019; Xia et al., 2018).

RF and gradient-boosting algorithms are popular in research (Lessmann et al., 2015; Siddiqi, 2016), but their practical application in credit scoring remains limited. Their prevalence in literature is attributable to their superior prediction accuracy compared to LR (Lessmann et al., 2015). This superiority is due to their ability to effectively handle high-dimensional data, non-linearity, and multicollinearity (Couronné et al., 2018; Jagric et al., 2011; Tomaschek et al., 2018). As Kotsiantis (2013) notes, tree-based models are particularly robust against issues with multicollinearity.

Lessmann et al. (2015) analysed over forty credit scoring research papers published between 2003 and 2014, providing valuable insights into the algorithms used in credit scoring during this period. The study highlighted a wide range of algorithms considered (2-35 per paper). A key finding from this benchmarking effort is that research in credit scoring is dominated by advanced algorithms, with limited interest in further developing LR. Instead, researchers have focused on leveraging advanced models to enhance accuracy. Notably, the study identified tree-based algorithms as

strong candidates for producing highly accurate credit scoring models. Subsequent research, such as studies by Ala'raj and Abbod (2016) and Xia et al. (2020), has supported these findings, reaffirming the predictive superiority of tree-based approaches.

Ala'raj and Abbod (2016) demonstrated that careful tuning of hyperparameters, such as the number of trees and predictor variables, is crucial for optimizing RF model performance. Optimising the best values of the model hyperparameters (for example, number of trees, variables, etc.) leads to the most accurate model (Charilaou & Battat, 2022). The study by Ala'raj and Abbod (2016) found that optimizing the number of trees (around 60) and variables (11-22) significantly improved accuracy across the datasets. Xia et al. (2020) explored a range of 50-500 trees, and a proportion of between 60% and 100% of the total number of predictor variables in the data to find hyperparameters that yield the most accurate RF model. Despite research findings (for example, Ala'raj & Abbod 2016; Xia et al. 2020) demonstrating that RF produces more accurate predictions, its adoption in practice remains limited due to its inability to provide human-understandable predictions (Arrieta et al., 2020).

Chen and Guestrin (2016) conducted a survey of machine learning algorithms used to win popular data science competitions in 2015, found that XGBoost to be the most common algorithm for achieving superior prediction accuracy across domains like credit risk, bankruptcy, and others. Furthermore, Wei et al. (2019) indicated that XGBoost can deal with imbalanced datasets much better than other advanced scoring methods. This is an important aspect, especially considering the premise that most credit datasets are inherently imbalanced (Brown & Mues, 2012). Imbalanced data occurs because of rare instances of defaulters as compared to a higher number of non-defaulters (Brown & Mues, 2012).

Munkhdalai et al. (2019) designed an experiment to benchmark classifiers such as LR, NN, SVM, RF and XGBoost on credit data. XGBoost demonstrated the highest accuracy. Importantly, to bridge the gap between research and practice, they compared their model to FICO scores, the industry standard for consumer risk in the U.S. Their study concluded that XGBoost outperformed FICO (Munkhdalai et al., 2019).

Recent studies, including those by Al Daoud (2019) and Lextrait (2023), demonstrate LightGBM's effectiveness in credit scoring. The study by Al Daoud (2019) applied XGBoost, LightGBM, and CatBoost on the Home Credit Group (2018) data, revealing that LightGBM exhibited the highest accuracy. Similarly, Lextrait (2023) demonstrated that LightGBM surpassed alternative methods, including XGBoost, CatBoost, SVM, and LR, in terms of predictive accuracy. Notably, Lextrait (2023) highlighted that LightGBM demonstrated the fastest training speed compared to other gradient-boosting algorithms such as XGBoost.

Previous research, including studies by Prokhorenkova et al. (2018) and Xia et al. (2020), highlights the superior performance of CatBoost compared to other tree-based methods. Specifically, Xia et al. (2020) demonstrated CatBoost's higher accuracy across diverse credit datasets when benchmarked against LightGBM, XGBoost, LR, SVM, and RF. Similarly, Prokhorenkova et al. (2018) found that CatBoost outperformed XGBoost and LightGBM on various datasets.

These advanced tree-based approaches share a common foundation in their use of ensemble structures, where multiple algorithms are combined into a single scoring model (Xia et al., 2018). Ensemble approaches are broadly categorized into two types:

- Homogeneous: Here, algorithms of the same type are combined. RF, XGBoost, LightGBM, and CatBoost are homogeneous ensembles, each utilizing multiple trees of a single algorithm type.
- Heterogeneous: This approach focuses on integrating diverse algorithms (Xia et al., 2018).

In both homogeneous and heterogenous approaches, the final prediction of the algorithms is done using either, hard or soft voting (Kumar, 2020).

- Hard voting: The class prediction (default or non-default) of each algorithm is considered, and the ensemble prediction is based on the highest number of algorithms producing the same prediction (Kumar, 2020).
- Soft voting: The average probability predictions of the algorithms are calculated, and the output is used to decide the class prediction (default or non-default). Voting is considered an effective method for improving prediction accuracy through combining the advantages of individual algorithms (Mahabub, 2020).

Soft voting is considered to yield better accuracy than hard voting due to the flexibility to assign different weights to individual models (Hernández Santa Cruz, 2021). Additionally, hard voting might result in a stalemate. This occurs when two individual algorithms within the ensemble fail to reach a consensus agreement, thereby generating disparate predictions. In essence, this divergence in predictions can impede the decision-making process and undermine the overall predictive accuracy and reliability of the model, as highlighted by Li et al. (2022).

2.5 Limitations of advanced credit scoring models

The widespread acknowledgement of the benefits of advanced models, such as RF, XGBoost, LightGBM and CatBoost, is evident in their superior prediction accuracy and robust performance on various credit datasets (Ala'raj & Abbod, 2016; Munkhdalai et al., 2019 ; Xia et al., 2020).

Despite their effectiveness, the practical implementation of advanced scoring models remains limited due to interpretability challenges (Hertza, 2018). The primary obstacle is their lack of easily interpretable predictions, a crucial requirement for credit decision-making (Arrieta et al., 2020; Hand & Henley, 1997; Siddiqi, 2016).

Siddiqi (2016) emphasizes, credit practitioners involved in the utilization of credit scorecards may not necessarily possess a technical background to understand the inner working of advanced credit scoring algorithms. For non-technical users, the comprehensibility of credit scorecards is paramount (Siddiqi, 2016). An understanding of the components of a credit scorecard is critical in a business setting (Siddiqi, 2016). This understanding is vital in business settings, where practitioners rely on interpreting model components to improve profitability and reduce credit losses (Hand & Henley, 1997).

2.6 Interpretability of credit scoring models

This section covers the interplay between regulatory mandates, the evolving landscape of credit scoring dynamics, and the imperative for transparent decision-making. As credit lenders explore advanced machine learning algorithms, the need for interpretability takes centre stage.

2.6.1 Significance of interpretability in credit scoring

Credit regulators play a crucial role in society by overseeing credit regulations that impact credit decisions made by lenders (Kelly-Louw, 2007; McCorkell & Smith, 2009). Regulations like South Africa's National Credit Act (NCA) of 2005 and the US's Fair Credit Reporting Act (FCRA) of 1970 highlight the importance of transparency in credit scoring (Hertza, 2018; Kelly-Louw, 2007; McCorkell & Smith, 2009).

The emphasis on transparency becomes increasingly relevant in the current landscape, where credit lenders are increasingly exploring advanced machine learning algorithms (Hertza, 2018). As lenders consider advanced machine learning for credit decisions, interpretability is critical to meet regulatory standards (Hertza, 2018). This research explores the intersection of credit scorecard interpretability, regulatory requirements, and the evolving dynamics in credit scoring, emphasizing the importance of transparent decision-making.

Benchmarking studies like Lessmann et al. (2015), have advocated for the adoption of advanced machine learning algorithms in credit scoring. Chen and Guestrin (2016) and Lessmann et al. (2015) demonstrated the superior predictive power of these algorithms compared to traditional models like LR. Despite their enhanced predictive capabilities, Hertza (2018) highlighted a crucial consideration for credit lenders – whether the predictions generated by advanced machine learning algorithms align with the transparency expectations set by credit regulators.

Arrieta et al. (2020) also emphasize that societal expectations extend beyond predictive accuracy. There is a growing demand for these algorithms to provide human-understandable reasons for their predictions. This presents a major obstacle to the widespread adoption of advanced machine

learning in credit scoring, as their complex models often struggle to meet interpretability demands set by both regulators and practitioners (Arrieta et al., 2020; Wei et al., 2019).

2.6.2 Existing methods for enhancing interpretability

Interpretability of models is usually provided in a form of global and local interpretability (Murdoch et al., 2019). Global interpretability provides insights into the overall behaviour of a model across all instances of the data (Murdoch et al., 2019). In contrast, local interpretability focuses on individual predictions, explaining specific decision for a particular instance (Dieber & Kirrane, 2022). Both approaches are crucial for building trust in machine learning models, particularly in applications requiring high accountability and transparency. This section explores methods that balance these interpretability perspectives, ensuring they address the needs of various stakeholders.

The treatment of predictor variables in credit scoring is crucial for both predictive accuracy and interpretability (Kritzinger & Van Vuuren, 2018). Siddiqi (2016) emphasizes the importance of variables being both predictive and easy to interpret, underlining variable binning as an approach to enhance interpretability. Variable binning, as advocated by Kritzinger and Van Vuuren (2018) and Siddiqi (2016), offers several benefits, including the removal of outliers, addressing non-linear relationships between target and predictor variables, and simplifying the understanding of relationships between predictor and target variables.

Moreover, beyond addressing non-linear relationships, predictor variable binning plays a crucial role in assessing the plausibility of credit scorecards, as demonstrated by Hsieh and Hung (2010). The process significantly contributes to increased accuracy in credit scoring models, thereby

positioning predictor variable binning as a valuable preprocessing step for the development of robust and reliable credit scoring models.

The challenge of providing human-understandable reasons for predictions generated by advanced machine learning algorithms has been a significant barrier to their widespread adoption in practice (Arrieta et al., 2020; Wei et al., 2019). Both societal expectations and regulations in places like the USA and RSA demand that models used in decision-making provide clear reasoning (Hertza, 2018; Kelly-Louw, 2007; McCorkell & Smith, 2009).

Siddiqi (2016) highlights the importance of explaining loan declines, low scores, and high scores to stakeholders. He demonstrated practical methods for representing reasons for declining a loan application, including the calculation of a neutral score and the weighted average of variable input values. However, such a framework for providing adverse reasons is non-existent in the literature for advanced scoring methods.

2.6.3 Shapley values in credit scoring

Lundberg and Lee (2017) introduced the Shapley Additive exPlanations (SHAP) framework, which is rooted in cooperative game theory, to enhance the interpretability of machine learning algorithms. Shapley values, originally developed by Shapley (1953), are designed to fairly allocate the total value of a cooperative game to individual players based on their contributions. In the context of machine learning, these “players” correspond to predictor variables, and the “game” represents the model's predictive output.

Given predictive models such as RF and extreme boosted-trees, SHAP calculates Shapley values by systematically considering all possible combinations of predictor variables. This process

ensures that each variable's contribution to a prediction is isolated and quantified, considering the interdependencies among variables (Lundberg et al., 2020). This makes Shapley values particularly suitable for complex, non-linear models like RF and XGBoost, where variable interactions play a significant role in predictions.

Another key advantage of Shapley values is their theoretical grounding, which guarantees consistency and fairness in attributing importance to features. This makes them especially valuable in credit scoring, where clear and explainable decisions are critical for both practitioners and regulators. For example, Shapley values not only quantify the impact of each variable on a prediction but also align these impacts with human-understandable explanations, making them ideal for use in high-stakes domains like finance and medicine (Samek et al., 2019). Additionally, Hickey et al. (2021) demonstrated how Shapley values can enhance fairness in credit scoring by identifying and mitigating bias from sensitive variables such as age and gender. Their framework links interpretability with fairness metrics like Statistical Parity Difference (SPD), reducing fairness gaps while maintaining strong predictive performance.

Many researchers, (for example, Arcadu et al., 2019; Elshawi et al., 2019; Wang & Gribskov, 2019; Zhou et al., 2020) have adopted SHAP in medical research to provide detailed explanations of complex machine models. Elshawi et al. (2019) indicated that the motivation for using this framework is to help increase the understanding of, and trust in, machine learning algorithms among clinicians. This is because clinicians need more information from the model than a simple binary prediction to support their diagnosis (Elshawi et al., 2019).

In the domain of credit scoring, studies conducted by Bracke et al. (2019), Bueff et al. (2022), and Bussmann et al. (2020) have leveraged the SHAP framework to enhance the interpretability of credit scoring models. For example, Bracke et al. (2019) delved into the intricacies of mortgage defaults, utilizing the SHAP framework to comprehensively understand the key drivers. Their analysis involved computing probabilities and log-odds within both a tree-based gradient boosting model and an LR model. The comparison revealed significant differences in predictions, emphasizing that the tree-based model predicts a higher number of cases at risk of defaulting.

Similarly, Bussmann et al. (2020) explored the explanatory power of the SHAP framework for predictions made by both tree-based gradient boosting and LR models on credit data for small and medium companies. Their findings echoed those of Bracke et al. (2019), highlighting that gradient boosting tends to predict a higher number of defaults. Despite this, the studies produced marginal probabilities of predictor variables, providing insights into each customer's credit risk.

In line with their counterparts, Bueff et al. (2022) developed tree-based gradient boosting and LR models, again affirming the superior predictive capacity of the former. While they utilized the SHAP framework for explanations, their unique approach incorporated counterfactuals, offering a perspective on changing input variables to alter model predictions.

Despite these insights, a notable gap exists. None of these studies directly align the marginal probabilities generated by SHAP with the scores used by practitioners. They also do not clarify how Shapley values could guide credit professionals in identifying specific variables contributing to low scores and rejections. This highlights the need for research bridging the interpretability offered by SHAP with the practical decision-making process.

2.6.4 Comparison of Shapley with other interpretability methods

Various interpretability methods have been proposed to address the challenges of understanding complex machine learning models. Some of the methods that are usually compared to the Shapley values include Local Interpretable Model-agnostic Explanations (LIME) and Partial Dependence Plots (PDP).

LIME explains individual predictions by fitting locally interpretable surrogate models around a specific instance (Ribeiro et al., 2016). This approach is particularly effective for instance-level explanations as it provides an intuitive understanding of how input features influence a single prediction. However, LIME has limitations including inconsistency where repeated experiments for the same instance might differ due to randomness in data perturbation (Dieber & Kirrane, 2022). In addition, it does not offer a global understanding of model behaviour (Dieber & Kirrane, 2022), which can be critical in domains like credit scoring.

In contrast, Shapley values provide consistent explanations derived from cooperative game theory principles, ensuring global fairness and reliability. Moreover, Shapley values offer both local and global interpretability, making them suitable for complex domains requiring transparency and accountability (Lundberg & Lee, 2017).

PDP visualizes the relationship between a predictor variable and the predicted outcome by marginalizing over all other variables (Molnar et al., 2023). While this method is useful for understanding global feature importance, it fails to capture interactions between variables (Molnar et al., 2023). SHAP, on the other hand, considers all possible combinations of variables, making it better suited for models like RF and XGBoost, where feature interactions are prevalent.

2.6.5 Summary

The interpretability of credit scoring models underscores the crucial role of transparency in credit regulations, examples of these include acts like the NCA in South Africa and the FCRA in the United States. The rise of advanced machine learning algorithms in credit scoring poses a challenge in meeting transparency expectations, necessitating interpretability to align practices with regulatory standards. While benchmarking studies advocate for the adoption of advanced machine algorithms, the limited implementation in practical applications stems from their difficulties in providing interpretable predictions. The treatment of predictor variables is highlighted as pivotal for interpretability, with variable binning suggested as an effective method. Existing challenges in offering human-understandable reasons for predictions by advanced models are discussed, emphasizing the regulatory mandate for interpretable models. Studies using SHAP in credit scoring were reviewed, revealing a critical gap in aligning marginal probabilities with credit scores and elucidating the practical application of Shapley values in guiding credit decisions, highlighting the need for further research in this area.

2.7 Performance measures of credit scoring models

To examine the performance of credit scorecards, several studies (for example, Barboza et al., 2017; Guo et al., 2019; Gurný & Gurný, 2013; Tang et al., 2019; Van Gool et al., 2012; Wei et al., 2019) have employed a combination of performance metrics to examine credit scorecard performance. These metrics include overall accuracy, sensitivity (true positive rate), specificity (true negative rate), and area under the curve (AUC).

Overall accuracy is widely used in practice and literature and is familiar to credit practitioners (Lessmann et al., 2015; Siddiqi, 2016). As defined by Siddiqi (2016), it measures the proportion of correctly classified defaults and non-defaults. In practice, improving profitability and reducing credit losses for banks are cited as the major reasons for the pursuit of higher overall accuracy (Hand & Henley, 1997). Studies comparing scoring models often use overall accuracy to determine the best model (for example, Guo et al., 2019 ; Lessmann et al., 2015; Tang et al., 2019).

However, Khemakhem et al. (2018) indicated that overall accuracy should not be used in isolation to make the final decision of which is a better credit scorecard. In typical credit data, the proportion of non-defaulting customers is generally significantly larger than that of defaulting customers, a situation commonly referred to as unbalanced data (Khemakhem et al., 2018). Therefore, overall accuracy might be influenced by one of the classes (Khemakhem et al., 2018).

To address this, most researchers tend to also focus on sensitivity and specificity (Barboza et al., 2017; Gurný & Gurný, 2013). Higher values of sensitivity and specificity are desired (Siddiqi, 2016). The assessment of sensitivity and specificity metrics assists in providing an indication of whether the scorecard is predictive of both non-defaulting and defaulting customers (Khemakhem et al., 2018; Siddiqi, 2016). Credit practitioners utilize both sensitivity and specificity to evaluate the anticipated cost of misclassifying a defaulting customer as non-defaulting (Siddiqi, 2016). However, the assessment of the suitability of credit scorecard is not based only on the overall accuracy or the accuracy of the individual classes (Siddiqi, 2016).

In addition to overall accuracy, sensitivity, and specificity, most credit scoring research evaluates performance using the AUC metric (Lessmann et al., 2015; see also Barboza et al., 2017; Gurný

& Gurný, 2013; Wei et al., 2019). The AUC is popular in research due to its ability to provide an indication of the ability of the credit scorecard to separate non-defaulting and defaulting customers (Lessmann et al., 2015). A higher AUC indicates a model that is better in separating non-defaulting and defaulting customers (Siddiqi, 2016).

However, the AUC metric has limitations (Halligan et al., 2015; Hand, 2009; Lobo et al., 2008). Firstly, it is possible that a poorly fitted credit model can either overestimate or underestimate predictions to show the discrimination power of non-defaulting and defaulting customers (Lobo et al., 2008). Secondly, it is difficult for practitioners to interpret the various performance thresholds that it provides (Halligan et al., 2015). Despite these limitations, the AUC remains popular in practice and research (Lessmann et al., 2015).

Previous studies such as Babaei et al. (2023), McKinney et al. (2020), and Qin and Hotilovac (2008) in various domains including credit scoring have commonly employed the DeLong et al. (1988) test as a robust statistical method for comparing the performance of different predictive models. Originating from the work of DeLong et al. (1988), the test demonstrates effectiveness in evaluating the statistical significance of differences between models, particularly in the context of comparing AUC values.

2.8 Predictor variables selection

To assess the predictive power of variables in credit scoring, practitioners commonly use methods such as Weight of Evidence (WoE), Information Value (IV), and the Gini Index (Kritzinger & Van Vuuren, 2018; Siddiqi, 2016). WoE quantifies the strength of the relationship between a predictor variable and the target variable by comparing the distribution of good and bad outcomes across

categories of the predictor (Wang et al. 2018). IV builds on WoE by providing a single measure to evaluate the predictive strength of a variable, where higher values indicate greater predictive power (Mushava & Murray, 2018). The Gini Index, widely used in machine learning, measures the inequality of a distribution and is often used to assess the discriminatory power of variables in credit scoring models (Manek et al., 2017).

In the study by Al Daoud (2019), an iterative process was employed to remove predictor variables with missing values at various proportions, and subsequent predictor variable rankings were determined. The ranking methodology relied on the Gain metric derived from the LightGBM model, as outlined in the study by Al Daoud (2019). This Gain metric, recognized as a valuable technique for identifying predictive variables, has been highlighted in studies such as that by Shi et al. (2019).

The model-X knockoffs framework serves as a powerful approach for predictor variable selection in high-dimensional data (Candès et al., 2018). Several knockoff methodologies have been proposed, Barber and Candés (2015) introduced fixed-X knockoffs for linear regression models. Subsequently, Dai and Barber (2016) extended the fixed-X knockoffs to group-knockoffs, performing group-wise predictor variable selection. However, these methods are limited to linear models (Zhu & Zhao, 2021). To address this limitation, Romano et al. (2020) proposed a model-X knockoffs framework employing deep generative models, known as deep knockoffs. Deep knockoffs stand out as a robust method for predictor variable selection, not relying on the distribution of the data (Dai et al., 2022). While they effectively mimic the relationships among original predictor variables, deep knockoffs may become infeasible in scenarios where the number of predictor variables exceeds the number of records in the data (Dai et al., 2022).

The predominant focus of research on the application of the model-X knockoffs framework has been in genome-wide association studies, examples of these studies include such as those conducted by Fu et al. (2022), He et al. (2021), and Shen et al. (2019). For instance, Fu et al. (2022) employed the model-X knockoffs technique for predictor variable selection, yielding superior results. Shen et al. (2019), in the context of identifying genes associated with a cure for cancer, utilized the model-X knockoffs framework to control false discoveries, demonstrating its favourable comparison to alternative methods. Furthermore, He et al. (2021) showcased the effectiveness of a model-X knockoffs framework in detecting both rare and common risk variants in whole-genome sequencing. Notably, the absence of the application of this framework in the domain of credit scoring represents a notable gap in the existing literature.

2.9 Hyperparameter tuning

Hyperparameter tuning optimizes the parameters of credit scoring models to enhance accuracy (Xia et al., 2017). Common methods include Grid Search, Random Search, and Bayesian Optimization (Yang & Shami, 2020). While Grid Search and Random Search explore fixed hyperparameter spaces independently, Bayesian Optimization dynamically adjusts hyperparameter selection based on previous evaluations (Yang & Shami, 2020).

Grid Search is known for its simplicity and ease of implementation, ensuring a thorough exploration of the hyperparameter space (Yang & Shami, 2020). However, Grid Search may be inefficient where there are high number of parameters to select from (Yang & Shami, 2020). On the other hand, Random Search mitigates these challenges by introducing randomness into the selection of hyperparameter combinations, proving particularly efficient for expansive search

spaces (Yang & Shami, 2020). Both Grid Search and Random Search share a common drawback - they treat each evaluation as independent, potentially leading to redundant function evaluations and overlooking well-performing regions (Yang & Shami, 2020). In contrast, Bayesian Optimization incorporates surrogate models and acquisition functions to guide the search based on previously observed results (Xia et al., 2017). Unlike methods such as Grid Search and Random Search, which can perform evaluations independently and in parallel, the Bayesian approach relies on the information gained from previous evaluations to guide its search (Yang & Shami, 2020).

In LR models, hyperparameter tuning is a process that involves the selection of the regularization method, the adjustment of the regularization strength, and the specification of the optimization algorithm to align with both the dataset characteristics and the desired model performance (Yang & Shami, 2020). Regularization, a foundational element in this context, introduces a penalty term to the model's objective function, thereby preventing overfitting and promoting generalization by discouraging overly complex models (Yang & Shami, 2020). It is noteworthy that explicit details about this process, especially in the context of credit scoring, are often scarce in research literature. Notably, existing studies, such as the work by Yang and Shami (2020), often provide insights into hyperparameter tuning but for research outside of credit scoring.

In tree-based methodologies such as RF, key parameters encompass the number of trees, tree depth, and the quantity of predictor variables within a tree. Furthermore, models like CatBoost, LightGBM, and XGBoost consider hyperparameters like the learning rate and the fraction of data samples utilized in model training (Xia et al., 2020). Earlier studies, (for example, Xia et al., 2017; Xia et al., 2020), used Bayesian Optimization for hyperparameter tuning in the context of credit

scoring. Significantly, these studies illustrated that their approach surpassed the performance of a majority of individual and homogeneous ensemble benchmark models.

Several other studies, such as those conducted by Liu et al. (2022a), Liu et al. (2022b), and Liu et al. (2023), employed a Grid Search methodology for hyperparameter tuning in RF, LightGBM, and XGBoost algorithms. Their findings demonstrated the effectiveness of their credit scoring models, showcasing notable improvements in accuracy compared to benchmark credit data.

The crucial aspect of hyperparameter tuning involves recognizing the specific hyperparameters to adjust, such as the number of trees, tree depth, learning rate, and so forth (Yang & Shami, 2020). However, numerous studies (for example, Al Daoud, 2019; Yu et al., 2021) do not clearly specify the hyperparameter methodology, if any, they choose to utilize for this tuning process, whether it be Grid Search, Random Search, Bayesian Optimization, or the like.

2.10 Alternative data in credit scoring

One key advantage of advanced machine learning algorithms is their ability to effectively model large, non-linear, and high-dimensional data – a type of data known as big data (Lessmann et al., 2015; Jagric et al., 2011). One of the distinct benefits of big data is that it provides alternative data in credit scoring, which banks have rarely considered previously to develop credit scoring models (Siddiqi, 2016). Some of the examples of alternative data sources include data provided by mobile telecommunication companies, Internet of Things (IoT) devices, wearable devices, social networks, and smartphones (Óskarsdóttir et al., 2019; Wei et al., 2016).

Alternative data offers several distinct advantages over traditional credit data. It captures behavioural, social, and regional patterns that are often absent from conventional datasets such as credit bureau records or income statements (Chen et al. 2022). For instance, social network metrics can reveal trustworthiness through peer interactions, while regional ratings provide insights into the economic stability of a borrower's community. This ability to leverage rich, multidimensional data is particularly valuable in assessing creditworthiness for individuals traditionally excluded from financial systems, such as the unbanked population (Aitken, 2017; Brevoort et al., 2016; Óskarsdóttir et al., 2019). These individuals may lack formal credit records but exhibit financial behaviours captured in mobile phone usage or social media interactions, allowing for a more comprehensive and inclusive approach to credit scoring.

Furthermore, alternative data mitigates limitations inherent in traditional data, such as biases in credit histories or insufficient data for emerging borrowers (Roa et al., 2021). By incorporating diverse data points, it can improve model robustness, enhance predictive accuracy, and support fairer lending practices. Research indicates that this data not only models credit risk effectively but also provides actionable insights for policymakers and practitioners, driving innovation in the financial sector (Berg et al., 2020).

However, a major gap exists in understanding the practical impact of alternative data on credit scoring models and how it might affect the use of traditional methods by banks. This research aims to address this gap, providing insights to inform industry policy. The following sections will explore the impact of alternative data on credit risk modelling, covering various data sources and their application in credit scoring.

2.10.1 Psychometric and email predictor variables

One of the earlier studies that tested the effectiveness of psychometric data in predicting repayment of loans was the work done by Meier and Sprenger (2011). They used empirical data to support behavioural economics literature, where Fehr (2002) had earlier demonstrated that self-control explains the use of credit by customers. Meier and Sprenger (2011), finding that impatience is linked to loan default.

Arráiz et al. (2017) and Klinger et al. (2013) studied the effectiveness of psychometric data in assessing the credit risk of Peruvian entrepreneurs. Both studies utilized the same data, although Arráiz et al. (2017) used a larger sample. The psychometric test takes close to half an hour to complete and has an objective to assess individual attitude, beliefs, and integrity (Arráiz et al., 2017; Klinger et al., 2013). Although Klinger et al. (2013) highlighted the response rates for the psychometric questions, it is not clear whether these predictor variables are more predictive than traditional credit bureau predictor variables. However, they indicated that the psychometric responses provide insights into the ability and willingness of debtors to pay their loans.

Both studies concluded that psychometric data is suitable for assessing individuals lacking a credit history. Importantly, these assessments rely on experienced credit analysts conducting interviews with applicants (Arráiz et al., 2017; Klinger et al., 2013).

Djeundje et al. (2021) conducted their research to gain an understanding of whether psychometric and email predictor variables are useful in discriminating between defaulting and non-defaulting customers. Until the study by Djeundje et al. (2021), email usage was not utilized previously to

predict the default of loan products. This is one of the few studies that focused specifically on the creditworthiness of consumer credit applicants.

Djeundje et al. (2021) provides examples of email usage predictor variables associated with higher risk of default:

1. Proportion of emails sent between 12am and 6am.
2. Proportion of sent or received emails from non-top financial product providers.

Djeundje et al. (2021) found that one of the psychometric predictor variables suggests that credit applicants that prefer funding now than in the future have a higher probability of default. These results agree with the finding by Meier and Sprenger (2011) that impatience is a driver of credit risk.

Djeundje et al. (2021) concluded that the psychometric predictor variables and email usage patterns are viable alternative data to predict credit risk. However, they found that one of the psychometric predictor variables they considered has a non-linear relationship with the probability of default.

However, acquiring psychometric data requires loan providers to conduct extensive interviews with potential customers (Arráiz et al., 2017; Klinger et al., 2013). Conversely, utilizing email data is contingent upon internet accessibility, a limitation highlighted by Billon et al. (2021), particularly in regions where access to information and communication technologies (ICT) services is constrained, predominantly in developing countries.

2.10.2 Social networking predictor variables

Wei et al. (2016) concluded that due to the scarcity of reliable credit history data, alternative data such as social data is an option to make credit available to more people. They termed this credit scoring approach as “social scoring”, where the connections and interactions between individuals within a social network are utilized to assess and assign credit scores to individuals.

The premise of the findings by Wei et al. (2016) is that individuals that have similar credit scores tend to associate with each other to form a social network. Wei et al. (2016) reported that social networks data are receiving interest from established credit scoring firms such as Experian. Wei et al. (2016) acknowledged that these types of scores are subject to manipulation where individuals form ties with a selective network to improve their scores. Conversely, when credit bureau data is employed, an enhancement in credit scores is driven by positive credit behaviour (Siddiqi, 2016).

Wei et al. (2016) focused on developing a theoretical framework using the assumption of homophily to demonstrate the importance of social networking in credit scoring (De Cnudde et al., 2019). Homophily states that people tend to associate with people they perceive to be like them (Óskarsdóttir et al., 2019). Although Wei et al. (2016) showed that social networking predictor variables play a role in credit scoring, their research is not explicit on how comparable these predictor variables are in ranking the probability to default on loans to traditional data sources.

Twitter is a social networking platform where users can engage with each other through text-based messages (Ge et al., 2017). Weibo is a social networking platform that has similar functionality to Twitter (Ge et al., 2017). Ge et al. (2017) studied the effects of Weibo predictor variables such as the number of messages sent, followers, friends, and fans on loan default. In addition, they created

an indicator of whether a customer disclosed their Weibo account or not. Ge et al. (2017) studied the effect of these predictor variables on loan default using LR. This study stands out as one of the few that specifically delves into the credit scoring model parameters associated with social media attributes. These parameters are important since they provide the magnitude of increase in the outcome for each 1-unit increase in the predictor variable (Stoltzfus, 2011).

Gül et al. (2018) opted to focus on the usefulness of Twitter data, with a specific focus on the sentiment of a company to find out if this is useful to produce a credit rating. Sentiment analysis focuses on the analysis and understanding of opinions, attitudes, and emotions towards a company, individual, or topic (Medhat et al., 2014). Ge et al. (2017) did not consider sentiments in their study. Gül et al. (2018) found that even though Twitter data is not as predictive as financial ratios to predict the credit rating of companies, however, it is useful.

The study by Suthanthiradevi et al. (2021) focused on using various algorithms to predict the sentiment of individuals against a bank using Twitter data. To predict the Twit score Suthanthiradevi et al. (2021) combined different algorithms to form an ensemble. They found that the ensemble of LR and RF produced the best accuracy. In addition, they combined the model with a NN model, resulting in the highest accuracy among the algorithms investigated. Furthermore, financial and credit scores contributed to the overall score. Suthanthiradevi et al. (2021) conclude d that data derived from Twitter is predictive of credit behaviour.

Suthanthiradevi et al. (2021) did not follow the conventional methodology of building a predictive model using traditional or advanced machine learning algorithms to produce a credit score. They instead used a linear function to combine both the Twit and credit scores.

Facebook is a social networking platform that allows users to connect with family and friends, share videos and photos (Bacaksiz et al., 2020). De Cnudde et al. (2019) investigated the use of Facebook data to predict the creditworthiness of customers using data provided by LenddoEFL. LenddoEFL is a company that uses alternative data to derive credit scores with a particular focus on customers that have none or limited credit bureau data (De Cnudde et al., 2019). The focus was to understand the types of relationships between Facebook users that influence credit scores. The first type of relationship is ordinary friendship connections between borrowers. The second type of relationship arose from the understanding of individuals that have similar interests and preferences, they referred to these as Look-a-likes (LALs). And lastly, relationships based on individuals that interact with each other, they referred to these as Best Friends Forever (BFFs). The LALs and BFFs define fine grained Facebook relationships between individuals (De Cnudde et al., 2019).

De Cnudde et al. (2019) modelled the social media predictor variables as bipartite graphs (bigraphs). This approach follows that of Wei et al. (2016) who used components from graph theory to model social media predictor variables. De Cnudde et al. (2019) applied an LR model to determine the default of a customer using LALs predictor variables. The algorithm formed an ensemble of LR and the SVM model De Cnudde et al. (2019) where the SVM was applied to oth er predictor variables to extract the probability of default. De Cnudde et al. (2019) concluded that interest-based data such as LALs is more predictive than social network data such as Facebook friendship connections.

Some of the limitations of using data from social networking sites such as Twitter, Weibo, and Facebook are that these networking sites are not accessible in certain countries (Niu et al., 2019).

Empirical data has also shown that some customers decline to provide the handle to their social networking account making it impossible for loan providers to access customers data (Ge et al., 2017). Other customers simply do not have a social networking site account (Niu et al., 2019).

Recent research has demonstrated the value of social networking predictor variables in enhancing credit scoring models. Social media platforms such as Twitter and Facebook provide unique insights into borrower behaviour, trustworthiness, and financial habits, offering data points that go beyond traditional credit metrics. For example, Bayesian network models, as explored by Cerchiello et al. (2017), have been applied to social media data to model complex relationships, such as bank risk contagion, effectively capturing interconnected risks and behaviours. Integrating social networking data into credit scoring frameworks can help financial institutions assess creditworthiness more comprehensively, particularly for individuals with limited or no formal credit histories.

2.10.3 Telecommunication predictor variables

The number of mobile phone users worldwide is close to six billion, and telecommunication service providers are increasingly making data accessible to their data commercial partners and researchers (Ots et al., 2020). The most common use of this type of data is on modelling and predicting the behaviour and personalities of people (De Oliveira et al., 2011; Gathergood, 2012). The understanding of behaviour and personalities of people assist companies to design marketing offers and promotions (Ots et al., 2020).

The earliest study on the use of large telecommunications calls detail records data to predict consumer default was done by Pedro et al. (2015). In telecommunication, a call detail record

(CDR) contains identities relating to where a call or short message service (SMS) originates from, its destination, duration, date, cost, mobile carrier, and details of the tower used to facilitate the communication (Zhu et al., 2011). Pedro et al. (2015) used three algorithms LR, SVM, and Gradient Boosted Trees (GBT) to predict default.

Pedro et al. (2015) extracted consumption, mobility, and social network related data from the CDRs, they describe these predictor variables as follows:

1. Consumption predictor variables, these relate to the frequency of calls and SMSs,
2. Social networks, these relate to the interconnections between customers, and,
3. Mobility relates the location information of customers.

Pedro et al. (2015) found that the GBT model outperformed the other two methods in the study. Their study also found that CDR data performed better than credit bureau data. Pedro et al. (2015) concluded that mobile telecommunication CDRs provide alternative data for credit scoring. Their study emphasised that tree-based advanced machine learning algorithms are more predictive than traditional scoring methods. Pedro et al. (2015) attributed the robustness of advanced machine learning algorithms to their ability to deal with non-linear data.

Agarwal et al. (2018) investigated the use of mobile telecommunication data relating to call duration, originating caller, and the identity of the caller to predict default. Unlike in Pedro et al. (2015), the study in Agarwal et al. (2018) did not use graph theory as part of feature engineering to provide an understanding of interconnections between individuals. Agarwal et al. (2018) instead focused on the socio-behaviour of customers. Agarwal et al. (2018) did not consider traditional

algorithms such as LR, instead, they used the XGBoost model. They concluded that mobile telecommunication provides an alternative data source viable for credit scoring.

Óskarsdóttir et al. (2019) followed a social network approach to gain an understanding of the effect of mobile telecommunication data in predicting default. As indicated in Pedro et al. (2015), this approach uses the graph theory framework to represent interconnections derived from CDR data to build predictive variables. In addition, call duration, number of incoming and outgoing calls, day and time of calls predictor variables were used in the study. Óskarsdóttir et al. (2019) developed LR, DT, and RF models.

Óskarsdóttir et al. (2019) concluded that constructing social networks by using CDR data is beneficial in credit scoring. Óskarsdóttir et al. (2019) found that RF was the best performing model and LR was the least performing model, they attribute this to the non-linearity in the predictor variables their used. This observation is consistent with the conclusion made by Pedro et al. (2015) that LR struggles when the data is non-linear.

In addition, Óskarsdóttir et al. (2019) concluded that CDR data provide the following benefits in credit scoring:

1. It complements existing data sources by improving the accuracy of predicting default,
2. It can be used solely to predict default, and,
3. The interconnections measured from the CDR data prove to be effective in predicting default.

While most researchers focused on using an enormous amount of data to investigate the effect of mobile telecommunication data on predicting default, Ots et al. (2020) took a different approach. Ots et al. (2020) investigated the consequence of this type of data when the sample size is small. The research bench-marked the following five modelling algorithms; LR, DT, RF, SVM, and NN.

Ots et al. (2020) followed a similar approach to Agarwal et al. (2018), they did not use graph theory to formulate predictor variables that focused on the interconnections between individuals. They instead derived predictor variables such as call duration, number of incoming and outgoing SMSs, number of missed, incoming, and outgoing calls. Ots et al. (2020) found that RF produced the best performance when compared to the other models. Although Ots et al. (2020) concluded that telecommunication data is a viable alternative data source in credit scoring, they found that the models on their smaller datasets underperform compared to studies that utilized larger datasets.

Telecommunication data is more readily available than social networking data (Niu et al., 2019). Telecommunication data such as call, and SMS records are more readily available than the alternative data sources explored in this study. However, this data has its challenges, for example, the privacy of users (Óskarsdóttir et al., 2019). Telecommunications companies must consider how this data is shared whilst prioritising the privacy of individuals (Óskarsdóttir et al., 2019).

2.10.4 Network models for credit scoring

Network models have emerged as an innovative approach in credit scoring, particularly in the context of peer-to-peer (P2P) lending. These models leverage network centrality measures to analyse relational data between borrowers and lenders, offering insights beyond traditional financial metrics (Chen et al. 2022). Chen et al. (2022) demonstrated that borrowers with higher

network centrality—measured through metrics such as degree, betweenness, and eigenvector centrality—tend to secure loans at lower interest rates, achieve higher funding success rates, and exhibit lower default probabilities. Similarly, lenders with central positions in the network displayed enhanced confidence in their decisions, resulting in faster and larger investments. These findings underscore the potential of network models in addressing information asymmetry in lending markets.

By integrating network centrality measures into credit scoring models, financial institutions can capture additional behavioural and relational dimensions of creditworthiness (Chen et al. 2022). This approach not only improves the accuracy of credit risk predictions but also enhances financial inclusion by providing a framework for evaluating borrowers in data-scarce environments (Chen et al. 2022).

2.10.5 Summary – alternative data in credit scoring

In the domain of credit scoring, the advent of advanced machine learning algorithms has unlocked the potential of big data, offering alternative data sources that were previously overlooked by traditional credit scoring models. This includes data from mobile telecommunication companies, IoT devices, wearable devices, social networks, and smartphones. Network models have emerged as a complementary approach, utilizing relational data from P2P lending platforms to improve credit risk assessment. By incorporating network centrality measures, such as degree and eigenvector centrality, these models capture behavioural and relational dimensions of creditworthiness, offering insights beyond conventional metrics.

While alternative data has shown promise in effectively modelling credit risk and scoring individuals excluded by conventional methods, there is a significant gap in the literature concerning the impact of these alternative data sources on improving the accuracy of credit scoring models. This research aims to bridge this gap by evaluating how alternative data can enhance the accuracy of credit scoring models within the banking industry, potentially leading to more informed lending decisions.

Social networking predictor variables, derived from platforms like Twitter and Facebook, have emerged as valuable alternative data sources in credit scoring. These variables provide unique behavioural and relational insights that complement traditional credit metrics, particularly for underbanked populations. Studies such as Cerchiello et al. (2017) demonstrate the potential of leveraging social media data through Bayesian network models to capture complex interconnections and enhance predictive accuracy. By integrating social networking data, credit scoring models can achieve greater inclusivity and robustness, addressing gaps in traditional methods while fostering more equitable lending practices.

2.11 Synergies between interpretability and accuracy

The integration of interpretability and accuracy in credit scorecards is an area of growing importance, as highlighted by recent research. Arrieta et al. (2020) emphasize the limited adoption of advanced algorithms in practice due to their interpretability challenges. This underscores the need to strike a balance between accuracy and interpretability to ensure practical utility. Additionally, Siddiqi (2016) stresses the significance of credit scorecards being accessible to non-technical users, emphasizing the importance of interpretability in a business setting.

Alternative data sources have been explored in the literature to enhance the accuracy of credit scoring models. Meier and Sprenger (2011), and the subsequent work by Djeundje et al. (2021) provide insights into the effectiveness of psychometric predictor variables in predicting loan default. These alternative data sources contribute to accuracy while offering interpretable insights into the behavioural aspects affecting creditworthiness. Furthermore, Wei et al. (2016) and Ge et al. (2017) delve into social networking predictor variables, demonstrating the potential for accuracy improvement in credit scoring models. Their studies, alongside de Cnudde et al. (2019) research on Facebook data, shed light on the intricate relationship between social attributes and creditworthiness.

In the context of predictor variables derived from telecommunication data, Pedro et al. (2015) and Óskarsdóttir et al. (2019) have made valuable contributions by showcasing the predictive power of mobile telecommunication data in credit scoring. Their studies emphasize the importance of accuracy in credit risk prediction while highlighting the challenges associated with maintaining interpretability, particularly in dealing with non-linear data. These insights collectively underscore the need for synergies between interpretability and accuracy to navigate the complexities introduced by alternative data sources in credit scoring.

2.12 Gaps in current research

The accuracy gains of advanced machine learning algorithms come with the challenge of interpretability (Arrieta et al., 2020). While much research highlights their accuracy gains, there is limited exploration of methods to achieve both accuracy and interpretability. Siddiqi (2016)

underscores the importance of accessibility for non-technical users in a business setting, pointing towards a need for research addressing the interpretability of complex models.

Alternative data sources (psychometrics, social networks, telecommunications) offer potential accuracy improvements (Meier & Sprenger, 2011; Pedro et al., 2015; Wei et al., 2016). However, their influence on enhancing the comprehensibility of credit scoring models is less understood. Understanding how such data affects the comprehensibility of credit risk assessments is key.

The emerging concept of the model-X framework, particularly deep knockoffs, for predictor variable selection in high-dimensional data, receives limited attention in current literature. There is a notable absence of its use, especially in selecting predictor variables within high-dimensional data to enhance accuracy. Subsequent research should examine the viability of implementing the deep knockoffs framework within credit scoring contexts, with a specific focus on exploring its potential in selecting predictor variables for improved model performance. This exploration may shed light on the efficacy of the deep knockoffs approach in addressing the challenges posed by high-dimensional data in credit scoring applications.

2.13 Research hypothesis

Previous studies, such as Lessmann et al. (2015) and Wei et al. (2019), have demonstrated the superior performance of tree-based models in handling non-linearity and complex data structures compared to logistic regression. Building on these findings, this study seeks to assess whether tree-based algorithms provide a significant improvement in credit scoring accuracy:

- $\boldsymbol{H_0}$**:** There is no significant difference in the accuracy of credit scoring models developed using tree-based algorithms compared to logistic regression credit scoring models.
- $\boldsymbol{H_1}$**:** Credit scoring models developed using tree-based algorithms exhibit higher accuracy than logistic regression credit scoring models.

Transparency is essential for facilitating decision-making within the banking industry, especially as advanced models gain adoption. The lack of transparency in complex models has been a significant barrier to their use in credit scoring (Fritz-Morgenthal et al., 2022; Hertza, 2018). This study hypothesizes that Shapley values can enhance transparency and understanding in decision-making processes:

- $\boldsymbol{H_0}$**:** Shapley values do not contribute to creating a transparent and understandable decision-making process in credit scoring.
- $\boldsymbol{H_1}$**:** Shapley values enhance transparency and understanding in credit scoring.

Alternative data, including psychometric assessments and telecommunications records, has been shown to improve predictive power and increase inclusivity, particularly for unbanked populations (Agarwal et al., 2018; Meier & Sprenger, 2011). This study investigates the impact of alternative data on credit scoring model accuracy:

- $\boldsymbol{H_0}$**:** Including alternative data does not result in a statistically significant improvement in the overall accuracy of credit scoring models.
- $\boldsymbol{H_1}$**:** Including alternative data results in a statistically significant improvement in the overall accuracy of credit scoring models.

The model-X knockoffs framework has been identified as an effective method for handling high-dimensional datasets, addressing multicollinearity, and ensuring robust variable selection (Barber et al., 2020; Dai et al., 2022). This study explores its contribution to variable selection in credit scoring:

- $\boldsymbol{H_0}$**:** The application of the model-X knockoffs framework does not contribute to the effectiveness of predictor variable selection in high-dimensional credit scoring data.
- $\boldsymbol{H_1}$**:** The application of the model-X knockoffs framework contributes to the effectiveness of predictor variable selection in high-dimensional credit scoring data.

2.14 Summary of the literature

The literature chapter offers an in-depth examination of pivotal aspects of credit scoring, placing a particular emphasis on the application of existing models in real-world scenarios. It covers the advantages and limitations inherent in classic credit scoring models, provides an overview and limitations of advanced credit scoring models, discusses the interpretability of credit scoring models, explores performance measures, delves into predictor variable selection, hyperparameter tuning, and investigates alternative data sources. Furthermore, the chapter explores the relationship between interpretability and accuracy in credit scoring models.

The exploration of alternative data sources reveals the increasing importance of big data, specifically data from mobile telecommunication companies, social networks, and wearable devices. Previous studies investigated the effectiveness of psychometric, social networking, and telecommunication predictor variables in improving credit scoring accuracy. The potential of alternative data to include individuals excluded by traditional models is highlighted, and the

chapter identifies gaps in research, calling for a more robust examination of the impact of alternative data on credit scoring models.

The integration of interpretability and accuracy in credit scoring emerges as a critical area. The review highlights the limited attention given to the model-X knockoffs framework in current literature and calls for future research to aid predictor variable selection within high-dimensional data. This chapter serves as a foundation for the subsequent empirical study, identifying gaps and paving the way for a more holistic understanding of credit scoring methodologies.

Chapter 3 Methodology

3.1 Introduction

In this chapter, the study explores diverse methodologies aimed at addressing the research questions. The primary focus lies in examining methods employed in both practical applications and existing literature to construct accurate credit scoring models. These methods utilize various modelling approaches and alternative data sources. Additionally, the study investigates methods geared towards generating interpretable predictions from credit scoring models.

Following a deductive approach (Saunders et al., 2009), this research employs a quantitative study design. A quantitative (deductive) study utilizes quantitative data to establish the causal relationships between variables. In the context of this research, a deductive approach is employed by utilizing quantitative data to develop credit scoring models.

3.2 Data and preprocessing

3.2.1 Data

To address research questions 1 and 2, which focus on developing a framework to enhance the interpretability of advanced machine learning algorithms, specifically tree-based models like RF, XGBoost, LightGBM, and CatBoost, and leveraging Shapley values, this study uses secondary credit card data sourced from Yeh (2009) and Home Credit Group (2018).

1. Yeh (2009): This dataset contains 30,000 credit card loan accounts with a 22.12% default rate. The data tracks monthly payments from April to September 2005 and includes:

- A target variable: “default payment” (Yes = 1, No = 0)
- Predictor variables (23): A mix of credit-related information such as loan amount, demographics (gender, education, etc.), past payment history (April-September 2005), bill statement amounts, and previous payment amounts.

2. Home Credit Group (2018): This dataset includes 356,255 home loan accounts with a 6.974% default rate. It integrates credit bureau data, alternative data, and demographic information. Key features of the dataset include:

- A target variable: Default or non-default on a home loan.
- Predictor variables: A wide range of variables, such as:
  - Credit-related data: Loan amounts, past due accounts, repayment behaviour, and credit scores.
  - Demographics: Age, income.
  - Alternative data: Derived variables such as ratios and time-based features.

The Home Credit Group (2018) dataset is notable for combining traditional credit bureau data and alternative data, which is rare due to privacy concerns (Óskarsdóttir et al., 2019). This combination provides an invaluable opportunity to assess the impact of alternative data on credit scoring.

Both datasets provide a strong foundation for addressing the research questions. Specifically:

- For research questions 1 and 2, the datasets support the development of a framework to enhance interpretability using Shapley values and tree-based algorithms.

- For research questions 3 and 4, the Home Credit Group (2018) dataset is reused to explore predictor variable selection using the model-X knockoffs framework and assess the accuracy impact of alternative data.

To ensure comparability with prior research and specifically to address research questions 3 and 4, this study employs commonly used credit scoring models - XGBoost, LightGBM, and CatBoost - on the Home Credit Group (2018) dataset (Al Daoud, 2019; Tounsi et al., 2020b).

3.2.2 Feature engineering

Waring et al. (2020) described feature engineering as a process of creating new predictor variables from existing data to provide useful insights. Transformation and aggregation of predictor variables are common techniques (Han et al., 2012). Han et al. (2012) provides in-depth illustrations of the most frequently used data aggregations. For both datasets, this research applies mean, summation, maximum, and minimum feature transformations. The process involved using each client record to calculate the mean, standard deviation, sum, maximum, and minimum of numeric predictor variables to create new predictor variables.

For the Yeh (2009) data, aggregating of numeric predictor variables, including repayments of previous months' loans and bill statement amounts, is employed. This resulted in an expansion of predictor variables from 23 to 59. This approach offers a unique contribution, as prior studies using this dataset (Chen et al., 2023; Alam et al., 2020) did not perform feature engineering.

Similarly, for the Home Credit Group (2018) dataset, this study transforms past repayments, loan balances, and credit application counts, increasing variables from 217 to 767. Interestingly, while Tounsi et al. (2020b) employed feature engineering on this dataset and achieved superior

performance, Chen et al. (2019) did not. This comparison highlights the potential impact of feature engineering on credit scoring model accuracy.

3.2.3 Variable binning

A crucial stage in credit scorecard development is variable binning, as highlighted by Siddiqi (2016). Variable binning entails the conversion of a continuous variable into categorical ones. This process offers numerous advantages:

- Interpretability: Bins facilitate a more straightforward interpretation of relationships from the credit practitioner's standpoint.
- Scorecard design: Bins facilitate understanding of variable relationships by practitioners.
- Robustness: Binning improves handling of outliers and non-linearity (Kritzinger & Van Vuuren, 2018).

For this research, predictor variables from the Yeh (2009) data and the data Home Credit Group (2018) dataset are binned when addressing research questions 1 and 2, which focus on interpretability. This aligns with the methodology in Siddiqi (2016), these bins play a crucial role in calculating credit scores based on the input values of predictor variables within the scorecard.

3.2.4 Multicollinearity

Multicollinearity or convergent validity occurs when variables are highly correlated with each other (Siddiqi, 2016). The measure of the level of this relationship is determined through a correlation coefficient (Tomaschek et al., 2018). In banking, multicollinearity is closely guarded to ensure that credit scorecard variables remain explainable (De Jongh et al., 2015). Studies

indicate problematic multicollinearity often occurs when correlation coefficients exceed 0.80 (Judge et al., 1988; Kalnins, 2018) and can distort explanatory variable parameters (Aidoo et al., 2021).

Furthermore, previous studies have indicated that multicollinearity has a more pronounced impact on the performance of the LR algorithm compared to tree-based algorithms such as RF, XGBoost, LightGBM, and CatBoost (Aidoo et al., 2021; Couronné et al., 2018; Coussement et al., 2010; Jagric et al., 2011; Tomaschek et al., 2018; Zhang et al., 2016). Consequently, for the analysis of the Yeh (2009) and Home Credit Group (2018) datasets when addressing research questions 1 and 2, this research constructs an LR credit scoring model, necessitating careful consideration of multicollinearity implications. Conversely, when working with the Home Credit Group (2018) data to address research questions 3 and 4, only tree-based credit scoring models are developed. Hence, the issue of multicollinearity was not a major concern when developing the credit scoring models.

In the data from Yeh (2009), the highest correlation amongst the predictor variables is 0.75 and 0.70 for the Home Credit Group (2018) data, this value is lower than the cut-off threshold suggested in Judge et al. (1988) and Kalnins (2018).

3.2.5 Outliers

This research utilizes numerical variables, where a key challenge is the potential presence of outliers – data points significantly deviating from the rest (Aguinis et al., 2013). Outliers can distort analyses, so their treatment is crucial.

To treat outliers, Aguinis et al. (2013) recommended setting the bottom and top values of all observations in a variable to the value at 2.5% and 97.5%, respectively. This research adopted the methodology proposed by Aguinis et al. (2013) to address outliers in both the Yeh (2009) and Home Credit Group (2018) datasets.

3.2.6 One-hot encoding

Section 3.2.3 highlighted the benefits of binning variables: it improves scorecard interpretability and aids in outlier handling. In line with traditional credit scoring practice, this research bins numeric predictor variables. However, machine learning algorithms such as RF, XGBoost and LightGBM require numeric predictor variables as inputs (Cerda et al., 2018).

To address this, the research employs one-hot encoding to encode the binned variables. One-hot encoding is a popular method for encoding input values of categorical variables (Cerda et al., 2018). Due to its simplicity, one-hot encoding is widely used in model development (Yu et al., 2022).

For example, consider the 'customer income' data illustrated in Table 3-1. If binned into three categories (0-1000, 1000-2000, 2000+), one-hot encoding creates three new variables, as shown in Table 3-2.

Using this representation, the research can then calculate the weight of evidence (WoE) of each binned variable as needed to compute credit scores using Equation (5) from Section 3.3.3.3.

Table 3-1 - Binned customer income

Customer Unique id	Income range
1	0 - 1,000
2	1,000 - 2,000
3	0 - 1,000
4	2,000+
5	1,000 - 2,000
...	...
998	2,000+
999	0 - 1,000

Table 3-2 - One-hot encoding customer income

Customer Unique id	Income_0_to_1000	Income_1000_to_2000	Income_2000plus
1	1	0	0
2	0	1	0
3	1	0	0
4	0	0	1
5	0	1	0
...	...	...	...

Customer Unique id	Income_0_to_1000	Income_1000_to_2000	Income_2000plus
998	0	0	1
999	0	1	0

This research applies one-hot encoding to the predictor variables of the datasets used to address research questions 1 and 2, prioritizing interpretability in the development of credit scoring models.

3.2.7 Missing values

The Home Credit Group (2018) data contains missing values in numerical variables, ranging from 0.00046% to 80.05%. Missingness appears highest in credit bureau variables and lowest in social-related predictor variables (max of 0.33%).

Missing values in numerical variables were replaced with the mean of non-missing values for each variable. This is a common and effective imputation technique (Jenghara et al., 2018). Following the approach of Shi et al. (2002), credit type records were transposed to reduce redundancy, resulting in a long table of credit bureau values for each type.

Records with missing bureau data were imputed with zeros since these customers are genuinely not within the bureau; using the mean here would be misleading.

In the Yeh (2009) data, missing values occur in financial ratio variables created through feature engineering. These include ratios such as the current month's payment divided by the average payment over three months. We imputed these missing values with zero, as this accurately reflects

a lack of recorded activity. The missingness rate for these variables ranged between 4.05% and 43.93%. Other variables in the dataset had no missing data.

3.2.8 Sampling

Sampling is crucial as it splits data into training and validation sets (Xu & Goodacre, 2018). Model validation is an important aspect of model development, as it assists in testing whether the model can generalize or produce similar accuracy when applied to data outside of the training data (Xu & Goodacre, 2018). Following common practice (Thien & Yeo, 2022), this study partitions data into training and validation samples to ensure sufficient data for model building and performance evaluation.

This study employs 5-fold cross-validation and sampling techniques to validate models and assess their generalizability (Trivedi, 2020). Both the Yeh (2009) and Home Credit Group (2018) datasets are sampled to balance class distributions and mitigate potential biases. The datasets are then divided into five subsets, with each subset serving as the validation set once while the others are used for training (Trivedi, 2020). This process is repeated five times, and the performance metrics are averaged to provide a comprehensive estimate. Although computationally intensive, cross-validation achieves a balance between rigor and efficiency (Candès et al., 2018).

3.3 Data analysis methods

This section discusses techniques for analysing data, which forms the foundation of developing credit scoring models. It begins with feature engineering which is utilized to create new predictor variables from the data. The data is then split into training and test sets for evaluation. Statistical

significance tests are utilized to identify redundant predictor variables. Variable importance is assessed using permutation feature importance and Gain metrics, guiding predictor variable selection. Variable binning enhances interpretability, and credit score calculation methods are introduced. The section covers handling multicollinearity, outliers, and missing values. It also details predictor variable selection, employing methodologies like the model-X knockoffs framework.

3.3.1 Statistical significance test

Statistical significance tests assist in identifying redundant predictor variables in the data (Feng et al., 2020). Abowitz and Toole (2010) referred to this type of test as statistical conclusiveness validity. Statistical conclusiveness validity seeks to establish if the relationships between a predictor variable and a target variable are statistically significant (Abowitz & Toole, 2010). A p-value is typically used to either confirm or reject the existence of a relationship between a dependent and independent variable (Thiese et al., 2016). The probability of either the existence or non-existence of such a relationship is provided by the p-value (Thiese et al., 2016). Typically, a p-value below 0.05 indicate strong evidence of a relationship, while those above suggest the opposite (Thiese et al., 2016).

Table 0-1 presents the results of significance tests conducted on the Yeh (2009) data, while Table 0-2 provides the results for the Home Credit Group (2018) dataset. Variables cover financial metrics, demographics, and repayment status. Several predictor variables (those with low p-values) demonstrate statistical significance, highlighting their importance in predicting the outcome. The “rejection” column indicates which variables have a p-value above 0.05 and lack

sufficient evidence to be included in the model. Consequently, 27 variables were excluded from further model development.

For the Home Credit Group (2018) data, statistical tests play a dual role. The model-X knockoffs framework relies on a test to assess copies of predictor variables during the selection process. Additionally, a Wald test (e Silva et al., 2020) was used to determine the significance of 22 alternative variables (Table 5-1). A p-value of 0.05 or lower was used to establish statistical significance.

3.3.2 Variable importance

This section introduces two methodologies for variable importance assessment: permutation feature importance and the Gain metric. These techniques are crucial for selecting predictive variables within credit scoring. The choice between them depends on data size. Permutation importance is well-suited for smaller datasets like Yeh (2009), while the Gain metric is preferable for high-dimensional datasets like Home Credit Group (2018). Understanding the strengths of each method ensures the selection of the most appropriate technique for each dataset.

3.3.2.1 Permutation feature importance

Permutation feature importance is a technique employed to assess the significance of predictor variables (Hooker et al., 2021). This method involves comparing shuffled versions of variables with their original counterparts to evaluate their impact on model performance. By assessing the model's performance with the original variable values and comparing it with the performance when the values are randomly rearranged, the importance of the predictor variable is determined (Hooker et al., 2021).

A decrease in model performance following permutation suggests that the predictor variable plays a pivotal role in the accuracy of the model. Conversely, minimal impact indicates that the predictor variable may not exert significant influence on predictions (Hooker et al., 2021). Permutation feature importance, while effective, may pose computational challenges, especially with large data (Hapfelmeier et al., 2023).

3.3.2.2 Gain metric

Feature importance based on Gain evaluates the importance of predictor variables by assessing their role in reducing impurity during the construction of decision trees (Shi et al., 2019). It is directly calculated during the construction of algorithms like XGBoost, LightGBM, and CatBoost (Ke et al., 2017; Shi et al., 2019). Higher cumulative Gain scores signify a greater degree of importance (Shi et al., 2019). This metric is valuable for both feature selection and understanding which variables drive model accuracy (Chen & Guestrin, 2016).

The Gain for a feature is determined by assessing the reduction in the objective function when incorporating that feature into the decision tree. Chen and Guestrin (2016) expressed the relative importance of predictor variables as follows, where the Gain *G* for a predictor variable *j* is expressed in Equation (1).

$$\mathrm{G}(j) = \left(\frac{\text{Gain in impurity or loss when predictor variable } j \text{ is chosen for a split}}{\text{Total sum of Gain for all predictor variables}}\right) \qquad (1)$$

Equation (1) as expressed by Chen and Guestrin (2016), calculates relative importance measure for predictor variables, where the Gain for a specific predictor variable is calculated as the ratio of

the reduction in impurity or loss when that variable is chosen for a split to the total sum of Gain for all predictor variables.

3.3.2.3 Variable importance assessment and method selection

This research employs different variable importance techniques based on dataset characteristics. For the Yeh (2009) data, with fewer than 100 predictor variables, permutation feature importance is used to rank variable significance (Hooker et al., 2021). This method allows for the identification of potentially redundant variables and is computationally suitable for smaller datasets.

In contrast, the Home Credit Group (2018) data comprises 767 predictor variables, classifying it as high-dimensional according to Yu et al. (2021). Given the characteristics of this data, the Gain metric is deemed suitable for assessing predictor variable importance (Jadhav et al., 2018). Consequently, the Gain metric is employed to rank the importance of predictor variables in the data.

3.3.3 Credit scores

This section delves into the intricacies of calculating credit scores, employing methodologies to evaluate and construct effective credit scorecards in practice. The exploration encompasses critical components such as Offset and Factor Parameters, elucidating their role in scaling credit score points. Weight of Evidence (WOE) emerges as a pivotal measure, quantifying the strength of predictor variables. The credit score derivation process is demystified, incorporating WOE and LR coefficients. Additionally, this section examines how reasons for a low credit score are derived, offering an understanding of the predictor variables influencing low credit scores. This research

calculates credit scores using data from Yeh (2009) and Home Credit Group (2018) to address research questions 1 and 2, which focus on the interpretability of credit scoring models.

3.3.3.1 Offset and Factor parameters

Equations (2) and (3) illustrate how to scale credit score points using Offset and Factor parameters.

Practical meaning: Suppose a business wants a credit score of 200 to represent 20:1 odds of non-default to default and wants these odds to double with every 20-point increase in the score.

Calculating Offset and Factor: To achieve this scaling, the Offset and Factor parameters are calculated as shown in the equations provided by Siddiqi (2016):

$$Factor = \left(\frac{20}{ln\,(2)}\right) \tag{2}$$

$Offset = (200 - Factor(ln\,(20)))$	(3)

The selection of the Offset and Factor parameters in this research was made to enhance the interpretability of the credit score points. Choosing a starting point of 200, along with the specified values for the Offset and Factor, was a deliberate decision to ensure the score points are easily comprehensible. It is important to note that these choices are somewhat arbitrary in the context of the research data, as the specific values that the institution providing the data might use for such parameters are unknown. Nonetheless, the chosen values were tailored to facilitate a clear understanding of the credit scoring system within the scope of this research.

3.3.3.2 Weight of evidence

The WOE is critical in the calculation of points allocated to predictor variables (Siddiqi, 2016). Formally, WOE is derived as follows:

$$WoE = ln\left(\frac{Distribution\ of\ good_i}{Distribution\ of\ bad_i}\right) \quad (4)$$

where $i = 1,2, \dots, m$ is associated with a value of bin i of some predictor variable with m bins. The distribution represents the proportion of either non-default or default for each category of a predictor variable.

3.3.3.3 Credit score calculation

The credit score associated with each bin of a predictor variable is calculated as follows:

$$Score_k = -\left(WOE_i * \beta_i + \frac{\beta_0}{n}\right) * Factor + \frac{Offset}{n} \quad (5)$$

where $k, k = 1,2,..m$ for a variable with m bins, β_i is the coefficient associated with a variable of an LR model, α is the intercept of an LR model. The reader is referred to Section 3.3.3.1 for more information on the Offset and Factor parameters. This research proposes to use $\phi_i, i = 1,2, \dots, M$ in Equation (16) in the place of $\beta_i, i = 1,2, \dots, n$ parameters from Equation (8) given a model that has n variables, where $n = M$. In Siddiqi (2016), the $\beta_i, i = 1,2, \dots, n$ refers to parameters of independent variables of an LR model. Furthermore, the proposal is to use the Shapley values parameter ϕ_0 in the place of the LR β_0.

3.3.3.4 Reasons for a low credit score

When a customer scores below the neutral score on certain predictor variables, those variables are flagged as likely reasons for a credit application decline (Siddiqi, 2016). This approach is commonly used in practice to explain adverse decisions to applicants.

Siddiqi (2016) demonstrated how reasons for declining a loan application are typically represented in practice. Firstly, where the odds of non-default to default are equal. The neutral score is calculated as follows:

$$-\left(\frac{\beta_0}{n} * Factor\right) + \frac{Offset}{n} \tag{6}$$

where β_0 is the intercept of an LR model, n is the number of variables in a scorecard.

Siddiqi (2016) outlines an alternative method for identifying variables contributing to a declined credit application: calculating the weighted average of the input values. To illustrate, consider a hypothetical LR scorecard with ten predictor variables, as shown in Table 3-3.

Table 3-3 - Scorecard points

Predictor Variables	**Score**
$Variable_1$	52
$Variable_2$	19
$Variable_3$	87
	…

$Variable_{10}$	78

To illustrate how an adverse score reason is derived using Table 3-3. Consider a customer with a score of 52 for $Variable_1$. To identify which variables contributed to this low score for this variable, Siddiqi (2016) suggests performing the following:

1. Calculate Weighted Average: For each variable, multiply the input value by its corresponding points, then sum across variables. This weighted average shows the overall contribution of input values to the final score. (See Table 3-4)

2. Identify Adverse Variables: Compare each input value score to this weighted average. Input values scoring lower than the average are flagged as reasons for the low score (Siddiqi, 2016).

Table 3-4 - Variable inputs points

Input values	**Distribution**	**Score**
0 - 10	10%	31
11 - 20	20%	52
21 - 30	40%	71
31 - 40	15%	86
41 - 50	15%	92
Weighted average		**69**

Table 3-4 illustrates how Siddiqi (2016) suggests identifying variables contributing to a low score. Consider a customer with a score of 52 (from Table 3-3). Here is a breakdown:

1. Calculate Weighted Average: Calculate a weighted average (69 in this example) by multiplying the distribution percentage of each input value by its corresponding points (Table 3-4) and summing those products.

2. Identify Adverse Variables: Input values that fall below this weighted average (like 31 and 52 here) are considered to have adversely affected the score (Siddiqi, 2016). In this example, the low score on variable $Variable_1$ (refer to Table 3-3 for details) likely contributes to the overall low score.

This approach is valuable for interpreting scorecards based on logistic regression (Siddiqi, 2016). However, there is a gap in the literature for providing reasons for declines with more advanced models like RF, XGBoost, LightGBM, and CatBoost. This research aims to address this gap by proposing a framework that can make these advanced models more interpretable for practical use.

3.3.4 Predictor variable selection

To improve the efficiency of models, the process of predictor variable selection plays a pivotal role in determining the most suitable variables for model development (Speiser et al., 2019). The careful identification of predictor variables significantly contributes to the accuracy of the resulting models (Speiser et al., 2019). In this section, the methodologies employed in the research are delved into, introducing the model-X knockoffs framework. This framework provides a robust approach to predictor variable selection, ensuring that the chosen variables align optimally with the goals of model development.

3.3.4.1 Model-X knockoffs

The model-X knockoffs framework is a robust method for predictor variable selection. It creates "knockoff" copies of predictor variables, preserving their relationship to the target while controlling the false discovery rate (FDR) – the risk of incorrectly identifying unimportant predictor variables as important (Barber et al., 2020). This is crucial in high-dimensional settings where chance findings become more likely. Deep knockoffs are particularly valuable with hundreds or thousands of potential predictor variables (Romano et al., 2020).

However, the application of Model-X knockoffs, particularly with high-dimensional data, can be computationally expensive (Dai et al., 2022). The process of constructing knockoff copies that satisfy the required properties involves demanding calculations and optimizations. To mitigate this computational burden, careful data pre-processing and dimensionality reduction techniques are essential (Candès et al., 2018).

Romano et al. (2020) introduced deep knockoffs, a model-X approach using sophisticated deep generative models. This method proves especially robust as it does not rely on distributional assumptions and works well with high-dimensional data (Dai et al., 2022). Key steps include:

1. Generate Knockoffs: Deep learning models like GANs learn the dependency structure between the original variables and the target variable. This is used to generate "knockoff" variables.
2. Re-Assess Importance: The algorithm recalculates variable importance with the original and knockoff variables included. This helps distinguish genuinely important variables.

3. Control FDR: By applying a threshold to importance scores, the method controls FDR, reducing false positives (Romano et al., 2020).

3.3.4.2 Selection of predictor variables in the research data

The following is the predictor variable selection process utilized for analysis of the Home Credit Group (2018) data, specifically to address research questions 3 and 4:

1. Identify Collinearity: Following Romano et al. (2020) and using the threshold established by Judge et al. (1988) and Kalnins (2018), variables with a correlation coefficient above 0.7 were grouped together. This resulted in 551 groups.
2. Choose Group Representatives: To reduce redundancy, the approach in Al Daoud (2019) was used. Within each group, a LightGBM model ranked variables by the Gain metric, and the highest-scoring variable became the group representative. This step reduced the number of predictor variables from 767 to 321.
3. Apply Deep Knockoffs: The deep knockoffs method (Romano et al., 2020) further refined the selection, resulting in the final 215 variables used for model construction.

For the Yeh (2009) data, a distinct approach was applied due to its lower dimensionality:

- Initial Importance: Permutation importance (Section 3.3.2.1) and statistical significance tests (Section 3.3.1) reduced the original 59 predictor variables to a final set of 7 variables for modelling.

For the Home Credit Group (2018) data (addressing research questions 1 and 2), the original 756 predictor variables was reduced to a final set of 11 variables for modelling following the approach in Hlongwane et al. (2024b).

3.4 Model performance evaluation

This section focuses on the methods utilized to evaluate the performance of credit scoring models. This study employs misclassification statistics, including the confusion matrix, to assess accuracy at both overall and category-specific levels. Additionally, the area under the curve metric, derived from the receiver operating characteristics curve is introduced.

3.4.1 Accuracy

To assess the efficacy of a credit scoring models, misclassification statistics are employed (Kozodoi et al., 2022). Misclassification can be assessed both overall and within specific categories (default vs. non-default). Overall misclassification encompasses all predictions the model inaccurately made, while category-level analysis provides a more granular view.

A confusion matrix (Table 3-5) is used for category-level misclassification assessment. The predicted probability of default determines the allocation of customers into the four cells in the matrix: true positive, false positive, false negative, and true negative. This matrix allows for quantitative analysis of the relationship between predicted and actual default behaviour (Kozodoi et al., 2022).

Table 3-5 - Confusion matrix

		Predicted	
		Non-default	Default
Actual	Non-default	True negative	False positive

		Predicted	
	Default	False negative	True positive

The true negative rate also referred to as specificity is the accuracy of the model on predicting non-defaulting customers (Trivedi, 2020). Conversely, the true positive rate also referred to as sensitivity is the accuracy of the model on predicting defaulting customers (Trivedi, 2020). The goal is to use the probability of default produced by a scorecard to minimize the rate of false negative and false positive through a probability cut-off (Siddiqi, 2016).

3.4.2 Area under the curve

The area under the curve (AUC) provides a metric for assessing overall scorecard performance (Moscato et al., 2021). The AUC is derived from the receiver operating characteristics (ROC) curve (Figure 3-1), and it is calculated using Equation (7) (see Figure 3-1 for areas A and B):

$$AUC = area\ (A) + area\ (B) \tag{7}$$

The blue line (Figure 3-1) represents the line of equality, forming a 45° angle, while the red line is known as the ROC curve. The combined areas A and B constitute the AUC, a key metric for measuring model performance. The sum of areas A and B serves as an indicator of the effectiveness of the model.

Area A change with the performance of the model and it is described as follows:

1. If the red line (ROC curve) overlaps the blue line, it suggests that the model's performance is akin to randomly selecting and classifying the default status of customers.

2. When the red line (ROC curve) extends above the 45° line, it indicates that the model outperforms random selection in classifying customers.

3. If the red line (ROC curve) covers both areas A and C, the model perfectly predicts the cases of interest.

The ROC curve plot:

- $y - axis$ (True positive rate): The proportion of defaulting customers that have been predicted correctly.

- $x - axis$ (False positive rate): The proportion of non-defaulters the model incorrectly flags as likely to default.

The goal of the model is to increase the true positive rate and decrease the false positive rate. An AUC above 0.5 is favourable as it indicates a model that outperforms a random prediction model (Acosta et al., 2020).

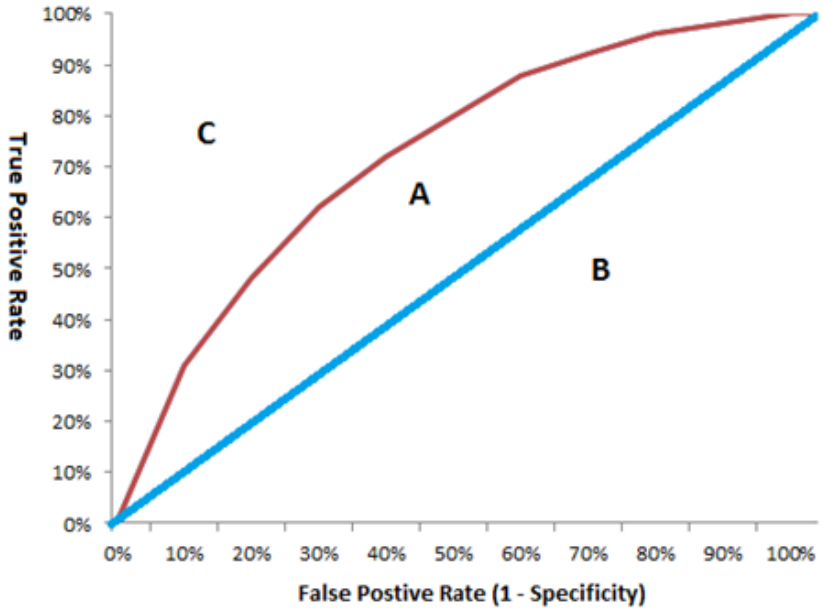

Figure 3-1 - Receiver Operating Characteristics

To gain an understanding of whether differences between two models are statistically significant, DeLong et al. (1988) introduced tests that compare the AUC of models (McKinney et al., 2020). This test helps determine the likelihood that the observed difference in AUC is not random but reflects a genuine improvement in one model's performance.

3.4.3 Model performance assessment in the study

This study employs a variety of metrics to assess credit scoring model performance:

- Overall misclassification: Measures the rate of incorrect predictions.
- Confusion matrix: Provides breakdown of misclassification types (false positives, false negatives), analysed for defaulting and non-defaulting customers.

Analysis for Yeh (2009) and Home Credit Group (2018) datasets (research questions 1 and 2):

- AUC comparison: The area under the ROC curve provides an overall performance measure for models built on this dataset.
- DeLong's test: Used to determine if AUC differences between models are statistically significant (DeLong et al., 1988).

Alternative data impact analysis on Home Credit Group (2018) (research questions 3 and 4):

- Misclassification: Assessed for models constructed both with and without alternative data.
- AUC comparison: Using DeLong's test (DeLong et al., 1988), models with and without alternative data will be compared.

3.5 Modelling approaches

This section outlines the modelling strategies employed in the study. To address research questions 1 and 2, five credit scoring models are developed, including tree-based algorithms such as XGBoost, RF, LightGBM, and CatBoost, along with an LR model. Furthermore, the Shapley values framework is introduced, contributing to the establishment of interpretability foundations for the tree-based algorithms. In addressing research questions 3 and 4, three tree-based credit scoring models, namely, XGBoost, LightGBM, and CatBoost, are implemented.

3.5.1 Logistic regression

In banking, LR is the most common algorithm for building credit scorecards (Siddiqi, 2016). Banks favour LR due to its simplicity, openness and to comply with credit regulations due to its ability to produce transparent predictions (Siddiqi, 2016).

Osborne (2017) defines an LR model as follows:

$$f(x) = ln(odds) = ln\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{i=1}^{M} \beta_i \, x_i \quad (8)$$

where β_0 is the intercept, $\beta_i, i = 1,2, \dots M$ are parameters of the predictor variables $x_i, i = 1,2, \dots M$. Although LR models produce the natural log of odds, probabilities of an event (such as default) can be easily calculated from them. Equation (8) provides the mathematical foundation for this conversion. For a more in-depth explanation of how this model works, see Osborne (2017).

As highlighted by Siddiqi (2016), LR models can produce easy to understand explanations of their predictions. This interpretability is crucial for our research, as it allows us to:

- Develop an LR model for comparison purposes.
- Contrast the explanation styles of the LR model with the tree-based models proposed in this research.

3.5.2 Decision trees

This research utilizes credit scoring models built with RF, LightGBM, XGBoost, and CatBoost. A fundamental understanding of decision trees is important as these algorithms are constructed using them as a foundation.

Decision trees create a series of rules to classify data, predicting a target variable like default status (Siddiqi, 2016). Graphically, they resemble an upside-down tree with decision nodes and branches (J. Han et al., 2012), Figure 3-2 illustrates this structure.

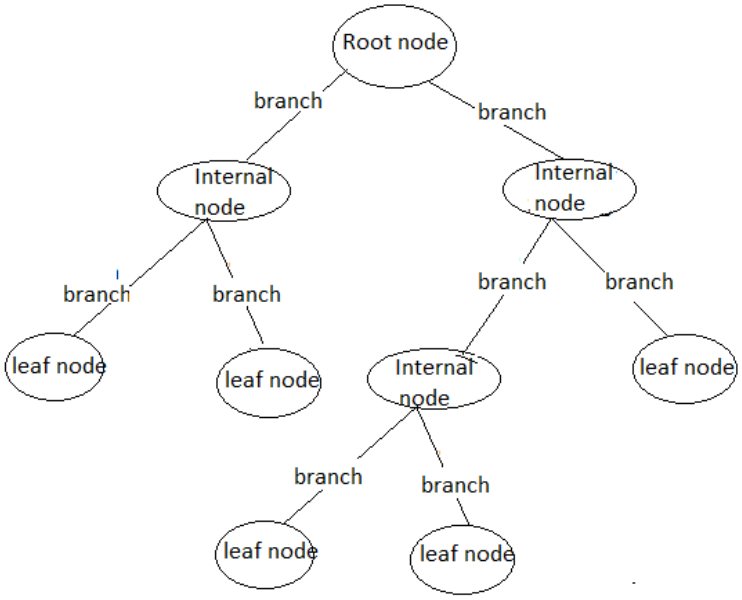

Figure 3-2 - Decision tree

Decision tree components (Han et al., 2012):

1. Root node: The initial decision point, based on the most predictive variable in the data.
2. Branches: Represent thresholds for splitting data based on variable values.
3. Internal nodes: Additional decision points using other variables, further sorting data.
4. Leaf nodes: Represent predictions made by the tree, for example: default or non-default.

The tree is built recursively in a top-down approach using any of the following algorithms:

- Iterative Dichotomiser 3 (ID3): A foundational algorithm. See Quinlan (1986) for details.
- ID3 (C4.5): A successor to ID3, designed to address some of its limitations. See Quinlan (1992).

- Classification and Regression Tree (CART): A versatile algorithm suitable for both categorical and numerical target variables (Han et al., 2012).

The algorithms are not limited to these three, however, these are the most common (Han et al., 2012).

3.5.3 Random forest

The RF model is one of the popular algorithms for credit scoring (Lessmann et al., 2015; Tsai et al., 2014; Wei et al., 2019; Xia et al., 2018). Developed by Breiman (2001), RF builds a "forest" by combining the predictions of multiple decision trees for improved accuracy. The algorithm focuses on the following components to build a forest:

1. Training data, D.
2. Number of instances in the data, N.
3. Number of predictor variables in the data, m.

The following steps describe how an RF model is built:

1. Two-thirds of the data is randomly selected (with replacement) to form a subset data, D_{s1}, through a process called bagging.
2. Approximately $\sqrt{m}$ of the total predictor variables are randomly selected through a process called attribute bagging.
3. Only the predictor variables selected in the previous step are retained in the subset data, D_{s1}. A decision tree is built from this data.

4. The remaining one-third of the data, D_{s2}, that is not selected in the first step is called the Out of Bag (OOB) sample and it is used to validate the decision tree.

At each iteration of developing a forest, the steps are repeated as more decision trees are added into the forest. The process continues until adding more trees does not improve the accuracy of the model.

Figure 3-3 illustrates the key principle of an RF model. The motivation for building multiple decision trees to form a forest is to prevent overfitting the model.

- Individual trees: Each tree makes a prediction based on its specific training data and variable subset.
- Majority voting: The predictions from all trees are combined. The class (e.g., 'default' or 'non-default') receiving the most votes across all trees becomes the final RF prediction (Tsai et al., 2014; Xia et al., 2018).

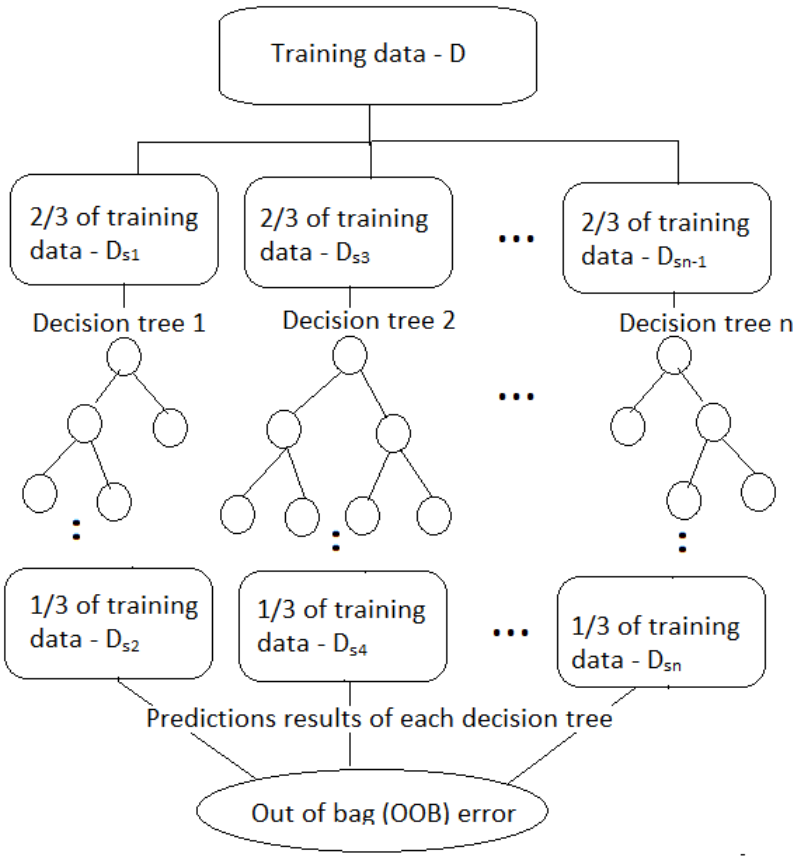

Figure 3-3 - Random forest

3.5.4 eXtreme gradient boosting

Chen & Guestrin (2016) introduced XGBoost, an algorithm for solving regression and classification problems. This algorithm is well known for producing winning solutions on machine learning competitions hosted by Kaggle (Chen & Guestrin, 2016). Kaggle is an online platform that hosts machine learning challenges, companies such as Mercedes-Benz, Facebook, Google, etc. have published data on the platform to allow data scientists to enter competitions to solve

various data science challenges. This algorithm is also used extensively in academic research (Li et al., 2019; Munkhdalai et al., 2019; Wei et al., 2019).

Like RF, XGBoost utilizes multiple decision trees. However, unlike RF where trees are built independently, XGBoost builds trees sequentially. Each new tree learns from the errors of the previous one, progressively improving predictions (Chen & Guestrin, 2016).

Suppose we have n predictor variables ($x_1, x_2,...,x_n$). The process of developing an XGBoost is as follows:

1. Initialization: Start with an initial prediction for each data point ('observed value') and set a threshold for minimum ‘Cover’ (a measure of node importance). A cover value is pre-initialized and computed by summing the weights assigned to all records corresponding to a specific tree node during the training process.
2. Root node: Calculate the residuals (difference between predicted and actual values). The root of the tree is formed from these residuals. Its similarity score is calculated (Equation (9)). At this stage, the tree is only made up of one leaf, i.e., the root.
3. Splitting: Consider splitting the root using a single predictor variable:
    - Sort the values of the predictor variable.
    - Test splits by taking the average of consecutive values, placing the highest value on the right branch, others on the left.
    - Splits are only retained if the 'Cover' of the resulting child nodes exceeds the initial threshold.
    - Calculate the similarity scores for the resulting left and right nodes (Equation (9)).

4. Calculating Gain:
    - For each potential split, calculate the Gain metric (Equation (10)). The split with the highest Gain is chosen.
5. Subsequent splits (up to desired depth):
    - If the desired tree depth is reached, stop.
    - Otherwise, repeat steps 3-4 using the subset of data in the new leaf node to identify further splits.
6. Pruning for complexity:
    - Calculate complexity (Equation (11)) for each branch, starting from the bottom of the tree.
    - Prune any branch that results in a negative complexity score.
        - This process is repeated until the root of the tree. In a case where all the branches have been removed and only the root remains, the tree is discarded since each tree must have a depth greater than zero. The depth of the root node is zero. However, in a case where at least one of the branches in the tree has a non-negative complexity value, irrespective of what the complexity value of the node is, the tree is kept.
7. Calculating leaf output values:
    - Use Equation 12 to determine the output values for each leaf node.
8. Making predictions:
    - For each data point, use Equations (13) and (14) to calculate the log-odds of an event.

- Calculate probabilities using Equation (15).

9. Updating residuals and building new trees:
    - The residuals from the predictions of this tree replace the predicted values from the previous iteration (step 1).
    - Build a new tree focusing on these residuals.
    - Repeat this process until the desired number of trees is reached or performance plateaus.

The goal is to minimize these output values. Additional trees are added into the model until the specified number of trees in the model is reached or the output values do not show an improvement.

$$Similarity_{score(T)} = \frac{\sum_{i=1}(observed_i - p_i)^2}{\lambda + \sum_{i=1} p_i(1-p_i)} = \frac{\sum_{i=1}(Residual_i)^2}{\lambda + Cover} \quad (9)$$

where p_i is the probability of the i^{th} observation, $observed_i$ is the observed probability of the i^{th} observation and λ is the regularization parameter.

$$Gain_B = (Similarity_{score}\ left\ leaf) + (Similarity_{score}\ right\ leaf) - (Similarity_{score}\ root) \quad (10)$$

$Gain_B - \gamma$	(11)

where $Gain_B$ is the $Gain$ of a branch B and $\gamma \geq 0$ is the minimum loss reduction term. A larger γ results in a conservative or less complex model.

$O_{value} = \frac{\sum_{i=1}(observed_i - p_i)}{\lambda + \sum_{i=1} p_i(1-p_i)} = \frac{\sum_{i=1}(Residual_i)^2}{\lambda + \sum_{i=1} p_i(1-p_i)}$	(12)

where p_i is the probability of the i^{th} observation, $observed_i$ is the observed probability of an event, for example, default or non-default, and $\lambda \geq 0$ is a regularisation term.

$$log(odds) = log\left(\frac{p_i}{1-p_i}\right)$$	(13)

where p_i is the probability of the i^{th} observation.

$$log(odds)\_prediction = log\left(\frac{p_i}{1-p_i}\right) + \eta * O_{value}$$	(14)

where p_i is the probability of the i^{th} observation, $\eta \in [0,1]$ is the learning rate and O_{value} is the output value for the leaf where the residual of the observation found.

$$Probability_{prediction} = \frac{exp\left(log\left(\frac{p_i}{1-p_i}\right)\right)}{1 + exp\left(log\left(\frac{p_i}{1-p_i}\right)\right)}$$	(15)

where p_i is the probability of the ith observation.

This illustration covers the details of developing an XGBoost for a simple case. In a case where the data is complex, XGBoost employs weighted quantile sketch to partition the data (Chen & Guestrin, 2016). For a more in-depth treatment of this algorithm, the reader is referred to Chen & Guestrin (2016).

3.5.5 Light gradient boosting

Introduced by Ke et al. (2017), LightGBM distinguishes itself from gradient boosting algorithms like XGBoost in its leaf-wise tree growth strategy.

The key differences between LightGBM and XGBoost:

- Leaf-wise growth: LightGBM constructs decision trees' leaf-wise, prioritizing nodes that yield the highest reduction in loss. This contrasts with XGBoost's depth-wise growth.
- Efficiency: The leaf-wise growth contributes to the efficiency of the LightGBM model, especially when dealing with high-dimensional data.

The similarities between LightGBM and XGBoost:

- Like other boosting algorithms, LightGBM starts with initial predictions and 'Cover' values.
- Trees are built iteratively by splitting nodes to further reduce residuals.
- Complexity is managed through pruning.
- Predicted probabilities are calculated from leaf output values (similar to Equations (13) to (15)).

For a more in-depth treatment of this algorithm, the reader is referred to Ke et al. (2017).

3.5.6 Categorical boosting

CatBoost, proposed by Prokhorenkova et al. (2018), shares similarities with XGBoost and LightGBM as a gradient boosting algorithm but introduces a few distinctive features.

Notably, CatBoost efficiently handles categorical predictor variables without the need for prior preprocessing, reducing the need for complex categorical feature encoding, such as one-hot encoding or label encoding. This simplification streamlines the model-building process and can potentially improve performance, especially when dealing with datasets containing many categorical variables.

For a more in-depth understanding, readers are referred to Prokhorenkova et al. (2018).

3.5.7 Shapley values

The SHAP framework (Lundberg & Lee, 2017) provides a robust method for interpreting the outputs of complex machine learning models, including tree-based algorithms such as RF and Gradient Boosting Machines (GBM). By calculating Shapley values for each predictor variable, the framework quantifies the contribution of each variable to a given prediction, ensuring a transparent understanding of model behaviour. This transparency is particularly valuable in credit scoring, where regulatory and operational requirements necessitate explainable decision-making.

Shapley values offer a powerful tool for enhancing transparency and accountability in credit risk modelling. By precisely quantifying each variable's contribution to individual creditworthiness predictions (Lundberg & Lee, 2017), they empower lenders to understand the drivers behind credit decisions. This granular insight facilitates improved decision-making, allowing practitioners to identify key factors influencing creditworthiness (Bussmann et al., 2020), and ensure alignment with regulatory expectations in the finance industry (Fritz-Morgenthal et al., 2022).

However, the computational complexity of calculating Shapley values, particularly for high-dimensional datasets, presents a challenge (Jethani et al., 2022). While approximation techniques like SHAP offer some relief, they introduce additional workflow steps, potentially increasing risks and costs (Fritz-Morgenthal et al., 2022). Moreover, ensuring the interpretability and actionability of SHAP explanations across diverse stakeholders, from regulators to consumers, requires careful consideration and adds complexity to implementation (Fritz-Morgenthal et al., 2022).

Lundberg and Lee (2017) highlight three key properties that make Shapley values uniquely suited for interpreting machine learning models:

i. Local accuracy: The predicted outcome for a specific instance can be fully explained by the contributions of its input features.
ii. Missingness: An absent variable in a model has zero contribution towards the prediction. This property is best described in Winter (2002), no payoffs is assigned to players who contributes nothing to the marginal contribution with respect to every alliance.
iii. Consistency or symmetry: Predictor variables that have the same contribution in a model contribute equally towards model prediction.

The prediction of each record is given by the following:

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i x_i \quad (16)$$

where ϕ_0 is the naive prediction i.e., prediction without any predictor variables, $\phi_i, i = 1,2, \dots M$ are the parameters of variables $x_i, i = 1,2, \dots M$ and x_i, $i = 1,2, \dots M$ are the model variables inputs. The reader should note that $\phi_i x_i, i = 1,2, \dots M$ are the Shapley values.

In Section 3.3.3.4, this research highlighted the methods provided by Siddiqi (2016) to determine predictor variables that lead to a customer scoring less than the required score to qualify for credit. Firstly, a neutral score is calculated using the intercept of an LR model, the number of variables in the model, the Offset and Factor parameters. The Offset and Factor parameters are independent of the model and have been described in Section 3.3.3.1. These two parameters are typically driven

by business rational to align the distribution of credit scores to a specific range, in this research, no guidelines exist therefore any score range is sufficient.

It is important to highlight that the scorecard intercept referred to in Siddiqi (2016) is derived from an LR model. The intercept of an LR model, β_0 in Equation (8) is the expected mean value of the prediction when all independent variables are equal to zero (Osborne, 2017).

Note the similarity to the SHAP framework (Equation (16)). The term ϕ_0 represents the average predicted value across the dataset and acts as the baseline when all features are absent. This makes ϕ_0 analogous to the LR intercept to β_0.

To calculate the neutral score from Siddiqi (2016) (Equation (6)), instead of using the intercept from a traditional LR model, this research proposes using ϕ_0 from Equation (16). We justify this because both terms represent a baseline prediction when all features are absent.

Additionally, this research proposes to use $\phi_i, i = 1,2, \dots, M$ from Equation (16) in the place of the LR parameters $\beta_i, i = 1,2, \dots, M$ from Equation (8). This proposal is informed by the premise that in an LR model,

$$f(x) = ln(odds) = ln\left(\frac{p}{1-p}\right) = \alpha + \sum_{i=1}^{M} \beta_i x_i \quad (17)$$

where α is the naive prediction, $\beta_i, i = 1,2, \dots M$ are the parameters of the variables $x_i, i = 1,2, \dots M$ and x_i, $i = 1,2, \dots M$ are the inputs values of the model variables. Similarly, $\phi_i, i = 0,1,2, \dots, M$ in Equation (16) can be expressed as either probability or log-odds (Lundberg & Lee,

2017). This research proposes to convert the Shapley values to log-odds, this assists in aligning to the log-odds output produced by the LR model.

This approach sets up an insightful comparison:

- Conventional method: First, credit scores will be calculated using method in Siddiqi (2016) based on an LR model. This will identify potential adverse factors using the standard LR coefficients.
- SHAP-based method: Next, a log-odds output will be calculated using Shapley values from the tree-based models (RF, XGBoost, LightGBM, and CatBoost). This log-odds value, interpreted through the SHAP values, will reveal the feature contributions specific to tree-based predictions.

Analysing these results side-by-side can potentially uncover differences in how features are weighted for decision-making between the LR approach and the tree-based ensemble models.

3.6 Hyperparameter tuning

Hyperparameter tuning is a crucial aspect of machine learning, as it directly impacts model accuracy (Xia et al., 2017). The various hyperparameter tuning techniques including:

- Bayesian Optimization: Uses probability models to guide the search.
- Grid Search: Systematically explores predefined hyperparameter combinations.
- Random Search: Randomly samples hyperparameter combinations.
- Manual Search: guided by human expertise or domain knowledge (Yang et al., 2022).

The choice of a specific method depends on factors such as computational resources and problem complexity, with each approach striking a balance between comprehensiveness and efficiency in the search of the optimal model configurations (Yang & Shami, 2020).

This study employs a Grid Search for hyperparameter tuning due to its proven effectiveness across machine learning algorithms (Pan et al., 2022). Grid Search provides a systematic and exhaustive exploration of predefined hyperparameter combinations, offering a thorough examination of the model's performance across a range of configurations.

3.7 Summary of the methodologies

In this methodology chapter, this study illustrated various approaches employed in the development and evaluation of credit scoring models with a focus on using two distinct datasets: Yeh (2009) and Home Credit Group (2018).

Data analysis methods are introduced, encompassing feature engineering to create additional predictor variables, sampling to split the data into training and test sets, and statistical significance tests to identify redundant predictor variables. The methods for calculating credit scores, mirroring practical applications, are also presented. Variable importance is assessed using permutation feature importance or the Gain metric, depending on the dimensionality of the data. Additionally, critical preprocessing steps, including variable binning, one-hot encoding, handling missing values, and addressing multicollinearity and outliers, are addressed.

In addition, the model-X knockoffs framework is introduced to aid in predictor variable selection. For the evaluation of model performance, misclassification statistics are employed, including the

confusion matrix and the AUC metric. Furthermore, the DeLong et al. (1988) test is incorporated to compare the AUC metrics of credit scoring models.

This section outlined the modelling strategies employed in the study. For the Yeh (2009) and Home Credit Group (2018) datasets (addressing research questions 1 and 2), five credit scoring models are developed, including tree-based algorithms such as XGBoost, RF, LightGBM, and CatBoost, along with an LR model. Furthermore, the Shapley values framework is introduced, contributing to the establishment of interpretability foundations for the tree-based algorithms. In the Home Credit Group (2018) data (addressing research questions 3 and 4), three tree-based credit scoring models, namely, XGBoost, LightGBM, and CatBoost, are implemented.

Chapter 4 Interpretable credit scorecards using Shapley values

This chapter serves as an important component within the thesis, focusing specifically on the results and analysis, as well as the proposed framework for calculating credit scores using tree-based algorithms. It encompasses sections extracted from the first manuscript published by the PLoS ONE peer-review journal. While the broader thesis extensively explores the literature and methodological aspects of the research, this chapter covers the findings presented in the first manuscript.

The manuscript addresses the imperative challenge of interpretability in credit scoring models, particularly focusing on advanced machine learning algorithms like XGBoost, RF, LightGBM, and CatBoost. Despite the superior accuracy of these models over the traditionally used LR algorithm, their inherent lack of interpretability has been a persistent concern. In response to this, the study introduces a novel framework that involves discretizing numerical variables and applying one-hot encoding, aligning the predictor variable representation with industry standards and the expectations of credit practitioners. The distinctive feature of this approach is the utilization of Shapley values, offering a transparent representation of predictor variables for each group in the context of credit scoring. The results showcase the framework's effectiveness by providing credit scores that are not only interpretable but also comparable to those derived from the conventional LR algorithm, thus ensuring a seamless integration of advanced machine learning algorithms into practical credit scoring scenarios.

In the broader context of the aims and objectives of this thesis, this paper contributes by demonstrating a methodology to enhance interpretability without sacrificing accuracy. By offering credit practitioners a comprehensible decisionmaking process, the framework aligns with the ove rarching goal of bridging the gap between credit scoring practice and advanced machine learning algorithms. The paper's focus on industry standards and its comparative analysis with the LR algorithm underscore its practical applicability, emphasizing the viability of using tree-based algorithms in real-world credit scoring scenarios. These findings, coupled with the proposed framework, provide insights into how to make credit scoring models transparent and relevant for industry professionals, ensuring informed lending decisions in the dynamic landscape of credit assessment.

4.1 Predictor variables

This research employs two datasets: the Taiwan Credit Card data from (Yeh, 2009), comprising 30,000 loan accounts (6,636 in default, a 22.12% default rate) from April to September 2005, and the Home Credit data from (Home Credit Group, 2018), containing 356,255 customers (24,845 classified as “bad” due to default, a 6.97% default rate), released on Kaggle in June 2018.

Methodologies outlined in Section 3.2.2 to generate new predictor variables through feature engineering. Redundant predictor variables were identified using the statistical significance test detailed in Section 3.3.1 and subsequently eliminated redundant predictor variables. To identify the most predictive variables, the study applied the permutation variable importance methodology, as outlined in Section 3.3.2.3. This process involved ranking the variables based on their

importance in predicting the outcome. The final list of predictor variables is provided in Table 4-1 for the Yeh (2009) dataset and Table 0-2 for the Home Credit Group (2018) dataset.

Table 4-1 - Predictor variables

Predictor variable	**Description**
AVG_PAY__SEP	Represents the average repayment status in September 2005, providing an overall measure o payment performance during that month.
CURRENT_OVER_3MAVG_PAY__SEP	Calculates the ratio of the current payment to the average of the previous three months, offering insight into the recent trend of payment behaviour compared to the preceding months.
STD_PAY__SEP	Measures the standard deviation of repayment status in September, indicating the variability in

Predictor variable	Description
	payment patterns during that month.
AVG_PAY__JUN	Represents the average repayment status in June 2005, offering a similar measure of payment performance for a different month.
AVG_BILL_AMT_SEP	Captures the average bill amount in September 2005, providing an indication of the typical credit card statement balances during that month.
AVG_PAY_AMT_SEP	Calculates the average amount of previous payments made in September, offering insight into the regularity and magnitude of payments.

Predictor variable	Description
CURRENT_OVER_3MAVG_BILL_AMT_SEP	Computes the ratio of the current bill amount to the average of the bill amounts over the previous three months, highlighting the current billing status relative to recent trends.

The final list of predictor variables in this table captures various aspects of individuals' bill amount and repayment behaviour.

4.2 Imbalanced data in credit scoring

Addressing class imbalance in the data is critical, especially in the credit scoring domain where the target variable is inherently imbalanced (Brown & Mues, 2012). Tree-based techniques such as XGBoost have demonstrated effectiveness in handling imbalanced datasets (Wei et al., 2019), the choice of these tree-based techniques assist further in ensuring the challenge of imbalanced data is addressed when compared to classic scoring methods such as LR. Additionally, adjusting the probability cutoff threshold in classification models can significantly impact performance on imbalanced data (Zou et al., 2016). By aligning the cutoff with the proportion of the minority class (the 'bad' rate in this instance), practitioners can improve the model's sensitivity to the minority

class, enhancing detection rates (Zou et al., 2016). The approach in this research is based on ensuring the cutoff is as close as possible to the bad rate in the portfolio.

4.3 Proposed framework for calculating credit scores

This framework illustrates a systematic approach for enhancing credit scoring models using Shapley values within the context of the Siddiqi (2016) methodology. It encapsulates the process of deriving credit scores, from the initial stage of predictor variable binning through to credit score calculation. By integrating Shapley values proposed in Lundberg and Lee (2017) into the methodology, this framework offers a comprehensive pathway to derive more transparent and insightful credit scores, ultimately contributing to informed credit decision-making and model refinement. The methodology proposed in this research starts from the binning phase, a foundational step in scorecard development (Siddiqi, 2016), given its paramount role in shaping the final scorecard.

This framework, as detailed in (Hlongwane et al., 2024a), introduces supplementary steps to traditional credit scoring methodologies. Specifically, as depicted in Figure 4-1, the proposed approach applies one-hot encoding to binned predictor variables before model fitting, a step designed to improve the handling of categorical data. Furthermore, Shapley values replace the LR parameters traditionally used, providing a more transparent and interpretable measure of each predictor variable's contribution to the model's output. This integration of Shapley values addresses the challenge of aligning advanced machine learning algorithms with interpretability standards required in credit scoring.

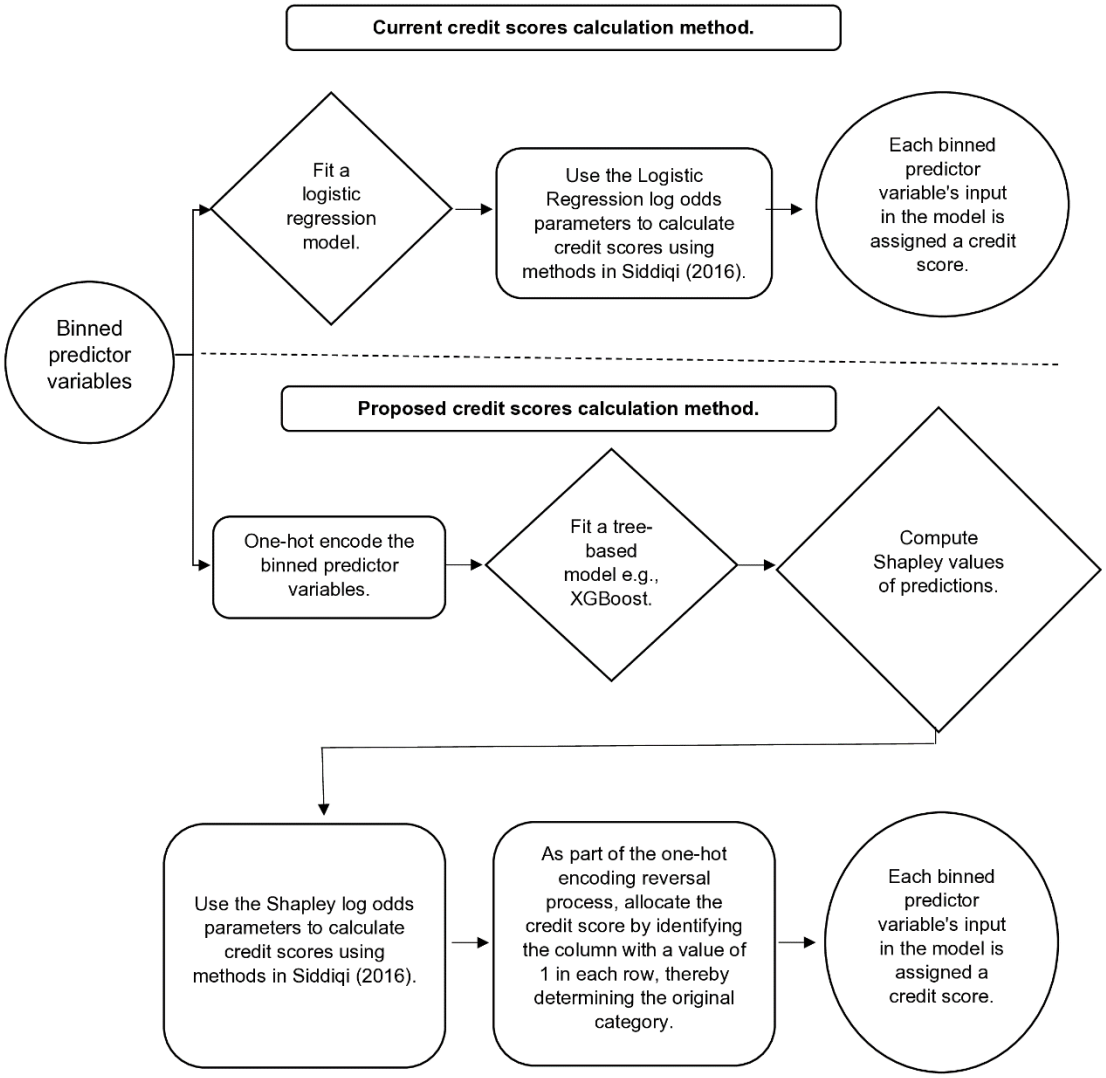

Figure 4-1 - Credit scores calculation process flow - current vs. proposed

4.4 Results and analysis

This section presents the outcomes of the credit scoring models and delves into their performance. This includes an in-depth examination of credit scorecards associated with each model, illustrating how individual predictor variables are practically represented. Through a detailed exploration of

these outcomes, this section offers insights into the effectiveness and real-world applicability of the developed models.

4.4.1 Performance of the models

Table 4-2 presents a comparison of the LR, RF, XGBoost, LightGBM, and CatBoost models in terms of AUC for the Yeh (2009) (Taiwan Credit Card) data. The RF model achieved the highest AUC, followed closely by XGBoost and LightGBM. However, the DeLong test (DeLong et al., 1988) indicates that the differences in AUC among these three models are not statistically significant.

Similarly, the AUC values for LR and CatBoost were not significantly different from each other. However, the p-values from the DeLong test show significant differences between the top-performing group (RF, XGBoost, LightGBM) and the lower-performing group (LR, CatBoost).

Notably, our models outperformed the benchmark AUC of 0.697 reported in previous research (Alam et al., 2020; D. Chen et al., 2023) that used the same dataset but without applying feature engineering approach. This suggests that feature engineering, which distinguished our study from previous work in terms of predictor variable utilization, contributed to the improved predictive performance.

Table 4-2. AUC and p-values of the models – Taiwan data

	p-value				
Model	**AUC**	**XGBoost**	**LightGBM**	**LR**	**CatBoost**

RF	0.75929	0.41580	0.15990	0.00021	0.00143
XGBoost	0.75766		0.47520	0.00315	0.00056
LightGBM	0.75690			0.00316	0.00310
LR	0.74891				0.81190
CatBoost	0.74793				

Table 4-3 presents the confusion matrices for the Taiwan Credit Card data models, highlighting the superior predictive power of the RF and XGBoost models. Both achieved the highest overall accuracy (75.717%) and lowest misclassification rate (24.283%), outperforming LightGBM, LR, and CatBoost.

Table 4-3 - Confusion matrices of the models – Taiwan data

RF		**Predicted**	
		Good	Bad
Actual	Good	3,765 (80.363%)	920
	Bad	537	778 (59.163%)
XGBoost		**Predicted**	
		Good	Bad

RF		**Predicted**	
		Good	Bad
Actual	Good	3,767 (80.406%)	918
	Bad	539	776 (59.011%)
LightGBM		**Predicted**	
		Good	Bad
Actual	Good	3,727 (79.552%)	958
	Bad	522	793 (60.304%)
LR		**Predicted**	
		Good	Bad
Actual	Good	3,688 (78.719%)	997
	Bad	533	782 (59.468%)
CatBoost		**Predicted**	
		Good	Bad
Actual	Good	3,657 (78.058%)	1,028
	Bad	515	800 (60.837%)

Table 4-4 presents the AUC values of the different models on the Home Credit Group (2018) (Home Credit) data. The XGBoost model achieved the highest AUC of 0.69766. The DeLong test

(DeLong et al., 1988) confirmed that the differences in AUC between XGBoost and all other models, were statistically significant (p-values < 0.05). The only comparison that did not reach statistical significance was between LightGBM and LR, suggesting their AUC values are not significantly different according to the DeLong test (DeLong et al., 1988).

Table 4-4 - AUC and p-values of the models – Home Credit data

	p-value				
Model	**AUC**	**XGBoost**	**LightGBM**	**LR**	**CatBoost**
RF	0.69280	0.00000	0.00044	0.00012	0.00002
XGBoost	0.69766		0.02081	0.00466	0.00000
LightGBM	0.69654			0.87847	0.00000
LR	0.69644				0.00000
CatBoost	0.68450				

Table 4-5 presents the confusion matrices of the Home Credit data models. The XGBoost model achieved the highest overall accuracy (70.335%) and the lowest misclassification rate (29.665%) compared to the other models.

Table 4-5 - Confusion matrices of the models – Home Credit data

RF		**Predicted**	
		Good	Bad
Actual	Good	39422 (69.775%)	17077
	Bad	2057	2947 (58.893%)
XGBoost		**Predicted**	
		Good	Bad
Actual	Good	40299 (71.327%)	16200
	Bad	2045	2959 (59.133%)
LightGBM		**Predicted**	
		Good	Bad
Actual	Good	40060 (70.904%)	16439
	Bad	2019	2985 (59.652%)
LR		**Predicted**	
		Good	Bad
Actual	Good	39545 (69.992%)	16954
	Bad	2006	2998 (59.912%)
CatBoost		**Predicted**	

RF		Predicted	
		Good	Bad
		Good	Bad
Actual	Good	39178 (69.343%)	17321
	Bad	2024	2980 (59.552%)

Overall, these results corroborate previous findings (Jabeur et al., 2021; Lessmann et al., 2015) demonstrating the superior performance of tree-based models compared to classic techniques like LR in credit risk assessment.

4.4.2 Interpretable credit models – Taiwan

Previous research, such as (Bracke et al., 2019; Bueff et al., 2022; Bussmann et al., 2020), focused on providing marginal probability or log-odds contributions of each variable in a model, shedding light on their statistical significance.

Figure 4-2 illustrates the type of interpretability offered by previous studies, showcasing the log-odds contributions of each predictor variable for a specific customer in the dataset. While statistically informative, this type of output, which focuses on log-odds or probabilities, may not be readily interpretable or actionable for credit practitioners who primarily rely on credit scores for decision-making (Siddiqi, 2016). This section aims to bridge this gap by drawing parallels between the parameters used in LR-based models and those derived from the SHAP framework, proposing to replace LR parameters with Shapley values for identifying top reasons for model

predictions. We compare the established method for determining top reasons for credit scorecard predictions (Siddiqi, 2016) with our proposed approach using the SHAP framework (Lundberg & Lee, 2017).

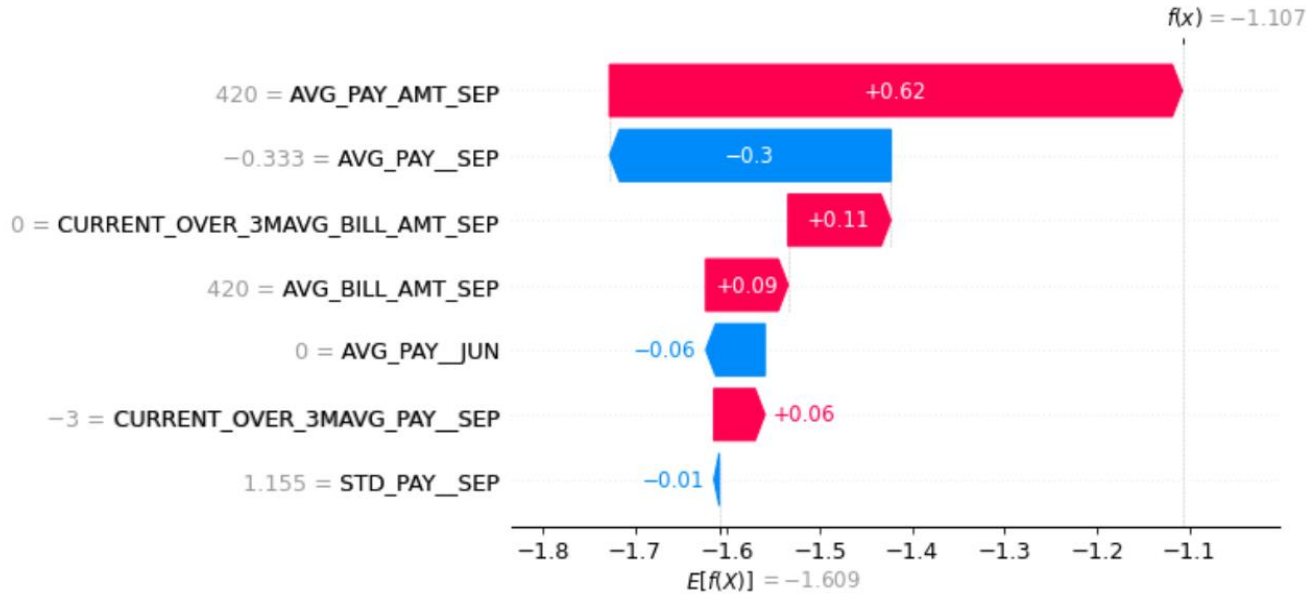

Figure 4-2 – Log-odds of the predictor variables

The following representations visually distinguish credit scores below the neutral score by shading them in grey. We provide side-by-side comparisons of credit scores based on both LR parameters and Shapley values. All five models were developed using seven predictor variables with consistent binning.

Table 4-6 to Table 4-12 illustrate the credit scores of the predictor variables on the Taiwan data. In most cases, the five models agree regarding the predictor variable bins that lie below the neutral credit score, thereby presenting potential explanations for customers receiving lower credit scores. Except for the predictor variable “Average Bill Amount (July, August, September)” in Table 4-7,

where the RF model suggests that only the bin (-inf, 13.50) could potentially be cited as a reason for an applicant receiving a lower credit score.

Table 4-6 - Average Payment Indicator - July, August & September

		LR	**XGBoost**	**RF**	**LightGBM**	**CatBoost**
Predictor Variable	Bin	Credit Score				
Average Payment Indicator (July, August, September)	[0.17, inf)	0.0000	77.7914	0.0000	0.0000	0.0000
	(-inf, 0.17)	227.9544	189.8177	203.2322	1145.1521	208.4717
Neutral Credit Score		181.5809	167.0278	161.8880	912.1900	166.0616

Table 4-7 - Average Bill Amount - July, August & September

		LR	**XGBoost**	**RF**	**LightGBM**	**CatBoost**

Predictor Variable	Bin	Credit Score				
Average Bill Amount (July, August, September)	(-inf, 13.50)	75.4879	84.5542	69.1873	0.0000	66.7293
	[13.50,49794.83)	119.9583	120.3305	119.6997	106.1236	119.5988
	[49794.83, inf)	154.1544	126.4671	126.5869	129.8170	165.6828
Neutral Credit Score		128.18	120.3961	119.2480	107.9997	131.0323

Table 4-8 - Average payment indicator – April, May, and June

		LR	**XGBoost**	**RF**	**LightGBM**	**CatBoost**
Predictor Variable	**Bin**	**Credit Score**				
	[1.50, inf)	0.0000	33.9752	0.0000	40.5293	0.0000

Average Payment Indicator (Apr, May, Jun)	[0.50, 1.50)	55.7196	79.5345	0.0000	82.6911	17.5708
	(-inf, 0.50)	172.5623	232.1287	263.2195	1,590.1881	237.1356
Neutral Credit Score		151.0169	205.6815	222.4819	1,354.2532	202.0626

Table 4-9 - Ratio Sep Payment divided by a 3-months Avg Payment (Jul, Aug, Sep)

		LR	**XGBoost**	**RF**	**LightGBM**	**CatBoost**
Predictor Variable	**Bin**	**Credit Score**				
Ratio September Payment Indicator over a 3 months	[1.23, inf)	72.9460	18.1803	82.6155	0.0000	0.0000
	[0.20, 1.23)	106.2807	88.8360	109.3608	74.6328	45.9130
	(-inf, 0.20)	171.8272	299.5807	149.1770	264.4279	622.8268

Average Payment Indicator (July,August, September)						
Neutral Credit Score		140.2733	201.8103	129.3548	175.2677	370.0489

Table 4-10 - Standard Deviation of Payment Indicator - July, August, September

		LR	**XGBoost**	**RF**	**LightGBM**	**CatBoost**
Predictor Variable	**Bin**	**Credit Score**				
Standard Deviation of Payment Indicator (July, August, September)	[0.79, inf)	69.4371	0.0000	93.2310	38.0470	27.2959
	(-inf, 0.79)	155.7639	204.9736	132.7333	219.7735	204.3044

Neutral Credit Score		137.4367	161.4577	124.3469	181.1929	166.7255

Table 4-11 - Average Payment Amount - July, August, September

		LR	**XGBoost**	**RF**	**LightGBM**	**CatBoost**
Predictor Variable	**Bin**	**Credit Score**				
Average Payment Amount (July, August, September)	(-inf, 31.17)	62.0784	0.0000	85.1439	0.0000	0.0000
	[31.17,2001.83)	91.5675	7.1893	103.2557	43.4340	0.0000
	[2001.83,4312.17)	131.3691	144.8357	127.4429	141.8376	130.2874
	[4312.17, inf)	181.3795	265.6866	156.7996	246.9170	174.6079
Neutral Credit Score		127.3518	121.2622	124.6605	128.1473	86.2906

Table 4-12 - Ratio Sep Bill Amt over a 3-months Avg Bill Amt - Jul, Aug, Sep

		LR	**XGBoost**	**RF**	**LightGBM**	**CatBoost**
Predictor Variable	**Bin**	**Credit Score**				
Ratio September Bill Amount over a 3 months Average Bill Amount (July, August, September)	[0.83, 1.06)	30.9308	91.7615	77.6660	94.3556	0.0000
	(-inf, 0.83)	149.9491	145.6139	127.7371	142.1031	146.3869
	[1.06, inf)	203.8937	191.2323	139.0207	180.9784	193.4898
Neutral Credit Score		109.3790	133.7848	107.1574	131.1313	90.5642

The consistency and similarity in predictor variable input values across models have yielded compelling results. The models largely agree on which input values fall below or above the neutral credit score, demonstrating consistency in identifying potential reasons for a credit application decline. A significant finding of this research is the successful substitution of LR parameters with Shapley values to derive credit scores using the methodology outlined in Siddiqi (2016), showcasing the practical applicability of Shapley values in credit scoring.

4.4.3 Interpretable credit models – Home Credit

Across the Home Credit data, Table 4-13 to Table 4-23 illustrate the credit scores of the eleven predictor variables. Notably, in all instances, the five models consistently agree on which predictor variable bins fall below the neutral credit score, thus providing potential explanations for why customers might receive lower scores.

Table 4-13 - Average – Approved annuity amount

		LR	**XGBoost**	**RF**	**LightGBM**	**CatBoost**
Predictor Variable	**Bin**	**Credit Score**				
Average	(-inf, 4160.83)	23.6388	104.1387	121.8600	107.3289	0.0000

– Approved annuity amount	[4160.83, 8934.75)	91.5894	116.3997	121.8615	117.3830	26.1096
	[8934.75, inf)	162.1021	128.6020	121.8626	129.5077	142.7540
Neutral Credit Score		135.5539	123.9826	121.8622	125.0247	103.1746

Table 4-14 - Max number days between application and payment

		LR	**XGBoost**	**RF**	**LightGBM**	**CatBoost**
Predictor Variable	**Bin**	**Credit Score**				
Maximum number of days (relative to the application	[-19.50, inf)	121.8622	106.8595	121.8618	115.0803	114.3648
	(-inf, -19.50)	124.4529	125.0546	121.8623	128.6087	123.2026

date) on which a payment was made for previous instalments						
Neutral Credit Score		124.0384	122.1440	121.8623	126.4446	121.7888

Table 4-15 - Maximum – Approved credit amount

		LR	**XGBoost**	**RF**	**LightGBM**	**CatBoost**
Predictor Variable	**Bin**	**Credit Score**				
Maximum – Approved	(-inf, 50954.04)	0.0000	0.0000	121.8580	100.4851	88.3454
	[50954.04, 898398.00)	123.4501	126.0522	121.8622	122.8748	123.8610

credit amount	[898398.00, inf)	138.4285	165.5761	121.8623	132.4269	142.7153
Neutral Credit Score		112.5787	117.2354	121.8618	121.5474	122.1043

Table 4-16 - Average of external scores (1, 2 & 3)

		LR	**XGBoost**	**RF**	**LightGBM**	**CatBoost**
Predictor Variable	**Bin**	**Credit Score**				
Average of external scores (1, 2 & 3)	(-inf, 0.42)	121.8622	28.2460	83.5474	35.5794	0.0000
	[0.42, inf)	190.7930	175.5838	138.3170	180.3757	288.2603
Neutral Credit Score		177.4769	147.1210	127.7366	152.4039	232.5741

Table 4-17 - Normalized score from external data source 3

		LR	XGBoost	RF	LightGBM	CatBoost
Predictor Variable	**Bin**	**Credit Score**				
Normalized score from external data source 3	(-inf, 0.31)	121.8622	0.0000	92.8836	94.9131	0.0000
	[0.31, inf)	162.9762	215.6522	129.7101	132.4732	160.9843
Neutral Credit Score		156.9643	184.1183	124.3251	126.9810	137.4443

Table 4-18 - Normalized score from external data source 2

		LR	XGBoost	RF	LightGBM	CatBoost
Predictor Variable	**Bin**	**Credit Score**				

Normalized score from external data source 2	(-inf, 0.35)	121.8622	0.0000	46.7201	0.0000	0.0000
	[0.35, inf)	172.4943	355.6024	153.1185	339.8045	215.9150
Neutral Credit Score		161.6762	279.6242	130.3853	267.2016	169.7825

Table 4-19 - Normalized score from external data source - 1

		LR	**XGBoost**	**RF**	**LightGBM**	**CatBoost**
Predictor Variable	**Bin**	**Credit Score**				
Normalized score from external data source 1	(-inf, 0.27)	121.8622	49.3349	121.8610	47.4972	117.9383
	[0.27, inf)	129.7068	216.1394	121.8623	197.6447	138.7986

Neutral Credit Score		129.1451	204.1939	121.8622	186.8921	137.3047

Table 4-20 - Days since last credit application

		LR	**XGBoost**	**RF**	**LightGBM**	**CatBoost**
Predictor Variable	**Bin**	**Credit Score**				
Maximum - How many days before current application did client apply for Credit Bureau credit	[-76.50, inf)	121.8622	80.3987	107.6043	87.7062	11.3244
	(-inf, -76.50)	125.7284	131.9663	123.8594	137.1557	136.2507

Neutral Credit Score		125.3388	126.7700	122.2214	132.1728	123.6623

Table 4-21 - Average - Days past due of Instalments

		LR	**XGBoost**	**RF**	**LightGBM**	**CatBoost**
Predictor Variable	**Bin**	**Credit Score**				
Average - Days past due of Instalments	[0.08, inf)	121.8622	102.7606	115.5192	95.9194	92.8871
	(-inf, 0.08)	138.9704	252.5404	127.8514	238.8954	150.6132
Neutral Credit Score		130.6599	179.7826	121.8609	169.4427	122.5719

Table 4-22 - Number of days in current employment before application

		LR	XGBoost	RF	LightGBM	CatBoost
Predictor Variable	**Bin**	**Credit Score**				
How many days before the application the person started current employment	[6123.75, inf)	121.8622	87.3832	121.8599	87.3148	112.2891
	(-inf, 6123.75)	151.7564	168.3528	121.8692	157.0443	172.7415
Neutral Credit Score		130.4769	110.7163	121.8626	107.4089	129.7098

Table 4-23 - Ratio of Annuity amount / Credit Amount

		LR	XGBoost	RF	LightGBM	CatBoost

Predictor Variable	Bin	Credit Score				
Ratio of Annuity amount / Credit Amount	[0.05, inf)	121.8622	65.6892	109.3365	71.7789	108.1582
	(-inf, 0.05)	141.7548	160.1158	145.3041	177.0741	143.2001
Neutral Credit Score		129.9096	103.8887	123.8869	114.3752	122.3341

The consistent agreement across all models regarding which predictor variable input values fall below or above the neutral credit score demonstrates the robustness of our approach and reinforces the potential of Shapley values as a viable alternative to LR parameters for deriving interpretable credit scores, as demonstrated in the Taiwan dataset. This finding further supports the applicability of the methodology outlined in Siddiqi (2016) for a broader range of credit scoring models.

4.5 Summary

The literature is explicit about the limitations of adopting advanced machine learning algorithms such as those proposed in this research. Hosaka (2019) and Wei et al. (2019) highlighted that the primary barrier to adoption is the lack of transparency in predictions, which is a critical requirement for credit regulators.

The findings in this research indicate that transparency need not be a reason for the slow adoption of advanced machine learning algorithms in practice. Based on the findings of this study, the

conclusion is drawn that credit scores derived from Shapley values align closely with credit scores produced by LR models, meeting regulatory expectations for interpretability.

Furthermore, this study demonstrates that the Shapley-based approach to producing credit scores aligns with industry practices, as outlined by Siddiqi (2016). In contrast, previous works such as Bracke et al. (2019), Bueff et al. (2022) and Bussmann et al. (2020) focus on presenting marginal probabilities or log-odds contributions that do not align with practical scorecard presentations used in the industry. This distinction underscores the relevance and applicability of the Shapley-based framework proposed in this research.

This research also establishes that Shapley values can efficiently extract reasons for unfavourable credit reports from predictor variables, with these explanations aligning with industry standards. The SHAP framework serves as a powerful tool for demystifying the black-box nature of machine learning systems, as evidenced in this study.

Additionally, the research corroborates earlier studies that showcase the superior performance of XGBoost and RF models compared to LR concerning accuracy. The consistent findings across various investigations lend further credibility to the efficacy of these advanced models in credit scoring applications. The alignment of results with existing research strengthens the evidence base, reinforcing confidence in leveraging XGBoost and RF as robust tools for credit risk assessment.

Chapter 5 Improving credit scoring accuracy with alternative data

This chapter, which forms an important component of the thesis, presents the results and analysis section of the second manuscript published by the PLoS ONE peer-reviewed journal.

The second manuscript focusses on understanding the role of alternative data in credit scoring. In the face of the challenges and opportunities presented by big data, the study navigates the inclusion of previously overlooked information in predictive models. The data utilized for this exploration, sourced from the Home Credit Group (2018), provides customer attributes, credit bureau data, and various attributes of the individuals' financial profiles. Through an analysis of the consequences of excluding alternative data, the research underscores the significance of predictor variables such as social contexts and regional information in enhancing the accuracy of credit scoring models.

Additionally, the manuscript introduces a distinctive dimension by incorporating the model-X knockoffs framework into the domain of credit scoring. Recognizing the high-dimensional nature of the Kaggle home credit dataset, this framework becomes instrumental in addressing the imperative need for predictor variable selection. The study showcases the efficacy of tree-based algorithms, specifically XGBoost, LightGBM, and CatBoost, in crafting predictive models on this data. Importantly, the proposed predictor variable approach leads to credit scoring models that surpass existing models on the Home Credit Group (2018) data.

In alignment with the broader objectives of the thesis, this manuscript significantly advances the understanding of alternative data's impact on credit scoring. It introduces a novel approach to

predictor variable selection, emphasizing practical implications by shedding light on the consequences of excluding alternative data and providing insights into effective modelling techniques. In doing so, the research contributes to the overarching goal of enhancing accuracy in credit scoring models, highlighting the potential of advanced machine learning algorithms in addressing the challenges posed by big data in credit scoring.

5.1 Results and analysis

This section presents the outcomes of the credit scoring models and delves into their performance. This includes an in-depth examination of credit scorecards associated with each model, illustrating how the predictor variables influence the performance of the models. Through a detailed exploration of these outcomes, this section offers insights into the effectiveness of the developed models.

5.1.1 Alternative predictor variables

To address research questions 3 and 4, this study employs data sourced from Home Credit Group (2018), comprising 356,255 customers that have been granted home loans. Among them, 24,845 customers are classified as defaulters due to non-payment on their home loan accounts, resulting in a default rate of approximately 6.97%. According to García et al. (2012), this percentage represents a high-class imbalance, necessitating careful model evaluation and preprocessing strategies. Approaches such as adjusting probability thresholds, leveraging tree-based models inherently suited for imbalanced data, and exploring oversampling or under-sampling techniques are considered to address this imbalance (Wei et al., 2019; Zou et al., 2016).

In this study, the approach focuses on setting the classification cutoff as close as possible to the bad rate in the portfolio while utilizing tree-based techniques, which are well-suited to handling imbalanced data. This dual strategy ensures that model performance effectively accounts for the class distribution in the dataset while maintaining accuracy.

The data pre-processing methodology described in Section 3.3.4.2, involved grouping predictor variables with a correlation of 0.7 and higher, following the recommendation by Romano et al. (2020), resulting in 551 groups. Following a methodology similar to Al Daoud (2019), a LightGBM model ranked and identified predictive variables within each group using the Gain metric evaluation. This process led to the removal of 230 redundant variables. Further refinement of predictor variable selection was achieved using the deep knockoffs proposed by Romano et al. (2020), reducing the number of predictor variables from 321 to a final set of 215. Out of these 215 predictor variables, 22 are alternative data variables. Table 5-1 provides an overview of these alternative predictor variables, capturing unique social and geographical.

Table 5-1 - Alternative predictor variables

Predictor variable	**Description**	**p-value**
CNT_CHILDREN	Number of children the client has	0.00004
CNT_FAM_MEMBERS	The client's family size	0.00000

Predictor variable	Description	p-value
OBS_30_CNT_SOCIAL_CIRCLE	The number of instances in the client's social surroundings with observed 30 days past due (DPD) default	0.00000
DEF_30_CNT_SOCIAL_CIRCLE	The number of instances in the client's social surroundings with observed 30 DPD default	0.00000
OBS_60_CNT_SOCIAL_CIRCLE	The number of instances in the of client's social surroundings with observed 60 DPD default	0.00000
DEF_60_CNT_SOCIAL_CIRCLE	The number of instances in the client's social surroundings with observed 60 DPD	0.00000

Predictor variable	Description	p-value
DAYS_LAST_PHONE_CHANGE	The number of days before the application when the client changed their phone.	0.00000
REGION_RATING_CLIENT_W_CITY	Rating of the region where client lives taking city into account (1,2,3)	0.00000
REGION_RATING_CLIENT	Rating of the region where client lives (1,2,3)	0.00000
REGION_POPULATION_RELATIVE	Normalized population of region where client lives (higher number means the client lives in more populated region)	0.00000
REG_CITY_NOT_LIVE_CITY	A flag of whether the client's permanent matches the contact address	0.00000

Predictor variable	Description	p-value
	(1=different, 0=same, at city level)	
REG_CITY_NOT_WORK_CITY	A flag of whether client's permanent address matches the work address (1=different, 0=same, at city level)	0.00000
LIVE_CITY_NOT_WORK_CITY	A flag of whether client's contact address matches the work address (1=different, 0=same, at city level)	0.00000
FLAG_DOCUMENT_3	Did client provide document 3	0.00000
AMT_ANNUITY	Loan annuity	0.00022
AMT_CREDIT	Credit amount of the loan	0.00000

Predictor variable	Description	p-value
AMT_GOODS_PRICE	For consumer loans it is the price of the goods for which the loan is given	0.00000
EXT_SOURCE_1	Normalized score from external data source	0.00000
EXT_SOURCE_2	Normalized score from external data source	0.00000
EXT_SOURCE_3	Normalized score from external data source	0.00019
APPS_ANNUITY_CREDIT_RATIO	Ratio of AMT_ANNUITY / AMT_CREDIT	0.00000

Furthermore, the Wald test was utilized to test the significance of the 22 alternative variables in predicting default. All predictor variables listed in Table 5-1 underwent the Wald test, and all p-values were found to be less than 5%, indicating their significance in the study.

The remaining 193 predictor variables cover diverse aspects associated with loan applications, financial histories, and client attributes. Within this data, there are indicators of creditworthiness,

including external sources and credit-related ratios, shedding light on the financial well-being of applicants. Demographic information, such as age, employment history, and identity document updates, offering personal profiles of clients is also present. Furthermore, predictor variables detailing payment behaviours, debt amounts, and credit line histories provide an indication of the financial habits of the clients, providing a holistic view of their financial outlook.

This research employs three algorithms: XGBoost, LightGBM, and CatBoost to construct credit models. These three models are the most common credit scoring models that previous research, such as Al Daoud (2019) and Tounsi et al. (2020b), used on the Kaggle home credit data, offering an opportunity to compare the performance of the models in this study to other previous studies. Each algorithm is applied to develop a model encompassing the complete set of predictor variables, following the elimination of non-predictive variables. Additionally, an evaluation will be conducted by excluding the 22 alternative predictor variables and subsequently reconstructing the models using the remaining predictor variables. This assessment aims to determine whether the exclusion of alternative predictor variables leads to credit models that are less predictive.

5.1.2 Performance of the models

Models were constructed using XGBoost, LightGBM, and CatBoost with and without the alternative features, allowing us to assess their impact on performance.

Initially, XGBoost, LightGBM, and CatBoost algorithms were employed to construct predictive models encompassing the complete set of 215 predictor variables. Subsequently, the modelling process was replicated after excluding the alternative predictor variables. The ensuing comparison aimed to assess the impact of their exclusion on model performance.

Table 5-2 presents the model performance results. Models constructed without alternative predictor variables showed reduced performance across all algorithms, as measured by the AUC. The DeLong et al. (1988) confirmed the statistical significance of these AUC differences (p-values < 0.05 for all comparisons).

The LightGBM model, utilizing the full set of predictor variables, achieved the highest AUC score of 0.79360. This performance aligns with findings from prior studies (Al Daoud, 2019; Coşkun and Turanli, 2023; Qiu et al., 2019) which also highlight the effectiveness of tree-based techniques. Furthermore, the models developed in this study outperformed previously reported benchmarks for LR on the same dataset, including AUC scores of 0.68031 (Chen et al., 2019) and 0.7574 (Yu et al., 2021).

Table 5-2 - AUC of the models

Model	**AUC**	**p-value**
XGBoost	0.78916	< 2.2e-16
XGBoost (alternative data excluded)	0.74499	
LightGBM	**0.79360**	

Model	AUC	p-value
LightGBM (alternative data excluded)	0.75073	< 2.2e-16
CatBoost	0.78897	< 2.2e-16
CatBoost (alternative data excluded)	0.74444	

Table 5-3 presents the results of models trained exclusively on traditional data and models trained exclusively on alternative data. Models trained on alternative data consistently achieved higher AUC scores across all tested algorithms (XGBoost, LightGBM, and CatBoost). The DeLong test confirmed the statistical significance of these AUC improvements (p-values < 0.05). These findings provide strong evidence for the predictive power of alternative data in credit scoring, highlighting its potential to enhance model accuracy and decision-making.

Table 5-3 - AUC (Traditional data vs Alternative data)

Model	AUC	p-value
XGBoost (alternative data only)	**0.76378**	4.925e-06
XGBoost (traditional data only)	0.74499	
LightGBM (alternative data only)	0.76177	0.008309
LightGBM (traditional data only)	0.75073	
CatBoost (alternative data only)	0.75932	0.0003902
CatBoost (traditional data only)	0.74444	

The confusion matrix in Table 5-4 shows that the LightGBM model (using all predictor variables) achieves the highest true negative rate (specificity) at 74.171%, while the CatBoost model has the highest true positive rate (sensitivity) at 83.459%.

Table 5-4 - Confusion matrix of the three models

XGBoost		**Predicted**	
		Good	Bad
Actual	Good	31038 (73.167%)	11383
	Bad	732	2974 (80.248%)
XGBoost (alternative data excluded)		**Predicted**	
		Good	Bad
Actual	Good	29841 (70.345%)	12580
	Bad	787	2919 (78.764%)

LightGBM		**Predicted**	
		Good	Bad
Actual	Good	31464 (74.171%)	10957
	Bad	638	3068 (82.785%)
LightGBM (alternative data excluded)		**Predicted**	
		Good	Bad
Actual	Good	30311 (71.453%)	12110
	Bad	711	2995 (80.815%)

CatBoost		**Predicted**	
		Good	Bad
Actual	Good	31368 (73.945%)	11053
	Bad	613	3093 (83.459%)
		Predicted	

CatBoost		**Predicted**	
		Good	Bad
CatBoost (alternative data excluded)		Good	Bad
Actual	Good	30457 (71.797%)	11964
	Bad	651	3055 (82.434%)

Table 5-5 shows that the LightGBM model using the full set of predictor variables achieved the lowest overall misclassification rate (25.137%).

Table 5-5 - Overall misclassification rate

Model	**Overall Misclassification Rate**
XGBoost	26,264%
XGBoost (alternative data excluded)	28,979%

Model	Overall Misclassification Rate
LightGBM	**25,137%**
LightGBM (alternative data excluded)	27,795%
CatBoost	25,291%
CatBoost (alternative data excluded)	27,348%

This study demonstrates the critical importance of alternative data, including financial, social, and geographic factors, for accurate credit scoring. Excluding these variables led to a significant decline in model performance.

5.1.3 Performance of alternative variables

Feature importance analysis highlights the significant impact of alternative data, with variables like APPS_ANNUITY_CREDIT_RATIO, AMT_ANNUITY, and the mean of EXT_SOURCE ranking among the top predictor variables. This emphasizes the value of non-credit bureau metrics, such as loan structure and application details, for improving model performance. The

EXT_SOURCE variables specifically demonstrate how diverse data can capture nuanced borrower behaviour. These findings align with and expand upon those published by Hlongwane et al. (2024b), who demonstrated that incorporating alternative data sources, including social network behaviours and telecommunication variables, significantly enhances the accuracy of credit scoring models.

This underscores the broader benefits of alternative data in predictive modelling. Expanding such data improves understanding of borrowers and loans, leading to better decision-making. The inclusion of alternative variables as top predictor variables reinforces the need to move beyond traditional credit bureau data alone. Integrating diverse data allows for more comprehensive models and ultimately enhances risk management strategies.

However, alongside these benefits, the use of alternative data raises important questions about fairness and potential biases. While alternative data can improve inclusivity by capturing insights about previously underrepresented groups, it may also introduce new biases that reflect systemic inequalities present in its sources. For instance, variables derived from social network behaviours or geographic data may inadvertently disadvantage certain demographic groups.

Future research should evaluate how fairness is maintained or improved when incorporating alternative data. Techniques such as counterfactual fairness testing can help identify whether the inclusion of these variables disproportionately impacts specific groups (Xiong et al., 2020). Bias mitigation strategies, including re-weighting or excluding sensitive features, can be employed to address these challenges. Additionally, leveraging tools like Shapley values for fairness evaluation

provides a systematic way to assess the contributions of individual predictor variables and identify any unintended biases in the model's decisions (Hickey et al., 2021).

By addressing these fairness considerations, credit scoring models can better balance predictive performance with ethical responsibility, ensuring that alternative data contribute to equitable decision-making processes. This multi-faceted approach is critical for advancing the responsible use of alternative data in credit scoring while maintaining stakeholder trust.

5.2 Summary

This research investigates the impact of alternative data, specifically social and geographic variables, on the accuracy of credit risk prediction models. It builds upon the concept of "social scoring" (Wei et al., 2016) by demonstrating the predictive power of these variables in assessing creditworthiness. Excluding these alternative predictor variables reduced model performance across all methods tested, highlighting their importance. These findings align with prior studies (Agarwal et al. 2018; De Cnudde et al., 2019; Pedro et al., 2015) and demonstrate the potential of alternative data for improving credit scoring models.

Using the model-X knockoffs framework for predictor variable selection, the LightGBM model achieved the highest reported AUC (0.79360) on the Kaggle Home Credit dataset. This emphasizes the framework's effectiveness for handling diverse data. Moreover, models trained on alternative data consistently achieved higher AUC scores across all tested algorithms (XGBoost, LightGBM, and CatBoost), with improvements confirmed as statistically significant by the DeLong test (p-values < 0.05). These findings provide strong evidence for the predictive power of alternative data in credit scoring, highlighting its potential to enhance model accuracy and decision-making.

Chapter 6 Discussion

Serving as the research's discussion section, this chapter integrates the analysis and interpretation of study results with existing literature to examine the implications and significance of the findings, emphasizing the study's strengths and contributions while also addressing its limitations, thus offering a balanced perspective on its scope and applicability.

6.1 Restating the aims and research questions

This study aims to improve credit scoring model accuracy and interpretability while addressing the trade-off between performance and transparency. These objectives align with industry demands for models that deliver high predictive power while meeting regulatory and stakeholder expectations.

To achieve this goal, the study outlines several key objectives. Firstly, it develops a comprehensive framework to enhance the interpretability of advanced machine learning algorithms, emphasizing tree-based models such as random forest (RF), eXtreme gradient boosting (XGBoost), light gradient-boosting machine (LightGBM), and categorical boosting (CatBoost). Secondly, it utilizes the SHAP framework to derive interpretable credit scores from tree-based models, following Siddiqi (2016) methodology. Thirdly, it applies the model-X knockoffs framework to high-dimensional credit scoring data, demonstrating its effectiveness in predictor variable selection. Finally, it investigates the impact of including alternative data on credit scorecard accuracy, employing various statistical methods to assess its additive value.

The research questions focus on model accuracy and interpretability. They examine the accuracy difference between tree-based and LR models, effective ways to leverage Shapley values for transparent decision-making, the contribution of model-X knockoffs to predictor variable selection in high-dimensional data, and the improvement in overall model accuracy brought by alternative data.

By addressing these research questions, the study aims to provide insights and recommendations for enhancing credit scoring model accuracy and interpretability, contributing to the advancement of credit risk assessment methodologies in the banking industry.

6.2 Key findings

The study highlights the superior performance of tree-based credit scoring models—RF, XGBoost, and LightGBM—compared to LR, especially in terms of AUC and reduced misclassification rates. These findings confirm the efficacy of tree-based models in capturing complex, non-linear relationships, thereby enhancing credit risk assessments and informing more robust lending decisions.

A novel framework is introduced, integrating Shapley values into traditional scorecard methodologies to address interpretability challenges in tree-based models. The study demonstrates a close alignment between credit scores derived from Shapley values and the LR model, mitigating transparency concerns associated with advanced machine learning algorithms. Using the DeLong et al. (1988) test, this study confirms the consistent superiority of tree-based models over LR, enriching discussions on applying machine learning to credit scoring.

Furthermore, the study yields several significant findings. Firstly, the application of the model-X knockoffs framework, particularly the deep knockoffs method, proves instrumental in enhancing the effectiveness of high-dimensional credit scoring data by identifying representative predictor variables and addressing dimensionality and correlated variables. Secondly, the inclusion of alternative data, encompassing social and geographical variables, demonstrates an impact on the overall accuracy of credit scoring models, consistently outperforming models without alternative data.

The LightGBM model, constructed using the full set of predictor variables identified through the model-X knockoffs framework and incorporating alternative data, achieves an AUC score of 0.79356, surpassing previous models in the literature. This highlights the importance of both the model-X knockoffs framework and the incorporation of alternative data in advancing credit scoring.

6.3 A comparison with literature

This study's findings align with and build upon existing research on credit scoring model performance and interpretability. Previous studies, such as Lessmann et al. (2015), Wei et al. (2019), and Xia et al. (2018), emphasize the superior predictive accuracy of tree-based algorithms like RF, XGBoost, and LightGBM compared to traditional methods like LR. Consistent with these studies, this research confirms the efficacy of tree-based models in capturing complex, non-linear relationships, leading to improved performance metrics such as AUC and reduced misclassification rates.

Additionally, the study contributes to the growing body of work on model interpretability in credit scoring. Siddiqi (2016) emphasized the importance of interpretability for regulatory compliance and practical implementation. This research integrates Shapley values into credit scoring frameworks to enhance transparency and address 'black box' concerns (Hertza, 2018). The findings are consistent with Lundberg and Lee (2017), who demonstrated Shapley values' effectiveness in explaining model predictions.

The inclusion of alternative data in the Home Credit dataset aligns with broader credit scoring trends highlighted by Ala'raj and Abbod (2016). These studies emphasized the potential of non-traditional data sources to improve predictive accuracy and financial inclusion. This research supports these findings by demonstrating that alternative data significantly enhances model performance while maintaining interpretability through the proposed framework.

While this study supports existing literature, it also highlights gaps requiring further exploration. For instance, while most studies focus on performance metrics, this research emphasizes the balance between accuracy and interpretability, addressing a critical trade-off in credit scoring practices. By combining advanced algorithms with interpretable frameworks, this study provides a more holistic approach to credit risk assessment.

6.4 Addressing research questions and hypothesis

For the first research question, "How does the accuracy of tree-based algorithms compare to that of logistic regression in credit scoring?", the study employed the DeLong et al.'s (1988) test to compare models using the AUC metric. The results highlighted significant differences in predictive accuracy.

Tree-based models, such as RF, XGBoost, and LightGBM, consistently demonstrated higher AUC scores compared to the LR model. These findings underscore the superior discriminatory power of tree-based models and suggest potential advantages of employing such methodologies in credit scoring, implying enhanced predictive capabilities and overall model efficacy.

The results of the DeLong et al. (1988) test reject the null hypothesis, affirming significant differences in accuracy between tree-based models and LR. The findings confirm the alternative hypothesis, demonstrating that tree-based algorithms yield higher accuracy than logistic regression models in credit scoring.

The consistent outperformance of tree-based models over LR aligns with previous studies (Jabeur et al., 2021; Lessmann et al., 2015; Son et al., 2019) and reinforces the growing recognition of advanced machine learning techniques in credit scoring. By leveraging the flexibility and nonlinearity of tree-based algorithms, these models effectively capture complex patterns and interactions within the data, translating into improved accuracy and discrimination between creditworthy and non-creditworthy individuals.

To address the second research question, "In what ways can Shapley values be effectively leveraged to represent predictor variables in the context of credit scoring, ensuring a transparent and understandable decision-making process?", this study demonstrates their practical application.

The findings of this study, rooted in the interpretability analysis of credit scorecards generated using Shapley values, provide compelling evidence supporting the alternative hypothesis. By

providing clear insights into the influence of predictor variables, Shapley values enhance both understanding and trust in credit scoring decisions.

The second hypothesis aimed to explore whether Shapley values contribute to enhancing transparency and comprehensibility in the decision-making process of credit scoring. The null hypothesis posited that Shapley values do not significantly contribute to creating a transparent and understandable decision-making process, while the alternative hypothesis suggested the opposite. The findings, rooted in the interpretability analysis of credit scorecards generated using Shapley values, provide evidence in support of the alternative hypothesis. Shapley values offer clear insights into the impact of each predictor variable on credit scoring decisions, providing a transparent view akin to the interpretability achieved with LR. This transparency, demonstrated through visualizations derived from Shapley values, enhances understanding and trust in the credit scoring process. Consequently, the null hypothesis is rejected, and the alternative is accepted, affirming that Shapley values indeed contribute to creating a transparent and understandable decision-making process in credit scoring.

In addressing the third research question, "How does the model-X knockoffs framework improve predictor variable selection in high-dimensional credit scoring data?", this study diverges from traditional approaches, aligning with recent advancements in predictor variable selection methods. Unlike some previous studies that might rely on conventional predictor variable selection methods, this research introduces the model-X knockoffs framework, specifically employing the deep knockoffs method proposed by Romano et al. (2020). This departure signifies a methodological

advancement, showcasing an effective way of handling high-dimensional credit scoring data. This distinctive aspect sets the study apart from traditional literature and positions it at the forefront of methodological change in credit scoring research.

The model-X knockoffs framework identifies relevant predictor variables while eliminating redundancies, addressing challenges posed by high-dimensional credit scoring datasets. This methodology contributes to the effectiveness of predictor variable selection in high-dimensional data, as validated through the test by DeLong et al. (1988).

For the third set of hypotheses concerning the application of the model-X knockoffs framework in predictor variable selection, the findings support the conclusion that the model-X knockoffs framework contributes positively to the effectiveness of predictor variable selection in high-dimensional credit scoring data. Consequently, the null hypothesis suggesting no significant contribution is rejected, and the alternative hypothesis affirming the impact of the model-X knockoffs framework on enhancing predictor variable selection in the context of high-dimensional credit scoring data is accepted.

To explore the fourth research question, "What impact does incorporating alternative data have on the accuracy of credit scoring models?", the study highlights the impact of variables such as social and geographical data. Through an analysis encompassing diverse alternative predictor variables, including family and social circle data, the findings underscore the nuanced contribution of these variables in enhancing credit scoring accuracy.

Excluding alternative data reduces model performance across algorithms, underscoring its importance in credit risk assessment. This aligns with studies by Agarwal et al. (2018), De Cnudde et al. (2019), and Pedro et al. (2015) and extends them by exploring a wide array of alternative predictor variables. Consequently, this research not only reinforces the value of alternative data in credit scoring but also highlights the necessity of incorporating such variables for a more accurate evaluation of creditworthiness.

In examining the fourth set of hypotheses related to the inclusion of alternative data in credit scoring models, the analysis, supported by the test introduced by DeLong et al. (1988) for statistical significance, reveals that alternative data indeed leads to a statistically significant improvement in the overall accuracy of credit scoring models. Therefore, the null hypothesis, positing no significant improvement, is rejected, and the alternative hypothesis indicating an enhancement in model performance with the incorporation of alternative data is accepted.

6.5 Implications of the study

The implications of this study are multifaceted, providing valuable insights for both academia and industry.

Firstly, the study highlights the superior performance of tree-based credit scoring models, such as RF, XGBoost, and LightGBM, compared to LR, aligning with findings from previous research. These findings directly benefit industry practitioners seeking more accurate and reliable credit scoring models. By demonstrating the effectiveness of these algorithms, this research emphasizes their potential to improve predictive accuracy in credit scoring, offering a robust alternative to traditional models.

The integration of Shapley values into the framework addresses the longstanding transparency challenges of advanced credit scoring models. This enhancement ensures that even complex algorithms provide interpretable outputs, meeting both regulatory requirements and operational demands. Notably, the close alignment between credit scores derived from Shapley values and those generated by LR offers a bridge between advanced algorithms and industry expectations for transparent credit scoring practices.

The study highlights the transformative impact of alternative data on credit scoring models. By incorporating diverse data sources, such as social and geographical variables, and employing the model-X knockoffs framework—particularly the deep knockoffs method—for predictor variable selection, the research demonstrates significant advancements over previous credit scoring approaches. The findings reveal that excluding alternative data leads to a decline in model performance across all three scoring models evaluated, underscoring the importance of a holistic approach that integrates diverse information sources.

This integration of alternative data provides a pathway for assessing the creditworthiness of individuals with limited or no credit history, promoting financial inclusion. The study advocates for the adoption of sophisticated variable selection methodologies and highlights the economic potential of such advancements, including reduced credit losses, optimized capital allocation, and enhanced customer satisfaction.

Finally, this research advocates for the practical application of advanced machine learning algorithms in credit scoring. By fostering a deeper understanding of the factors influencing creditworthiness, these algorithms aid decision-making processes within the credit industry.

Moreover, the proposed methodology bridges the gap between accuracy and interpretability, supporting the development of predictive, inclusive, and transparent credit scoring models.

6.6 Limitations

Reliance on two datasets may limit the generalizability of the findings to other contexts or data environments. As Guo et al. (2019) note, reliance on specific datasets may introduce biases and limit result transferability to broader scenarios. Future research should test the framework on diverse datasets to enable more robust and generalizable conclusions.

Additionally, while the use of specific machine learning algorithms has provided valuable insights, it does not encompass the full range of available techniques. Xia et al. (2017) highlight the importance of considering alternative algorithms and hyperparameter tuning strategies to fully evaluate the potential of advanced credit scoring methodologies. Expanding the analysis to include other methods could further validate the effectiveness of the proposed framework.

Another limitation lies in the scope of alternative data sources utilized in this study. Óskarsdóttir et al. (2019) suggest that socio-economic and cultural factors play a critical role in shaping credit behaviours, and their inclusion could enhance the accuracy and fairness of credit scoring models. Future research should investigate the integration of diverse data sources to better capture the nuances of borrower characteristics and behaviours across different populations.

Addressing these limitations will not only strengthen the generalizability and robustness of the findings but also contribute to a more nuanced and comprehensive understanding of credit scoring practices in varying contexts.

Chapter 7 Conclusion

This chapter concludes the research by delving into the findings and overarching contributions, summarizing the insights from the analyses in Chapters 4 and 5 to reflect on the broader significance of the study.

7.1 Overall findings

This study analyses credit scoring models, demonstrating the superior performance of tree-based approaches. These approaches include RF, XGBoost, and LightGBM, which outperform traditional logistic regression methods. For instance, models using tree-based algorithms consistently achieved higher AUC scores and lower misclassification rates, confirming findings from prior studies (e.g., Lessmann et al., 2015).

A key contribution of this study is integrating Shapley values into credit scoring. This integration enhances interpretability without sacrificing accuracy. By aligning credit scores derived from Shapley values with those from logistic regression, the study addresses transparency concerns often associated with machine learning models, providing credit professionals with actionable and comprehensible insights into decision-making processes.

Additionally, the incorporation of alternative data, such as social and geographical variables, significantly improved model accuracy. Models utilizing these variables outperformed those relying solely on traditional data, demonstrating the importance of diverse data streams in refining credit risk assessment and promoting financial inclusion.

The study demonstrates the importance of alternative data in improving credit scoring accuracy. This includes social and geographic variables. By utilizing the model-X knockoffs framework, the study shows the significant impact of alternative data sources on model performance. Models incorporating these variables consistently outperform those without, emphasizing the need to consider diverse data streams in credit risk assessment. Overall, the findings contribute to advancing credit scoring methodologies. They provide empirical evidence supporting the effectiveness of tree-based models, Shapley values, and alternative data in enhancing the accuracy and transparency of credit risk assessment processes.

7.2 Overall contributions

This study makes significant contributions to the field of credit scoring, both academically and practically, by addressing key challenges related to accuracy, interpretability, and the integration of alternative data.

Academic Contributions

1. **Advancing Machine Learning in Credit Scoring:** The study demonstrates the superior performance of tree-based models, including RF, XGBoost, and LightGBM, compared to traditional LR. This contributes to the growing body of research advocating for advanced machine learning algorithms in credit risk assessment.
2. **Novel Framework for Interpretability:** A novel framework integrating Shapley values into traditional scorecard methodologies is introduced. This addresses the pressing interpretability challenges associated with advanced machine learning models, particularly

tree-based algorithms, bridging the gap between their complexity and the transparency requirements crucial for the credit industry.

3. **Alignment with Logistic Regression:** By aligning credit scores derived from Shapley values with those from LR models, the study ensures transparency while preserving the superior predictive power of advanced models. The use of statistical tests, such as the DeLong et al. (1988) test, reinforces the reliability of these findings and provides robust insights into model performance.

4. **Model-X Knockoffs Framework:** The research introduces and validates the effectiveness of the model-X knockoffs framework, specifically the deep knockoffs method, in handling high-dimensional credit scoring data. By reducing predictor variables and addressing dimensionality and correlation issues, this framework enhances the precision and reliability of credit scoring models.

Practical Contributions

1. **Improved Credit Scoring Practices:** The study offers practical insights into the application of Shapley values for identifying predictor variable categories that contribute to low credit scores. This provides actionable tools for credit professionals to better understand and address drivers of creditworthiness, improving the implementation of advanced models in real-world scenarios.

2. **Incorporation of Alternative Data:** The research highlights the value of incorporating alternative data, including social and geographical variables, to enhance model accuracy. By systematically comparing models with and without these predictor variables, the study

demonstrates consistent performance improvements when alternative data is considered. This has significant implications for promoting financial inclusion by enabling the assessment of individuals with limited or no credit history.

3. **Refinement of Credit Scoring Models:** The findings emphasize the importance of advanced predictor variable selection methodologies, such as the model-X knockoffs framework, in refining credit scoring models. The study advocates for integrating diverse data sources to develop models that are both accurate and reliable, ultimately aiding more effective creditworthiness assessments.

Overall Impact

By addressing two pivotal research questions, this study provides a holistic framework that enhances the accuracy, interpretability, and inclusivity of credit scoring models. These contributions not only advance academic understanding but also equip industry practitioners with tools to implement sophisticated and transparent credit scoring methodologies, ensuring better decision-making and customer outcomes.

7.3 Future research

This study opens several avenues for future research to further advance the field of credit scoring. Key areas for exploration include the practical implementation of interpretability methods, the integration of alternative data, and the financial impact of advanced credit scoring models.

1. Practical Implementation of Interpretability Methods

Future research should focus on the real-world application of the proposed interpretability framework, particularly on its integration into industry practices. Evaluating the framework's impact on decision-making processes will provide insights into its feasibility and operational effectiveness. Additionally, investigating the adaptability of this framework to enhance the interpretability of other advanced models, such as neural networks (NN), could expand its utility across diverse machine learning techniques.

Another valuable direction involves testing the framework across various industries and datasets to assess its generalizability and robustness. For instance, exploring its application in industries like insurance or e-commerce may yield a comprehensive understanding of its scalability in different contexts.

2. Feature Engineering and Alternative Predictor Variables

Advancing feature engineering for alternative predictor variables is a crucial area of focus. Future studies should aim to optimize feature reduction methods to align with practical scorecard development processes. This includes refining methodologies to identify the most relevant predictor variables from large datasets while maintaining interpretability.

Additionally, integrating new data sources, such as telecommunication customer records and social media platforms like Twitter and Facebook, could enrich credit scoring models. These data sources offer unique insights into customer behaviour and potentially enhance the comprehensiveness and accuracy of predictions.

3. Comparative Analysis of Data Sources

Future research should explore the integration of both credit bureau and alternative data sources. Although research combining these datasets is rare, it could provide valuable insights into the comparative value of alternative data in credit scoring models. This analysis would help identify scenarios where alternative data offers the most significant advantages over traditional credit bureau metrics.

4. Financial Impact of Alternative Data

Examining the financial implications of incorporating alternative data is another critical area for future investigation. Research should focus on the potential of alternative data to reduce misclassification rates and improve predictive accuracy in credit scoring models. This analysis could help lenders optimize models for increased profitability, reduced credit losses, and improved customer satisfaction.

5. Bias Detection and Mitigation

Future studies should investigate how Shapley values can be utilized to detect and mitigate biases in credit scoring models, particularly with respect to sensitive variables such as gender, income, and demographic factors. By analysing the contributions of these predictor variables, researchers can identify disproportionate impacts and explore strategies, such as re-weighting or excluding sensitive variables, to enhance fairness. This direction would further align credit scoring models with ethical and regulatory standards, promoting equitable lending practices.

6. Other Advanced Learning Techniques

Future could expand on this work by exploring the application of advanced machine learning techniques in credit scoring. For example, deep neural networks (DNNs) have shown potential in

capturing complex, non-linear relationships in financial data (Kraus et al., 2020), potentially uncovering hidden patterns and improving predictive accuracy. Ensemble learning, particularly cascade methods (Liu et al., 2023), can enhance both accuracy and robustness by combining the strengths of multiple models. Transfer learning holds promise for leveraging knowledge from related domains, especially with limited labelled credit data (Han et al., 2021). Future investigations should compare these techniques with established tree-based methods like Random Forest and XGBoost to understand their relative advantages and limitations, considering factors like predictive accuracy, interpretability, and computational efficiency in real-world credit risk applications.

References

Abowitz, D. A., & Toole, T. M. (2010). Mixed method research: Fundamental issues of design, validity, and reliability in construction research. *Journal of Construction Engineering and Management*, *136*(1), 1–23. https://doi.org/10.1061/(asce)co.1943-7862.0000026

Agarwal, R. R., Lin, C.-C., Chen, K.-T., & Singh, V. K. (2018). Predicting financial trouble using call data—On social capital, phone logs, and financial trouble. *PLoS ONE*, *13*(2), e0191863. https://doi.org/10.1371/journal.pone.0191863

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. In *Organizational Research Methods*, *16*(2), 270-301. https://doi.org/10.1177/1094428112470848

Aidoo, E. N., Appiah, S. K., & Boateng, A. (2021). Brief research report: A Monte Carlo simulation study of small sample bias in Ordered Logit Model under multicollinearity. *Journal of Experimental Education*, *89*(4), 742-750. https://doi.org/10.1080/00220973.2019.1708233

Aitken, R. (2017). "All data is credit data": Constituting the unbanked. *Competition and Change*, *21*(4), 274-300. https://doi.org/10.1177/1024529417712830

Al Daoud, E. (2019). Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset. *International Journal of Computer and Information Engineering*, *13*(1), 6-10.

Alam, T. M., Shaukat, K., Hameed, I. A., Luo, S., Sarwar, M. U., Shabbir, S., Li, J., & Khushi, M. (2020). An investigation of credit card default prediction in the imbalanced datasets. *IEEE Access, 8,* 201173-201198. https://doi.org/10.1109/ACCESS.2020.3033784

Ala'raj, M., & Abbod, M. F. (2016). A new hybrid ensemble credit scoring model based on classifiers consensus system approach. *Expert Systems with Applications, 64,* 36-55. https://doi.org/10.1016/j.eswa.2016.07.017

Alonso-Robisco, A., & Carbó, J. M. (2022). Can machine learning models save capital for banks? Evidence from a Spanish credit portfolio. *International Review of Financial Analysis, 84,* 102372. https://doi.org/10.1016/j.irfa.2022.102372

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance, 23*(4), 589-609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

Arcadu, F., Benmansour, F., Maunz, A., Willis, J., Haskova, Z., & Prunotto, M. (2019). Deep learning algorithm predicts diabetic retinopathy progression in individual patients. *npj Digital Medicine, 2*(1), 92. https://doi.org/10.1038/s41746-019-0172-3

Arráiz, I., Bruhn, M., & Stucchi, R. (2017). Psychometrics as a tool to improve credit information. *The World Bank Economic Review, 30*(Suppl. 1), S67–S76. https://doi.org/10.1093/wber/lhw016

Babaei, G., Giudici, P., & Raffinetti, E. (2023). Explainable fintech lending. *Journal of Economics and Business, 125–126,* 106126. https://doi.org/10.1016/j.jeconbus.2023.106126

Bacaksiz, F. E., Eskici, G. T., & Seren, A. K. H. (2020). “From my Facebook profile”: What do nursing students share on Timeline, Photos, Friends, and About sections? *Nurse Education Today*, *86*, 104326. https://doi.org/10.1016/j.nedt.2019.104326

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, *54*(6), 627-635. https://doi.org/10.1057/palgrave.jors.2601545

Barber, R. F., & Candés, E. J. (2015). Controlling the false discovery rate via knockoffs. *Annals of Statistics*, *43*(5), 2055-2085. https://doi.org/10.1214/15-AOS1337

Barber, R. F., Candès, E. J., & Samworth, R. J. (2020). Robust inference with knockoffs. *Annals of Statistics*, *48*(3), 1409-1431. https://doi.org/10.1214/19-AOS1852

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, *83*, 405-417. https://doi.org/10.1016/j.eswa.2017.04.006

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

Berg, T., Burg, V., Gombović, A., & Puri, M. (2020). On the rise of fintechs: Credit scoring using digital footprints. *The Review of Financial Studies*, *33*(7), 2845-2897. https://doi.org/10.1093/rfs/hhz099

Billon, M., Crespo, J., & Lera-Lopez, F. (2021). Do educational inequalities affect Internet use? An analysis for developed and developing countries. *Telematics and Informatics*, *58*, 101521. https://doi.org/10.1016/j.tele.2020.101521

Bishop, C. M. (2006). Pattern recognition and machine learning. In M. Jordan, J. Kleinberg, & B. Schölkopf (Eds.), *Information Science and Statistics Series*. Springer.

Bracke, P., Datta, A., Jung, C., & Sen, S. (2019). *Machine Learning explainability in finance: An application to default risk analysis*. Bank of England Working Paper No. 816. https://doi.org/10.2139/ssrn.3435104

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Brevoort, K. P., Grimm, P., & Kambara, M. (2016). Credit invisibles and the unscored. *Cityscape: A Journal of Policy Development and Research, 18*(2), 9-24. https://doi.org/10.2139/ssrn.2743007

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, *39*(3), 3446-3453. https://doi.org/10.1016/j.eswa.2011.09.033

Bueff, A. C., Cytryński, M., Calabrese, R., Jones, M., Roberts, J., Moore, J., & Brown, I. (2022). Machine learning interpretability for a stress scenario generation in credit scoring based on counterfactuals. *Expert Systems with Applications, 202,* 117271. https://doi.org/https://doi.org/10.1016/j.eswa.2022.117271

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020). Explainable AI in fintech risk management. *Frontiers in Artificial Intelligence*, *3*, 26. https://doi.org/10.3389/frai.2020.00026

Camana Acosta, M. R., Ahmed, S., Garcia, C. E., & Koo, I. (2020). Extremely randomized trees-based scheme for stealthy cyber-attack detection in smart grid networks. *IEEE Access, 8,* 19921-19933. https://doi.org/10.1109/ACCESS.2020.2968934

Candès, E., Fan, Y., Janson, L., & Lv, J. (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *80*(3), 551-577. https://doi.org/10.1111/rssb.12265

Cerchiello, P., Giudici, P., & Nicola, G. (2017). Twitter data models for bank risk contagion. *Neurocomputing*, *264*, 50-56. https://doi.org/10.1016/j.neucom.2016.10.101

Cerda, P., Varoquaux, G., & Kégl, B. (2018). Similarity encoding for learning with dirty categorical variables. *Machine Learning*, *107*, 1477-1494. https://doi.org/10.1007/s10994-018-5724-2

Charilaou, P., & Battat, R. (2022). Machine learning models and over-fitting considerations. *World Journal of Gastroenterology*, *28*(5), 605-607. https://doi.org/10.3748/wjg.v28.i5.605

Chen, D., Ye, J., & Ye, W. (2023). Interpretable selective learning in credit risk. *Research in International Business and Finance*, *65*, 101940. https://doi.org/10.1016/j.ribaf.2023.101940

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, 785-794. https://doi.org/10.1145/2939672.2939785

Chen, X., Chong, Z., Giudici, P., & Huang, B. (2022). Network centrality effects in peer to peer lending. *Physica A: Statistical Mechanics and Its Applications*, *600*, 127546. https://doi.org/10.1016/j.physa.2022.127546

Chen, X., Liu, Z., Zhong, M., Liu, X., Song, P. (2019). A deep learning approach using DeepGBM for credit assessment. *Proceedings of the 2019 International Conference on Robotics, Intelligent Control and Artificial Intelligence (RICAI '19)*, 774-779. https://doi.org/10.1145/3366194.3366333

Cierniak-Emerych, A., Mazur-Wierzbicka, E., & Rojek-Nowosielska, M. (2021). Corporate social responsibility in Poland. In S. O. Idowu (Eds.), *Current global practices of corporate social responsibility: CSR, sustainability, ethics & governance*. Springer. https://doi.org/10.1007/978-3-030-68386-3_13

Coşkun, S. B., & Turanli, M. (2023). Credit risk analysis using boosting methods. *Journal of Applied Mathematics, Statistics and Informatics, 19*(1), 5–18. https://doi.org/doi:10.2478/jamsi-2023-0001

Costa e Silva, E., Lopes, I. C., Correia, A., & Faria, S. (2020). A logistic regression model for consumer default risk. *Journal of Applied Statistics, 47*(13–15), 2879-2894. https://doi.org/10.1080/02664763.2020.1759030

Couronné, R., Probst, P., & Boulesteix, A. L. (2018). Random forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinformatics, 19*(1), 270. https://doi.org/10.1186/s12859-018-2264-5

Coussement, K., Benoit, D. F., & Van den Poel, D. (2010). Improved marketing decision making in a customer churn prediction context using generalized additive models. *Expert Systems with Applications, 37*(3), 2132-2143. https://doi.org/10.1016/j.eswa.2009.07.029

Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research, 183*(3), 1447-1465. https://doi.org/10.1016/j.ejor.2006.09.100

Dai, C., Lin, B., Xing, X., & Liu, J. S. (2022). False discovery rate control via data splitting. *Journal of the American Statistical Association, 118(544),* 2503-2520. https://doi.org/10.1080/01621459.2022.2060113

Dai, R., & Barber, R. F. (2016). The knockoff filter for FDR control in group-sparse and multitask regression. *Proceedings of the 33rd International Conference on Machine Learning, 48,* 1851-1859.

De Cnudde, S., Moeyersoms, J., Stankova, M., Tobback, E., Javaly, V., & Martens, D. (2019). What does your Facebook profile reveal about your creditworthiness? Using alternative data for microfinance. *Journal of the Operational Research Society, 70*(3), 353-363. https://doi.org/10.1080/01605682.2018.1434402

De Jongh, P., De Jongh, E., Pienaar, M., Gordon-Grant, H., Oberholzer, M., & Santana, L. (2015). The impact of pre-selected variance inflation factor thresholds on the stability and predictive power of logistic regression models in credit scoring. *ORiON, 31*(1), 17-37. https://doi.org/10.5784/31-1-162

De Oliveira, R., Karatzoglou, A., Concejero Cerezo, P., Lopez de Vicuña, A. A., & Oliver, N. (2011). Towards a psychographic user model from mobile phone usage. *Extended Abstracts on Human Factors in Computing Systems (CHI EA '11),* 2191-2196. https://doi.org/10.1145/1979742.1979920

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics, 44*(3), 837-845. https://doi.org/10.2307/2531595

Dieber, J., & Kirrane, S. (2022). A novel model usability evaluation framework (MUsE) for explainable artificial intelligence. *Information Fusion, 81*, 143-153. https://doi.org/10.1016/j.inffus.2021.11.017

Djeundje, V. B., Crook, J., Calabrese, R., & Hamid, M. (2021). Enhancing credit scoring with alternative data. *Expert Systems with Applications, 163*, 113766. https://doi.org/10.1016/j.eswa.2020.113766

Elshawi, R., Al-Mallah, M. H., & Sakr, S. (2019). On the interpretability of machine learning-based model for predicting hypertension. *BMC Medical Informatics and Decision Making*, *19*(1), 146. https://doi.org/10.1186/s12911-019-0874-0

Fehr, E. (2002). The economics of impatience. *Nature*, *415*, 269-272. https://doi.org/10.1038/415269a

Feng, G., Giglio, S., & Xiu, D. (2020). Taming the factor zoo: A test of new factors. *Journal of Finance*, *75*(3), 1327-1370. https://doi.org/10.1111/jofi.12883

Finlay, S. (2010). Credit scoring for profitability objectives. *European Journal of Operational Research*, *202*(2), 528-537. https://doi.org/10.1016/j.ejor.2009.05.025

Fischer, M. L., & Moore, K. (1986). An improved credit scoring function for the St. Paul Bank for Cooperatives. *Journal of Agricultural Cooperation*, *1*, 11-21.

Fritz-Morgenthal, S., Hein, B., & Papenbrock, J. (2022). Financial risk management and explainable, trustworthy, responsible AI. *Frontiers in Artificial Intelligence*, *5*, 779799. https://doi.org/10.3389/frai.2022.779799

Fu, H., Nicolet, D., Mrózek, K., Stone, R. M., Eisfeld, A.-K., Byrd, J. C., & Archer, K. J. (2022). Controlled variable selection in Weibull mixture cure models for high-dimensional data. *Statistics in Medicine*, *41*(22), 4340-4366. https://doi.org/10.1002/sim.9513

García, V., Sánchez, J. S., & Mollineda, R. A. (2012). On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems*, *25*(1), 13-21. https://doi.org/10.1016/j.knosys.2011.06.013

Gathergood, J. (2012). Self-control, financial literacy and consumer over-indebtedness. *Journal of Economic Psychology*, *33*(3), 590-602. https://doi.org/10.1016/j.joep.2011.11.006

Ge, R., Feng, J., Gu, B., & Zhang, P. (2017). Predicting and deterring default with social media information in peer-to-peer lending. *Journal of Management Information Systems*, *34*(2), 401-424. https://doi.org/10.1080/07421222.2017.1334472

Giudici, P. (2001). Bayesian data mining, with application to benchmarking and credit scoring. *Applied Stochastic Models in Business and Industry: Special Issue: Inference and Prediction on Financial Risk Management*, *17*(1), 69-81. https://doi.org/10.1002/asmb.425

Gül, S., Kabak, Ö., & Topcu, I. (2018). A multiple criteria credit rating approach utilizing social media data. *Data and Knowledge Engineering*, *116*, 80-99. https://doi.org/10.1016/j.datak.2018.05.005

Guo, S., He, H., & Huang, X. (2019). A multi-stage self-adaptive classifier ensemble model with application in credit scoring. *IEEE Access*, *7*, 78549-78559. https://doi.org/10.1109/ACCESS.2019.2922676

Gurný, P., & Gurný, M. (2013). Comparison of credit scoring models on probability of default estimation for US banks. *Prague Economic Papers*, *22*(2), 163-181. https://doi.org/10.18267/j.pep.446

Halligan, S., Altman, D. G., & Mallett, S. (2015). Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: A discussion and proposal for an alternative approach. *European Radiology*, *25*, 932-939. https://doi.org/10.1007/s00330-014-3487-0

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). A volume in The Morgan Kaufmann Series in Data Management Systems. Elsevier. https://doi.org/10.1016/C2009-0-61819-5

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Yao, Y., Zhang, A., Zhang, L., Han, W., Huang, M., Jin, Q., Lan, Y., Liu, Y., Liu, Z., Lu, Z., Qiu, X., Song, R., … Zhu, J. (2021). Pre-trained models: Past, present and future. *AI Open*, *2*, 225-250. https://doi.org/10.1016/j.aiopen.2021.08.002

Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning*, *77*(1), 103-123. https://doi.org/10.1007/s10994-009-5119-5

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *160*(3), 523-541. https://doi.org/10.1111/j.1467-985X.1997.00078.x

Hapfelmeier, A., Hornung, R., & Haller, B. (2023). Efficient permutation testing of variable importance measures by the example of random forests. *Computational Statistics and Data Analysis*, *181*, 107689. https://doi.org/10.1016/j.csda.2022.107689

He, Z., Liu, L., Wang, C., Le Guen, Y., Lee, J., Gogarten, S., Lu, F., Montgomery, S., Tang, H., Silverman, E. K., Cho, M. H., Greicius, M., & Ionita-Laza, I. (2021). Identification of putative causal loci in whole-genome sequencing data via knockoff statistics. *Nature Communications*, *12*(1), Art. 3152. https://doi.org/10.1038/s41467-021-22889-4

Hernández Santa Cruz, J. F. (2021). An ensemble approach for multi-stage transfer learning models for COVID-19 detection from chest CT scans. *Intelligence-Based Medicine, 5*. 100027. https://doi.org/10.1016/j.ibmed.2021.100027

Hertza, V. A. (2018). Fighting unfair classifications in credit reporting: Should the united states adopt GDPR-inspired rights in regulating consumer credit? *New York University Law Review*, *93*(6), 1707-1741.

Hickey, J. M., Di Stefano, P. G., & Vasileiou, V. (2021). Fairness by explicability and adversarial SHAP learning. In F. Hutter, K. Kersting, J. Lijffijt, & I. Valera (Eds.), *Machine learning and knowledge discovery in databases*. ECML PKDD 2020. Lecture Notes in Computer Science, vol. 12459. https://doi.org/10.1007/978-3-030-67664-3_11

Hiller, J. S., & Jones, L. S. (2022). Who's keeping score?: Oversight of changing consumer credit infrastructure. *American Business Law Journal*, *59*(1), 61-121. https://doi.org/10.1111/ablj.12199

Hlongwane, R., Ramaboa, K. K. K. M., & Mongwe, W. (2024a). A novel framework for enhancing transparency in credit scoring: Leveraging Shapley values for interpretable credit scorecards. *PLOS ONE*, *19*(8), e0308718. https://doi.org/10.1371/journal.pone.0308718

Hlongwane, R., Ramaboa, K. K. K. M., & Mongwe, W. (2024b). Enhancing credit scoring accuracy with a comprehensive evaluation of alternative data. *PLoS ONE, 19*(5), e0303566. https://doi.org/10.1371/journal.pone.0303566

Home Credit Group. (2018). *Home credit default risk dataset*. Kaggle. https://www.kaggle.com/c/home-credit-default-risk/data [Accessed: 13 July 2022].

Hooker, G., Mentch, L., & Zhou, S. (2021). Unrestricted permutation forces extrapolation: Variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, *31*(6), 82. https://doi.org/10.1007/s11222-021-10057-z

Hosaka, T. (2019). Bankruptcy prediction using imaged financial ratios and convolutional neural networks. *Expert Systems with Applications*, *117*, 287-299. https://doi.org/10.1016/j.eswa.2018.09.039

Hsieh, N.-C., & Hung, L.-P. (2010). A data driven ensemble classifier for credit scoring analysis. *Expert Systems with Applications*, *37*(1), 534-545. https://doi.org/10.1016/j.eswa.2009.05.059

International Monetary Fund. (2023). *2023 Global Debt Monitor*. https://www.imf.org/-/media/Files/Conferences/2023/2023-09-2023-global-debt-monitor.ashx [Accessed: 10 January 2024].

Jabeur, S. Ben, Gharib, C., Mefteh-Wali, S., & Arfi, W. B. (2021). CatBoost model and artificial intelligence techniques for corporate failure prediction. *Technological Forecasting and Social Change, 166, 120658*. https://doi.org/10.1016/j.techfore.2021.120658

Jadhav, S., He, H., & Jenkins, K. (2018). Information gain directed genetic algorithm wrapper feature selection for credit rating. *Applied Soft Computing Journal*, *69*. 541-553. https://doi.org/10.1016/j.asoc.2018.04.033

Jagric, V., Kracun, D., & Jagric, T. (2011). Does non-linearity matter in retail credit risk modeling? *Finance a Uver - Czech Journal of Economics and Finance*, *61*(4), 384-402.

Jenghara, M. M., Ebrahimpour-Komleh, H., Rezaie, V., Nejatian, S., Parvin, H., & Yusof, S. K. S. (2018). Imputing missing value through ensemble concept based on statistical measures. *Knowledge and Information Systems*, *56*(1), 123-139. https://doi.org/10.1007/s10115-017-1118-1

Jethani, N., Sudarshan, M., Covert, I., Lee, S. I., & Ranganath, R. (2022). FastSHAP: Real-time Shapley value estimation. ArXiv, abs/2107.07436. https://api.semanticscholar.org/CorpusID:235899304

Jones, S., Johnstone, D., & Wilson, R. (2017). Predicting corporate bankruptcy: An evaluation of alternative statistical frameworks. *Journal of Business Finance and Accounting*, *44*(1–2), 3-34. https://doi.org/10.1111/jbfa.12218

Judge, G. G., Hill, C., Griffiths, W. E., Lütkepohl, H., & Lee, T.-C. (1988). Introduction to the theory and practice of econometrics. *Journal of the American Statistical Association*, *83*(404), 1229. https://doi.org/10.2307/2290184

Kalnins, A. (2018). Multicollinearity: How common factors cause Type 1 errors in multivariate regression. *Strategic Management Journal*, *39*(8), 2362-2385. https://doi.org/10.1002/smj.2783

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems.* December. https://www.semanticscholar.org/paper/LightGBM%3A-A-Highly-Efficient-Gradient-Boosting-Tree-Ke-Meng/497e4b08279d69513e4d2313a7fd9a55dfb73273

Kelly-Louw, M. (2007). Introduction to the National Credit Act. *Juta's Business Law*, *15*(4), 147–159. https://hdl.handle.net/10520/EJC52610

Khemakhem, S., Ben Said, F., & Boujelbene, Y. (2018). Credit risk assessment for unbalanced datasets based on data mining, artificial neural network and support vector machines. *Journal of Modelling in Management*, *13*(4), 932-951. https://doi.org/10.1108/JM2-01-2017-0002

Klinger, B., Khwaja, A. I., & LaMonte, J. (2013). *Improving credit risk analysis with psychometrics in Peru.* Technical Note No. IDB-TN-587. Inter-American Development Bank.

Kotsiantis, S. B. (2013). Decision trees: A recent overview. In *Artificial Intelligence Review, 39*(4), 261-283. https://doi.org/10.1007/s10462-011-9272-4

Kozodoi, N., Jacob, J., & Lessmann, S. (2022). Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research*, *297*(3), 1083-1094. https://doi.org/10.1016/j.ejor.2021.06.023

Kraus, M., Feuerriegel, S., & Oztekin, A. (2020). Deep learning in business analytics and operations research: Models, applications and managerial implications. *European Journal of Operational Research*, *281*(3), 628-641. https://doi.org/10.1016/j.ejor.2019.09.018

Kritzinger, N., & Van Vuuren, G. W. (2018). An optimised credit scorecard to enhance cut-off score determination. *South African Journal of Economic and Management Sciences*, *21*(1), a1571. https://doi.org/10.4102/sajems.v21i1.1571

Kumar, A., & Jain, M. (Ed.). (2020). *Ensemble learning for AI developers: Learn bagging, stacking, and boosting methods with use cases*. Apress. https://doi.org/10.1007/978-1-4842-5940-5

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, *247*(1), 124-136. https://doi.org/10.1016/j.ejor.2015.05.030

Lextrait, B. (2023). Scaling up SMEs' credit scoring scope with LightGBM. *Applied Economics*, *55*(9), 925-943. https://doi.org/10.1080/00036846.2022.2095340

Li, T., Fong, S., Mohammed, S., Fiaidhi, J., Guan, S., & Chang, V. (2022). Empowering multi-class medical data classification by Group-of-Single-Class-predictors and transfer optimization: Cases of structured dataset by machine learning and radiological images by deep learning. *Future Generation Computer Systems*, *133*(7), 7-22. https://doi.org/10.1016/j.future.2022.02.022

Li, W., Ding, S., Chen, Y., Wang, H., & Yang, S. (2019). Transfer learning-based default prediction model for consumer credit in China. *Journal of Supercomputing*, *75*(2), 862-884. https://doi.org/10.1007/s11227-018-2619-8

Liu, W., Fan, H., & Xia, M. (2022). Credit scoring based on tree-enhanced gradient boosting decision trees. *Expert Systems with Applications*, *189*, 116034. https://doi.org/10.1016/j.eswa.2021.116034

Liu, W., Fan, H., & Xia, M. (2023). Tree-based heterogeneous cascade ensemble model for credit scoring. *International Journal of Forecasting*, *39*(4), 1593-1614. https://doi.org/10.1016/j.ijforecast.2022.07.007

Liu, W., Fan, H., Xia, M., & Xia, M. (2022). A focal-aware cost-sensitive boosted tree for imbalanced credit scoring. *Expert Systems with Applications, 208,* 118158. https://doi.org/10.1016/j.eswa.2022.118158

Lobo, J. M., Jiménez-Valverde, A., & Real, R. (2008). AUC: A misleading measure of the performance of predictive distribution models. In *Global Ecology and Biogeography*, *17*(2), 145-151. https://doi.org/10.1111/j.1466-8238.2007.00358.x

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56-67. https://doi.org/10.1038/s42256-019-0138-9

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17),* 4768-4777.

Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., & Niu, X. (2018). Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. *Electronic Commerce Research and Applications*, *31*, 24-39. https://doi.org/10.1016/j.elerap.2018.08.002

Mahabub, A. (2020). A robust technique of fake news detection using Ensemble Voting Classifier and comparison with other classifiers. *SN Applied Sciences, 2*(4), 525. https://doi.org/10.1007/s42452-020-2326-y

Manek, A. S., Shenoy, P. D., Mohan, M. C., & Venugopal, K. R. (2017). Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier. *World Wide Web: Internet and Web Information Systems, 20*(2), 135-154. https://doi.org/10.1007/s11280-015-0381-x

McCorkell, P. L., & Smith, A. M. (2009). Fair Credit Reporting Act update—2008. *The Business Lawyer*, *64*(2), 579-592. http://www.jstor.org/stable/41552808

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., … Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, *577*, 89–94. https://doi.org/10.1038/s41586-019-1799-6

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093-1113. https://doi.org/10.1016/j.asej.2014.04.011

Meier, S., & Sprenger, C. (2011). *Impatience and credit behavior: Evidence from a field experiment*. Federal Reserve Bank of Boston Working Paper No. 07-3. https://doi.org/10.2139/ssrn.982398

Molnar, C., Freiesleben, T., König, G., Herbinger, J., Reisinger, T., Casalicchio, G., Wright, M. N., & Bischl, B. (2023). Relating the partial dependence plot and permutation feature importance to the data generating process. In L. Longo (Ed.), *Explainable artificial intelligence, xAI 2023. Communications in Computer and Information Science (CCIS), vol. 1901*. Springer. https://doi.org/10.1007/978-3-031-44064-9_24

Moscato, V., Picariello, A., & Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications, 165, 113986*. https://doi.org/10.1016/j.eswa.2020.113986

Munkhdalai, L., Munkhdalai, T., Namsrai, O.-E., Lee, J. Y., & Ryu, K. H. (2019). An empirical comparison of machine-learning methods on bank client credit assessments. *Sustainability (Switzerland)*, *11*(3), 699. https://doi.org/10.3390/su11030699

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(44), 22071-22080. https://doi.org/10.1073/pnas.1900654116

Mushava, J., & Murray, M. (2018). An experimental comparison of classification techniques in debt recoveries scoring: Evidence from South Africa's unsecured lending market. *Expert Systems with Applications*, *111*, 35-50. https://doi.org/10.1016/j.eswa.2018.02.030

Niu, B., Ren, J., & Li, X. (2019). Credit scoring using machine learning by combing social network information: Evidence from peer-to-peer lending. *Information (Switzerland)*, *10*(12), 397. https://doi.org/10.3390/INFO10120397

Osborne, J. W. (2017). *Best practices in logistic regression*. Sage. https://doi.org/10.4135/9781483399041

Óskarsdóttir, M., Bravo, C., Sarraute, C., Vanthienen, J., & Baesens, B. (2019). The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. *Applied Soft Computing Journal*, *74*, 26-39. https://doi.org/10.1016/j.asoc.2018.10.004

Ots, H., Liiv, I., & Tur, D. (2020). Mobile phone usage data for credit scoring. In T. Robal, H. M. Haav, J. Penjam, & R. Matulevičius (Eds.), *Databases and information systems: DB&IS 2020.Communications in computer and information science, vol. 1243.* Springer. https://doi.org/10.1007/978-3-030-57672-1_7

Pan, S., Zheng, Z., Guo, Z., & Luo, H. (2022). An optimized XGBoost method for predicting reservoir porosity using petrophysical logs. *Journal of Petroleum Science and Engineering, 208*(Part C), 109520. https://doi.org/10.1016/j.petrol.2021.109520

Pedro, J. S., Proserpio, D., & Oliver, N. (2015). MobiScore: Towards universal credit scoring from mobile phone data. In F. Ricci, K. Bontcheva, O. Conlan, & S. (Eds.), *User modeling, adaptation and personalization: UMAP 2015*. *Lecture notes in computer science (including subseries lecture notes in Artificial Intelligence and lecture notes in Bioinformatics), vol. 9146.* Springer. https://doi.org/10.1007/978-3-319-20267-9_16

Polanska, A., Price, M. A., Spurio Mancini, A., & McEwen, J. D. (2023). Learned harmonic mean estimation of the marginal likelihood with normalizing flows. *Physical Sciences Forum: Proceedings of the 42nd International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, 9*(1), 10. https://doi.org/10.3390/psf2023009010

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). *CatBoost: Unbiased boosting with categorical features*. https://doi.org/10.48550/arXiv.1706.09516

Qin, G., & Hotilovac, L. (2008). Comparison of non-parametric confidence intervals for the area under the ROC curve of a continuous-scale diagnostic test. *Statistical Methods in Medical Research*, *17*(2), 207-221. https://doi.org/10.1177/0962280207087173

Qiu, Z., Li, Y., Ni, P., & Li, G. (2019). Credit risk scoring analysis based on Machine Learning models. *2019 6th International Conference on Information Science and Control Engineering (ICISCE)*, 220–224. https://doi.org/10.1109/ICISCE48695.2019.00052

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81-106. https://doi.org/10.1007/bf00116251

Quinlan, J. R. (1992). C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. [Book review]. *Machine Learning*, *16,* 235-240. https://doi.org/10.1007/BF00993309

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, 1135-1144. https://doi.org/10.1145/2939672.2939778

Roa, L., Correa-Bahnsen, A., Suarez, G., Cortés-Tejada, F., Luque, M. A., & Bravo, C. (2021). Super-app behavioral patterns in credit risk models: Financial, statistical and regulatory implications. *Expert Systems with Applications*, *169*, 114486. https://doi.org/10.1016/j.eswa.2020.114486

Romano, Y., Sesia, M., & Candès, E. (2020). Deep knockoffs. *Journal of the American Statistical Association, 115*(532), 1861-1872. https://doi.org/10.1080/01621459.2019.1660174

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Muller, K.-R. (Eds.). (2019). Explainable AI: Interpreting, explaining and visualizing deep learning. *Lecture Notes in Computer Science (LNCS), 11700*. Springer. https://doi.org/10.1007/978-3-030-28954-6

Saunders, M., Lewis, P., & Thornhill, A. (2009). Research methods for business students (5th ed). Prentice Hall.

Shapley, L. S. (1953). A value for n-person games. Contributions to the Theory of Games II (1953) 307-317. In H. W. (Ed.), *Classics in Game Theory*. Princeton University Press, 69-79. https://doi.org/10.1515/9781400829156-012

Shen, A., Fu, H., & He, K., & Jiang, H. (2019). False discovery rate control in cancer biomarker selection using knockoffs. *Cancers, 11*(6), 744. https://doi.org/10.3390/cancers11060744

Shi, S., Tse, R., Luo, W., D'Addona, S., & Pau, G. (2022). Machine learning-driven credit risk: A systemic review. *Neural Computing and Applications, 34*, 14327-14339. https://doi.org/10.1007/s00521-022-07472-2

Shi, X., Wong, Y. D., Li, M. Z.-F., Palanisamy, C., & Chai, C. (2019). A feature learning approach based on XGBoost for driving assessment and risk prediction. *Accident Analysis and Prevention, 129*, 170-179. https://doi.org/10.1016/j.aap.2019.05.005

Shi, Y., Peng, Y., Xu, W., & Tang, X. (2002). Data mining via multiple criteria linear programming: Applications in credit card portfolio management. *International Journal of Information Technology & Decision Making*, *01*(01), 131-151. https://doi.org/10.1142/s0219622002000038

Siddiqi, N. (2016). Scorecard development. In *Intelligent credit scoring: Building and implementing better credit risk scorecards* (2nd ed.). John Wiley & Sons. https://doi.org/10.1002/9781119282396.ch2

Son, H., Hyun, C., Phan, D., & Hwang, H. J. (2019). Data analytic approach for bankruptcy prediction. *Expert Systems with Applications, 138,* 112816. https://doi.org/10.1016/j.eswa.2019.07.033

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications, 124*, 93-101. https://doi.org/10.1016/j.eswa.2019.05.028

Stoltzfus, J. C. (2011). Logistic regression: A brief primer. *Academic Emergency Medicine*, *18*(10), 1099-1104. https://doi.org/10.1111/j.1553-2712.2011.01185.x

Suthanthiradevi, P., Srividhyasaradha, K., & Karthika, S. (2021). Modelling a behavioral scoring system for lending loans using Twitter. *ITM Web of Conferences, 37, 01012*. https://doi.org/10.1051/itmconf/20213701012

Tang, L., Cai, F., & Ouyang, Y. (2019). Applying a nonparametric random forest algorithm to assess the credit risk of the energy industry in China. *Technological Forecasting and Social Change, 144*, 563-572. https://doi.org/10.1016/j.techfore.2018.03.007

Thien, T. F., & Yeo, W. S. (2022). A comparative study between PCR, PLSR, and LW-PLS on the predictive performance at different data splitting ratios. *Chemical Engineering Communications*, *209*(11). https://doi.org/10.1080/00986445.2021.1957853

Thiese, M. S., Ronna, B., & Ott, U. (2016). P value interpretations and considerations. *Journal of Thoracic Disease, 8*(9), 1439-1456. https://doi.org/10.21037/jtd.2016.08.16

Tomaschek, F., Hendrix, P., & Baayen, R. H. (2018). Strategies for addressing collinearity in multivariate linguistic data. *Journal of Phonetics*, *71*, 249-267. https://doi.org/10.1016/j.wocn.2018.09.004

Tounsi, Y., Anoun, H., & Hassouni, L. (2020). CSMAS: Improving multi-agent credit scoring system by integrating big data and the new generation of gradient boosting algorithms. *Proceedings of the 3rd International Conference on Networking, Information Systems & Security (NISS '20), 32*, 1-7. https://doi.org/10.1145/3386723.3387851

Trivedi, S. K. (2020). A study on credit scoring modeling with different feature selection and machine learning approaches. *Technology in Society*, *63*, 101413. https://doi.org/10.1016/j.techsoc.2020.101413

Tsagkarakis, M.-P., Doumpos, M., & Pasiouras, F. (2021). Capital shortfall: A multicriteria decision support system for the identification of weak banks. *Decision Support Systems, 145,* 113526. https://doi.org/10.1016/j.dss.2021.113526

Tsai, C.-F., Hsu, Y.-F., & Yen, D. C. (2014). A comparative study of classifier ensembles for bankruptcy prediction. *Applied Soft Computing Journal*, *24,* 977-984. https://doi.org/10.1016/j.asoc.2014.08.047

Van Gool, J., Verbeke, W., Sercu, P., & Baesens, B. (2012). Credit scoring for microfinance: Is it worth it? *International Journal of Finance and Economics*, *17*(2), 103-123. https://doi.org/10.1002/ijfe.444

Wang, J., & Gribskov, M. (2019). IRESpy: An XGBoost model for prediction of internal ribosome entry sites. *BMC Bioinformatics*, *20*(1), 409. https://doi.org/10.1186/s12859-019-2999-7

Wang, Z., Jiang, C., Ding, Y., Lyu, X., & Liu, Y. (2018). A Novel behavioral scoring model for estimating probability of default over time in peer-to-peer lending. *Electronic Commerce Research and Applications*, *27,* 74-82. https://doi.org/10.1016/j.elerap.2017.12.006

Waring, J., Lindvall, C., & Umeton, R. (2020). Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artificial Intelligence in Medicine, 104,* 101822. https://doi.org/10.1016/j.artmed.2020.101822

Wei, S., Yang, D., Zhang, W., & Zhang, S. (2019). A novel noise-adapted two-layer ensemble model for credit scoring based on backflow learning. *IEEE Access, 7,* 99217-99230. https://doi.org/10.1109/ACCESS.2019.2930332

Wei, Y., Yildirim, P., Van Den Bulte, C., & Dellarocas, C. (2016). Credit scoring with social network data. *Marketing Science, 35*(2), 201-340. https://doi.org/10.1287/mksc.2015.0949

Winter, E. (2002). Chapter 53 The Shapley value. In R. Aumann, & S. Hart (Eds.), *Handbook of game theory with economic applications, vol. 3*. Elsevier, 2025-2054. https://doi.org/10.1016/S1574-0005(02)03016-3

Xia, Y., Liu, C., Da, B., & Xie, F. (2018). A novel heterogeneous ensemble credit scoring model based on bstacking approach. *Expert Systems with Applications, 93,* 182-199. https://doi.org/10.1016/j.eswa.2017.10.022

Xia, Y., Liu, C., Li, Y. Y., & Liu, N. (2017). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications, 78,* 225-241. https://doi.org/10.1016/j.eswa.2017.02.017

Xia, Y., Zhao, J., He, L., Li, Y., & Niu, M. (2020). A novel tree-based dynamic heterogeneous ensemble method for credit scoring. *Expert Systems with Applications, 159,* 113615. https://doi.org/10.1016/j.eswa.2020.113615

Xiong, X., Liu, S., Li, D., Cai, Z., & Niu, X. (2020). A comprehensive survey on local differential privacy. In M. de Rey (Acad. Ed.), *Security and communication networks.* John Wiley & Sons. https://doi.org/10.1155/2020/8829523

Xu, Y., & Goodacre, R. (2018). On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of Analysis and Testing*, *2*(3), 249-262. https://doi.org/10.1007/s41664-018-0068-2

Yang, F., Qiao, Y., Qi, Y., Bo, J., & Wang, X. (2022). BACS: Blockchain and AutoML-based technology for efficient credit scoring classification. *Annals of Operations Research,* 1-21. https://doi.org/10.1007/s10479-022-04531-8

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415,* 295-316. https://doi.org/10.1016/j.neucom.2020.07.061

Yao, J., Wang, Z., Wang, L., Liu, M., Jiang, H., & Chen, Y. (2022). Novel hybrid ensemble credit scoring model with stacking-based noise detection and weight assignment. *Expert Systems with Applications, 198,* 116913. https://doi.org/10.1016/j.eswa.2022.116913

Yeh, I.-C. (2009). Default of credit card clients [Dataset]. UCI Machine Learning Repository. https://doi.org/https://doi.org/10.24432/C55S3H

Yu, L., Yu, L., & Yu, K. (2021). A high-dimensionality-trait-driven learning paradigm for high dimensional credit classification. *Financial Innovation, 7*(1), Art. 32. https://doi.org/10.1186/s40854-021-00249-x

Yu, L., Zhou, R., Chen, R., & Lai, K. K. (2022). Missing data preprocessing in credit classification: One-Hot encoding or imputation? *Emerging Markets Finance and Trade*, *58*(2), 472-482. https://doi.org/10.1080/1540496X.2020.1825935

Zaidi, N. A., Webb, G. I., Carman, M. J., Petitjean, F., & Cerquides, J. (2016). ALR^n: Accelerated higher-order logistic regression. *Machine Learning*, *104*, 151-194. https://doi.org/10.1007/s10994-016-5574-8

Zhang, Q., Yang, L. T., & Chen, Z. (2016). Deep computation model for unsupervised feature learning on big data. *IEEE Transactions on Services Computing*, *9*(1), 161-171. https://doi.org/10.1109/TSC.2015.2497705

Zhang, W., He, H., & Zhang, S. (2019). A novel multi-stage hybrid model with enhanced multi-population niche genetic algorithm: An application in credit scoring. *Expert Systems with Applications*, *121*, 221-232. https://doi.org/10.1016/j.eswa.2018.12.020

Zhou, B., Bartholmai, B. J., Kalra, S., & Zhang, X. (2020). Predicting lung mass density of patients with interstitial lung disease and healthy subjects using deep neural network and lung ultrasound surface wave elastography. *Journal of the Mechanical Behavior of Biomedical Materials*, *104*, 103682. https://doi.org/10.1016/j.jmbbm.2020.103682

Zhu, G., & Zhao, T. (2021). Deep-gKnock: Nonlinear group-feature selection with deep neural networks. *Neural Networks, 135,* 139-147. https://doi.org/10.1016/j.neunet.2020.12.004

Zhu, T., Wang, B., Wu, B., & Zhu, C. (2011). Role defining using behavior-based clustering in telecommunication network. *Expert Systems with Applications, 38*(4), 3902-3908. https://doi.org/10.1016/j.eswa.2010.09.051

Zou, Q., Xie, S., Lin, Z., Wu, M., & Ju, Y. (2016). Finding the best classification threshold in imbalanced classification. *Big Data Research, 5,* 2-8. https://doi.org/10.1016/j.bdr.2015.12.001

Appendix

Table 0-1 provides a view of the data sourced from Yeh (2009), accompanied by their corresponding p-values and a rejection status. This includes a diverse range of financial metrics, demographic details, and repayment statuses. Several predictor variables, such as those related to average payment amounts, standard deviations, and ratios, exhibit statistically significant p-values, indicating their importance in predicting the outcome. Conversely, certain variables, particularly those associated with repayment statuses in later months and specific bill amounts, show p-values of 1.000, suggesting a lack of statistical significance. The rejection column provides clarity on whether the null hypothesis is rejected based on the p-value, offering insights into the relevance of each predictor in the context of the study.

Table 0-1 - Yeh, (2009) data - variables and their p-values

Variable	p-value	Reject
AVG_PAY__SEP	0.000	No
CURRENT_OVER_3MAVG_PAY__SEP	0.000	No
STD_PAY__SEP	0.000	No
AVG_PAY__JUN	0.008	No
AVG_BILL_AMT_SEP	0.000	No
AVG_PAY_AMT_SEP	0.000	No
CURRENT_OVER_3MAVG_BILL_AMT_SEP	0.000	No
LIMIT_BAL	0.000	No

SEX	0.000	No
EDUCATION	0.000	No
MARRIAGE	0.000	No
AGE	0.036	No
PAY_1	0.000	No
PAY_2	1.000	Yes
PAY_3	1.000	Yes
PAY_4	1.000	Yes
PAY_5	1.000	Yes
PAY_6	0.006	No
BILL_AMT1	0.003	No
BILL_AMT2	0.000	No
BILL_AMT3	1.000	Yes
BILL_AMT4	1.000	No
BILL_AMT5	1.000	Yes
BILL_AMT6	1.000	Yes
PAY_AMT1	0.000	No
PAY_AMT2	1.000	Yes
PAY_AMT3	1.000	Yes
PAY_AMT4	1.000	Yes
PAY_AMT5	1.000	Yes

PAY_AMT6	1.000	Yes
AVG_PAY_AMT_JUN	1.000	Yes
STD_PAY_AMT_JUN	0.000	No
CURRENT_OVER_3MAVG_PAY_AMT_JUN	0.000	No
AVG_BILL_AMT_JUN	1.000	Yes
STD_BILL_AMT_JUN	0.520	Yes
CURRENT_OVER_3MAVG_BILL_AMT_JUN	0.733	Yes
STD_PAY_JUN	0.001	No
CURRENT_OVER_3MAVG_PAY_JUN	0.000	No
AVG_PAY_AMT_JUL	0.000	No
STD_PAY_AMT_JUL	0.005	No
CURRENT_OVER_3MAVG_PAY_AMT_JUL	0.000	No
AVG_BILL_AMT_JUL	0.000	No
STD_BILL_AMT_JUL	0.460	Yes
CURRENT_OVER_3MAVG_BILL_AMT_JUL	0.389	Yes
AVG_PAY_JUL	1.000	Yes
STD_PAY_JUL	0.795	Yes
CURRENT_OVER_3MAVG_PAY_JUL	0.000	No
AVG_PAY_AMT_AUG	1.000	Yes
STD_PAY_AMT_AUG	0.327	Yes

CURRENT_OVER_3MAVG_PAY_AMT_AUG	0.000	No
AVG_BILL_AMT_AUG	0.000	No
STD_BILL_AMT_AUG	0.188	Yes
CURRENT_OVER_3MAVG_BILL_AMT_AUG	0.105	Yes
AVG_PAY_AUG	1.000	Yes
STD_PAY_AUG	0.007	No
CURRENT_OVER_3MAVG_PAY_AUG	0.000	No
STD_PAY_AMT_SEP	0.001	No
CURRENT_OVER_3MAVG_PAY_AMT_SEP	0.098	Yes
STD_BILL_AMT_SEP	0.584	Yes

The predictor variables used to address research questions 1 and 2 from the Home Credit Group (2018) dataset, as detailed in Table 0-2, align with the predictor variables derived from Hlongwane et al. (2024a).

Table 0-2 - Home Credit Group (2018) predictor variables

Variable	**p-value**	**Reject**
AVG_EXT_SOURCE	0.042	No
EXT_SOURCE_3	0.000	No
EXT_SOURCE_2	0.000	No
EXT_SOURCE_1	0.000	No

BUREAU_DAYS_CREDIT_MAX	0.000	No
INSTAL_DPD_MEAN	0.000	No
DAYS_EMPLOYED	0.000	No
APPS_ANNUITY_CREDIT_RATIO	0.000	No
APPROVED_AMT_ANNUITY_MEAN	0.000	No
INSTAL_DAYS_ENTRY_PAYMENT_MAX	0.000	No
APPROVED_AMT_CREDIT_MAX	0.000	No