Machine learning for retail credit risk scoring: a systematic literature review with insights for South African banks

V Sithebe

orcid.org0009-0005-9904-8171

Mini-dissertation accepted in partial fulfillment of the requirements for the degree *Master of Commerce in Applied Risk Management* at the North-West University

Supervisor: Mr TAP van den Berg

Co-supervisor: Dr JF Goede

Graduation: May 2026

PREFACE

This mini-dissertation is the final deliverable for the Master of Commerce (MCom) in Applied Risk Management. The mini-dissertation was written in article format and consists of three sections: Research project overview; Article; and Reflection.

This mini-dissertation is the student's work. The student was responsible for the final concept, set up, execution of the research project and writing of the mini-dissertation. The supervisory team members contributed in an advisory and technical support capacity to the study's conception and design, analysis and interpretation of data, and critical revision of the manuscript. The mini-dissertation was language edited before submission for examination. However, the student is responsible for making these edits and for the grammatical correctness of the final document.

I declare that this mini-dissertation was done according to the NWU Guidelines on Responsible and Ethical Use of Artificial Intelligence (https://www.nwu.ac.za/sites/www.nwu.ac.za/files/files/i-governance-management/policy/2024/November-2024/2P_2.4.3.2_Policy-on-Academic-Integrity.pdf)

The primary study supervisor permitted the student to submit this mini-dissertation for examination.

ABSTRACT

Credit scorecards remain central to bank lending, yet modern credit datasets increasingly demand models that capture non-linear, high-dimensional patterns. Traditional models such as logistic regression are increasingly constrained by their linear assumptions, particularly in emerging markets like South Africa where many borrowers have thin or fragmented credit histories. This study systematically reviews 32 peer-reviewed papers to examine how machine learning (ML) can be applied to retail credit scoring while maintaining the transparency requirements mandated by regulators and required by auditors. The study also conducts a performance comparison between ML approaches and barriers to adoption. Findings show that tree-based ensemble methods (Random Forest, XGBoost, LightGBM and CatBoost) consistently outperform traditional approaches in accuracy and stability while neural networks and support vector machines also perform well but raise transparency challenges. Explainable artificial intelligence (AI) techniques, especially SHapley Additive exPlanations (SHAP), emerge as practical tools to bridge predictive power with auditability. The review concludes that South African banks can adopt a staged, hybrid approach: using ML in data preparation and segmentation while retaining interpretable decision layers, thereby enhancing predictive accuracy and financial inclusion without undermining regulatory transparency.

Keywords: machine learning, retail credit risk, risk scoring models

ACKNOWLEDGEMENTS

All praise is to the Almighty, the Sovereign. He advances whom He wills by His grace and delays whom He wills by His justice. I would like to thank the woman beneath whose feet lies paradise, my mother. I would like to thank my beautiful brothers for their patience and love for me even in my most stressful times. My colleagues who were always excited to share their views and my employer for seeing value in me and funding my studies. My supervisors who were always generous with their time and insight. Thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

Acronym	Definition
AdaBoost	Adaptive Boosting
AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under the Receiver Operating Characteristic Curve
BCBS	Basel Committee on Banking Supervision
CatBoost	Categorical Boosting
CNN	Convolutional Neural Network
DHET	Department of Higher Education and Training
FSB	Financial Stability Board
FSCA	Financial Sector Conduct Authority
IFRS	International Financial Reporting Standards
KS	Kolmogorov–Smirnov Statistic
Light GBM	Light Gradient Boosting Machine
LIME	Local Interpretable Model-agnostic Explanations
ML	Machine Learning
MLP	Multilayer Perceptrons
NCA	National Credit Act
NWU	North-West University
PD	Probability of Default
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RF	Random Forest
RNN	Recurrent Neural Network
SARB	South African Reserve Bank
SHAP	SHapley Additive exPlanations
XAI	Explainable Artificial Intelligence
XGBoost	Extreme Gradient Boosting
AdaBoost	Adaptive Boosting

RESEARCH PROJECT OVERVIEW

Credit scoring lies at the heart of the credit risk management process, as it enables banks to evaluate the creditworthiness of applicants and make informed lending decisions. This systematic literature review investigates the application of Machine Learning (ML) techniques in credit scoring, with a focus on how they compare to traditional statistical models and the challenges that limit their adoption in South African banks. The South African banking environment is highly regulated, and while global institutions are increasingly experimenting with ML-based credit models, local banks have adopted a more cautious approach due to regulatory, governance, and transparency considerations.

The study was borne out of the researcher's professional experience in credit risk, where exposure to the inner workings of credit models and governance frameworks highlight a growing tension between innovation and compliance. Further, observing how international institutions were advancing towards ML-driven credit scoring while South African banks largely relied on traditional logistic regression models prompted the researcher to explore this gap further. The motivation for this study therefore stemmed from both professional curiosity and the practical need to understand how ML can be responsibly introduced into local banking credit processes without undermining regulatory expectations on transparency.

The study aims to provide a balanced and evidence-based view of ML's role in modern credit scoring by highlighting its advantages in predictive accuracy, the barriers to its use, and how explainable AI techniques can make ML outputs interpretable for auditors and supervisors. The paper also provides a practical roadmap for South African banks -including the South African Reserve Bank- to consider hybrid adoption strategies that leverage ML in data preparation and feature engineering, while maintaining transparency in decision-making.

This mini-dissertation is presented in three parts. The first section, this Research Project Overview, introduces the research topic, explains the rationale for selecting it, and outlines how each section of the dissertation contributes to a coherent storyline. The second section, the Research Article, presents the PRISMA-aligned systematic literature review of 32 peer-reviewed studies, detailing how ML models perform relative to traditional methods, and concludes with recommendations tailored to the South African context. The third section, the Reflection, evaluates the study's contribution to the South African banking sector, summarises the key findings in a dashboard format, and discusses the researcher's personal learning journey and lessons for future research.

Fit within the field of risk management

Credit scoring directly influences credit risk modelling, provisioning, and capital adequacy –these are core elements of a bank's risk management framework. As banking data becomes increasingly complex, the ability to use ML for deeper insights, improved default prediction, and enhanced portfolio segmentation introduces a new frontier in managing risk. This study therefore sits at the intersection of data science and financial risk management, contributing to the ongoing discussion on how innovation and regulation can coexist responsibly.

Additionally, this study provided the researcher with an opportunity to apply academic principles to a practical problem faced by the financial sector, demonstrating mastery of applied research at master's degree level within a research team. The responsibilities of the different role players in this research project are described in the table below:

Table 1: Research Roleplayers

#	Team Member	Role
1	Researcher	Led the research design, conducted the systematic literature review, applied thematic coding and synthesis, developed the academic article, and integrated supervisor feedback.
2	Supervisor: Mr TAP van den Berg	Provided academic direction, ensuring methodological soundness and alignment with research objectives, and guided the framing of the study within the field of risk management.
	Co-supervisor: Mr JF Goede	Offered specialised methodological and technical guidance, particularly in research design, synthesis methods, and ensuring academic coherence and rigour.
3	Editor	Conducted a grammar-only edit of the dissertation before submission for examination.

ARTICLE

Machine learning for retail credit risk scoring: a systematic literature review with insights for South African banks

Introduction

In South Africa, where household debt has grown at an average annual rate of 5.2 per cent between 2015 and 2022, the role of retail lending by banks is particularly significant, accounting for 75 per cent of this debt (Muneri & Kuhn, 2023). Central to this is the bank's ability to distinguish between low and high-risk borrowers as accurate risk assessments directly impact profitability, capital adequacy and compliance with responsible lending requirements under the National Credit Act (NCA) (Agung Dharmawan, 2020; Çallı & Coşkun, 2021; Chen et al., 2016; Gross, 2020; South African Reserve Bank, 2005).

Retail lending is simply the loans banks offer to everyday consumers, instead of corporates, to fulfil a financial need (Mohua, 2006), and credit risk refers to the potential that borrowers will fail to meet their debt obligations, posing a risk to the lender (Neal, 1996). If we combine the two, retail credit risk is the credit risk that arises when banks lend to everyday consumers whereas wholesale credit risk deals with large, negotiated loans offered to businesses (Allen et al., 2004). This study focuses on the retail credit risk. The retail lending process involves the bank and the applicant; the process goes from the initial loan application to assessment to the loan's successful repayment or default (the customer fails to make the repayments) (Kočenda & Vojtek, 2011).

Assessing lending risk

To manage this risk, lenders rely on credit scoring models, often in the form of application scorecards, which assess an applicant's creditworthiness by predicting their probability of default through basing their outcome from the data in the loan application form, behavioural data from the applicant's transactional bank account and, oftentimes, the credit bureau (Abdou & Pointon, 2011; Bijak & Thomas, 2012; Crook et al., 2007). An effective scoring system ensures that deserving applicants receive credit while protecting the lender from potential losses (Baesens et al., 2016; Blöchlinger & Leippold, 2006).

Accurate credit scoring plays a strategic role in customer segmentation, which Makuch (2001) defines as "the process of identifying homogeneous populations with respect to their predictive relationship". In credit risk terms, this involves grouping applicants by shared characteristics, such as credit history length (cohorts) or the depth of available data (thin vs. thick credit files) (Bijak & Thomas, 2012; Makuch, 2001). Effective segmentation enables tailored risk-based strategies, such as differentiated treatment policies for high and low-risk customers, ultimately improving portfolio performance. By understanding the distinct risk profiles of each segment, banks can also refine product offerings and personalise interactions to better meet borrower needs (Hand & Henley, 1997). Moreover, accurate scoring supports risk-based pricing, allowing borrowers with low probability to

default to access competitive interest rates while higher-risk clients are charged rates that offset potential losses (Rosenberg & Gleit, 1994), thereby optimising the risk–return trade-off.

However, traditional scoring methods, such as logistic regression, often rely on linear assumptions and structured data, which may not fully capture the complex, non-linear behaviours of modern consumers (Dastile et al., 2020; Leo et al., 2019). This limitation is particularly pronounced in emerging markets like South Africa. The nation's unique socio-economic landscape, which is characterised by high income inequality, a large informal economy and fragmented consumer credit histories, present distinct challenges for conventional credit risk models (Mutsonziwa & Fanta, 2021; Simatele & Maciko, 2022; Stats SA, 2025). Consequently, there is a risk that existing models may fail to accurately assess affordability or may inadvertently exclude viable applicants who operate outside the formal financial system, a concern amplified by the National Credit Act's (NCA) mandate for responsible lending (South African Reserve Bank, 2005).

In response to these challenges, and considering the importance of accurate scoring, the global financial industry has increasingly turned to machine learning (ML) (Leo et al., 2019). Machine learning algorithms offer the potential to analyse vast and complex datasets, identify non-linear patterns and deliver more accurate predictions than their traditional counterparts (Lessmann et al., 2015; Martens et al., 2007). However, the adoption of ML is not without obstacles: the primary concern is the 'black-box' nature of many advanced models, which makes it difficult to interpret how ML decisions are made and is a major drawback to its adoption (Bücker et al., 2022; Martens et al., 2007). This lack of transparency poses a challenge in a highly regulated environment like South Africa where concerns raised by the Financial Sector Conduct Authority (FSCA) on model transparency (difficulty in decomposing the output of an ML model into the underlying drivers of its decision) are well understood, given the opaque nature of these models (Hlophe, 2024). Similarly, the Financial Stability Board (FSB) stated that "the use of complex algorithms could result in a lack of transparency to consumers. This 'black box' aspect of machine learning algorithms may in turn raise concerns." (Schindler et al., 2017).

This dissertation, therefore, identifies a critical tension: while ML offers a powerful solution to improve credit scoring accuracy in the complex South African market, its adoption is hindered by regulatory and practical concerns regarding interpretability. There remains a need for a harmonisation of existing literature on how ML can enhance accuracy and transparency in South Africa, considering the aforementioned concerns raised by the FSB and FSCA (Hlophe, 2024; Kumar et al., 2021; Schindler et al., 2017). This study will investigate how this tension can be resolved by exploring how ML techniques that are applied in developed markets can inform and potentially improve credit risk scoring models in emerging markets, specifically in South Africa.

To achieve this, the study will systematically review literature on the application of ML in retail credit scoring, with a focus on emerging markets and South Africa. It seeks to (1) identify limitations of traditional credit scoring models, (2) assess ML's ability to improve accuracy while maintaining transparency and (3) propose recommendations for the responsible adoption of ML in South Africa. Therefore, the central research question is, "How can machine learning enhance traditional retail credit risk scoring models' accuracy in South Africa while maintaining transparency?"

The paper begins by outlining the background to credit scoring and the adoption of machine learning, followed by a description of the systematic literature review methodology. It then presents and discusses the key findings before concluding with practical recommendations and directions for future research.

Background

Globally, household debt has been on an upward trend in recent decades (Edelberg, 2006), while factors such as indebtedness ratios, unemployment risk and household size have been linked to increased default rates in emerging economies (Madeira, 2023). Further, the September 2023 bulletin published by the South African Reserve Bank (SARB) highlights that the outstanding balance of household debt in South Africa grew at an average annual rate of 5.2 per cent from 2015 to 2022, of which loans granted by banks accounted for 75 per cent (Muneri & Kuhn, 2023). Therefore, in an emerging economy such as South Africa, managing retail credit risk is fundamental, and building a scorecard that accurately predicts the risk is crucial.

Determining the credit history of a borrower dates back to the 1940s and 50s (Dastile et al., 2020; Poon, 2007); this emerged from the need to assess credit risk systematically (Abdou & Pointon, 2011). Initially, credit decisions were made by using a judgmental approach, known as the 5Cs: Character, Capital, Collateral, Capacity and Condition (Dastile et al., 2020; Thomas et al., 2017). This approach proved inadequate for processing large volumes of applications, leading to the development of statistical scoring methods (Dastile et al., 2020). The first commercial credit scorecards emerged in the US in various retail and financial services firms, the most notable being introduced in 1958 by Fair, Isaac & Company, initially for American Investment before expanding its use in the company's 800 operations nationwide (Poon, 2007). The drive towards automated credit scoring was accelerated by the dramatic growth in consumer credit and the need for faster, more consistent decision-making processes, and these automated scorecards are the traditional scorecards being used today (Abdou & Pointon, 2011; Li et al., 2019).

Traditional scorecards are structured, statistically-derived models built to translate applications and bureau data into a single risk score – logistic regression being the industry standard and the basis for most points-based scorecards being used in retail lending (Siddiqi, 2017, p. 26). They typically draw on a limited set of structured inputs (application form fields, bureau records and selected internal performance indicators) that are grouped (or binned) into discrete attributes, and each group is assigned a weight or points contribution that sums to the final risk score (Baesens et al., 2016, p. 75; Siddiqi, 2017, p. 6).

Scorecards are designed to be operationally simple: a single numeric score permits cut-off rules for accept/decline decisions, risk-based pricing and portfolio segmentation and supports straightforward explanations to customers, auditors and regulators (Siddiqi, 2017, p. 155). However, these conventional models have several practical limitations when applied to modern, data-rich and volatile

credit markets, including that they typically assume linear or additive relationships between predictors and outcomes, which can miss important non-linear interactions present in borrower behaviour (Baesens et al., 2016, p. 111; Siddiqi, 2017, p. 56).

Understanding the stages of scorecard development enables the identification of the stages ML can be effectively embedded to improve accuracy and address transparency challenges. Therefore, to understand where ML can be effectively embedded, it is essential to first outline the traditional scorecard development process, which generally takes on 5 main steps:

1. Data gathering and preparation

This is where historical loan performance data is collected and cleaned, and definitions like what distinguishes a "good" from a "bad" applicant are drawn out by overlaying the performance data with other sources, such as other client information that is contained in the application form and credit bureau data (Martin, 2013; Ralf et al., 2024; Siddiqi, 2017). Data quality is paramount, as evidenced by the lessons learnt from the 2007-2009 Global Financial Crisis, which revealed significant inadequacies in banks' data architectures for risk management (Abdou & Pointon, 2011; Basel Committee on Banking Supervision, 2013); therefore, gathering data is a resource-intensive process where the data is not only assessed for completeness and consistency but also for its relevance, and this challenge is amplified in South Africa by concerns around data sufficiency and the potential for identity theft (Basel Committee on Banking Supervision, 2013; Siddiqi, 2017). Today, banks must adhere to strict data quality standards, including the Basel Committee on Banking Supervision's (BCBS) Principles for effective risk data aggregation and risk reporting (Basel Committee on Banking Supervision, 2013).

2. Definition of default

A clear and consistent definition of what constitutes a "default" is fundamental in developing credit scoring models as it determines which customers are classified as "bad" and directly shapes the model's predictive target. In practice, definitions of default vary depending on regulatory frameworks and company policy; for instance, some institutions may define default as 90 days past due while others may use the write-off status or legal recovery actions as thresholds (Harris, 2013; Siddiqi, 2017). Regulatory frameworks such as Basel II/III and IFRS 9 emphasise the importance of a precise, objective and consistent definition of default to ensure comparability and robustness across all scoring models (Basel Committee on Banking Supervision, 2004; International Accounting Standards Board, 2014); therefore, this stage is critical as an inconsistent or overly broad definition can distort the predictive power and reliability of the final scorecard.

3. Segmentation

Segmentation involves dividing the application population into distinct groups based on their shared characteristics to improve model accuracy. Generally, a separate scorecard is developed for each segment (Bijak & Thomas, 2012); however, in smaller markets where customer volume is insufficient to motivate for multiple scorecards, the segments are identified, yet one unified model is developed (Siddiqi, 2017, p. 46). Segmentation is another step in the building process, where ML can offer advantages by detecting non-obvious clusters through unsupervised learning techniques like k-means or hierarchical clustering (Lessmann et al., 2015).

4. Feature selection and transformation

Feature selection follows what is important in identifying the most predictive variables and improves model generalisability by reducing overfitting (Ashofteh & Bravo, 2021). Feature importance and selection refers to techniques used to measure how useful or valuable each input variable (feature) is in predicting the target variable (Liu & Schumann, 2005); so, in credit scoring, a feature could be income, credit history and number of credit enquiries for, example, while importance will tell you which of these inputs contributes the most to the model's prediction (Laborda & Ryoo, 2021). Credit analysts typically evaluate 50-60 potential features before ultimately selecting 8-12 that provide the most predictive combination (Huang et al., 2018). This step is often intensive for traditional methods that employ feature importance scores or statistical tests to remove redundant or irrelevant variables (Xia et al., 2017). In some advanced models, feature selection is automated and can adapt to changing data distributions (Ashofteh & Bravo, 2021), and the literature has shown that ML meaningfully enhances and automates this step by identifying the most relevant variables and improving model performance (Tripathi et al., 2019; Tripathi et al., 2018) .

5. Model development and validation

Model development comprises selecting a statistical or ML approach that is suited to the problem structure and data characteristics. Typically, techniques such as logistic regression or linear discriminant analysis are used (Bücker et al., 2022), and, more recently, ML models such as random forests, neural networks and ensemble methods have gained attention for their potential to improve predictive performance (Dastile et al., 2020; West, 2000; Xia et al., 2017). Recent literature explores hybrid approaches that combine ML with traditional models to maintain transparency while improving predictive power (Bücker et al., 2022; Wang & Lin, 2021). The developed model is then validated by using metrics such as the Gini coefficient/ Area Under the Curve for predictive power (Lessmann et al., 2015; Martin, 2013; Óskarsdóttir et al., 2019) and stability indices like the Characteristic Stability Index and Population Stability Index for monitoring changes in data or model performance (Martin, 2013; Whittaker et al., 2007).

6. Learning machines to assess risk

A more detailed understanding of patterns and correlations in large datasets is made possible by ML approaches (Lessmann et al., 2015; Mestiri, 2024). Machine learning, a subset of artificial intelligence (AI), involves algorithms that can learn from and make decisions based on data without being explicitly programmed to perform specific tasks, such as credit assessment or fraud detection. In the context of credit scoring, ML models can analyse complex datasets with high accuracy by capturing non-linear relationships and adapting to new information more efficiently than the traditional methods mentioned above (Wenbing et al., 2019).

Research has shown how ML may increase prediction accuracy and lower false positives in credit evaluations (Lessmann et al., 2015). For instance, research by Lessmann et al. (2015) compared several ML models, including support vector machines, neural networks and ensemble methods and found that these models markedly outperformed conventional scoring techniques. Other global studies, such as those by Addo et al. (2018), explored the use of deep learning (DL) in credit risk management and highlighted its ability to handle large, unstructured datasets and improve long-term predictions. The datasets used by Lessmann et al. (2015) were drawn mainly from European and Australian sources, while Addo et al. (2018) used data from European organisations. These studies show us that ML methods have proven to outperform traditional techniques, such as logistic regression, by offering more precise risk prediction (Dastile et al., 2020; Dumitrescu et al., 2022).

Among the most commonly applied ML techniques in credit scoring are Random Forests, Gradient Boosting Machines and Artificial Neural Networks (ANNs) (Breiman, 2001; Brown & Mues, 2012; Lee et al., 2002). Random Forests, an ensemble method, aggregate the outputs of multiple decision trees to improve generalisation and reduce overfitting (Breiman, 2001) while gradient boosting, another ensemble technique, builds models sequentially by correcting previous errors to improve performance (Brown & Mues, 2012; Friedman, 2002). Artificial Neural Networks have also been applied to credit scoring due to their ability to uncover intricate patterns in large datasets (Addo et al., 2018); although they have been criticised for their “black-box” nature (Bücker et al., 2022; Lessmann et al., 2015).

Notwithstanding their strengths in accuracy, a central concern regarding ML models remains their transparency and explainability, thus the rise of explainable AI (XAI) techniques, such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) has allowed practitioners the opportunity to better understand which input features are driving individual predictions, thereby increasing transparency (Houda et al., 2022; Lundberg & Lee, 2017), which is

particularly important in a regulated domain such as credit risk. Accordingly, ML is increasingly viewed not as a replacement for traditional models, but as a complementary layer that is used for feature engineering, segmentation or boosting prediction accuracy within hybrid scoring frameworks (Boughaci et al., 2020; Wang & Lin, 2021).

Traditional model enhancement opportunities

The growth of digital financial services has meant that large volumes of customer data, including transactional data and behavioural patterns, are now accessible to lenders (Gomber et al., 2017). This has resulted in a change in the credit scoring landscape by presenting an opportunity to develop more complex models that can consider the complexity of modern credit markets (Roa et al.). As credit markets become more complex and consumer data becomes more fragmented, indeed, there is a growing need for models that can analyse large, unstructured datasets and uncover non-linear relationships (Martens et al., 2007). Machine Learning models offer us a potential solution to these challenges; their ability to handle vast datasets, detect patterns in customer behaviour and predict future trends makes them a valuable tool for modern credit risk assessment (Okeleke et al., 2024).

Credit risk managers need to find ways to balance performance with interpretability; solutions like hybrid models, which combine the traditional models with ML components, have been proposed as a way to mitigate the explainability issues while still benefiting from the strengths of ML (Bücker et al., 2022; Martens et al., 2007).

Hybrid models that combine the predictability of ML with the interpretability of conventional credit scoring methods have been explored in recent research (Dumitrescu et al., 2022; Wang & Lin, 2021). In these models, the ML component typically handles the more complex tasks, such as segmentation (dividing borrowers into distinct groups based on characteristics, then identifying homogenous subsets within the larger population) and classification (assigning borrowers to predefined categories such as good/bad credit, or, approve/decline) of individuals while the traditional component provides a clearer, rule-based framework for decision-making (Boughaci et al., 2020). Research by Dumitrescu et al. (2022) developed the Penalised Logistic Tree Regression (PLTR), a hybrid credit scoring approach that combines decision tree-derived predictors with logistic regression and found that this method improves predictive performance while maintaining interpretability (Dumitrescu et al., 2022); the model has also shown cost reductions ranging from 18-22 per cent compared to linear logistic regression.

These hybrid models present a promising way forward for credit risk managers in markets like South Africa, where regulatory oversight is stringent, and credit behaviour is highly diverse due to historical inequalities, varying income levels and widespread financial exclusion (Simatele & Maciko, 2022). Hybrid models combine the best features of both methods to provide more precise and nuanced risk evaluations without compromising explainability (Lessmann et al., 2015); this is especially important when evaluating subpopulations whose credit histories are sparse or whose credit behaviour patterns are inconsistent since they might be better represented by ML's capacity to interpret unstructured data.

South African regulatory landscape

The application of any scoring model in South Africa is governed by the National Credit Act (NCA) gazetted in 2005 (South African Reserve Bank). The NCA was introduced to promote responsible lending practices, ensure consumer protection and foster a fair and transparent credit market. Central to the NCA are specific provisions that influence the development and implementation of credit scorecards: Section 81 ('prevention of reckless credit') is particularly salient as it prohibits credit providers from entering into a credit agreement without taking reasonable steps to assess the applicant's affordability. Section 170 ('provider to keep records') emphasises maintaining accurate and comprehensive records of credit agreements and applications to ensure transparency and accountability in lending decisions (South African Reserve Bank, 2005). As such, the Act does not prohibit specific technologies like ML; rather, its principles-based approach requires that any method used to assess an applicant's affordability must be fair, transparent and justifiable. Naturally, the NCA has strong implications for the type of credit scoring models that can be implemented by credit providers.

This legal framework operates within a challenging socio-economic environment. Many consumers in South Africa operate within the informal economy (Stats SA, 2025), making it challenging to assess creditworthiness by using conventional variables, such as payslips or tax records; additionally, credit histories are often thin or fragmented (known as "thin" credit files), thereby challenging the effectiveness of scorecards that rely heavily on credit bureau data (Bijak & Thomas, 2012; Mutsonziwa & Fanta, 2021; Siddiqi, 2017, p. 48). In light of these regulatory requirements and data limitations, it becomes necessary to explore modelling approaches that accommodate the complexities of South African credit markets while aligning with transparency expectations. The combination of evolving data ecosystems, regulatory scrutiny and the distinct characteristics of emerging-market consumers makes South Africa ideal for investigating the applicability and value of ML in enhancing credit scoring.

Method

Research approach

The study employs a systematic literature review approach to address the following question:

"What learning can be obtained from academic literature between 2000 and 2025 on the key findings, challenges and best practices in applying Machine Learning to retail credit risk scoring, particularly in enhancing model accuracy and transparency in emerging markets, and how might these insights be applied in the South African context?"

This question guided the selection of studies, the identification of themes and the synthesis of findings to provide insights for South African banks. A review protocol was developed according to the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) framework (www.prisma-statement.org) and used to track the identification, screening, eligibility and inclusion of sources. The PRISMA approach is widely accepted best practice for conducting systematic literature reviews and ensures reviews are transparent, replicable and comprehensive (Page et al., 2021; Parums, 2021).

Search strategy

A comprehensive search was conducted on 14 June 2024. Scopus was selected as the primary database as it contained comprehensive coverage of peer-reviewed literature in the fields of finance, ML and credit risk. Although other databases were considered (JSTOR, EBSCO), Scopus yielded a substantial number of relevant articles that aligned to the study's inclusion criteria and was thus deemed sufficient for the purpose of this review. The following terms were used as free-text words in the article title, abstract and keywords: 'machine' and 'learning' and 'credit' and 'scoring' and 'models'. The search was limited to articles published after January 2000, and only articles in English were accepted.

Selection criteria

A multi-stage screening process, tracked using a flow diagram, was used to identify eligible studies. For inclusion in the review, articles needed to meet all of the following criteria (refer **Figure 1**): (1) the title contains 'credit scoring', (2) the study address retail credit lending (e.g. personal loans and credit cards), (3) the study is a primary empirical study and not a literature review or conceptual paper, (4) the study does not introduce a novel methodology/ framework and, (5) the study is accredited by the Department of Higher Education and Training (DHET). As such, the exclusion criteria were the inverse – non-retail credit, literature reviews, papers proposing novel frameworks and not being DHET accredited.

Selection procedure

The initial search results from Scopus were exported to Microsoft Excel. First, the article titles were filtered for words containing 'credit scoring'. Following this, the author screened the remaining titles and abstracts against the full inclusion criteria. A tracking sheet was maintained for all records that details the status of each article, including the reason for exclusion where applicable (refer **Appendix A**). Full-text articles were retrieved via the North-West University (NWU) online library; papers that were inaccessible were excluded.

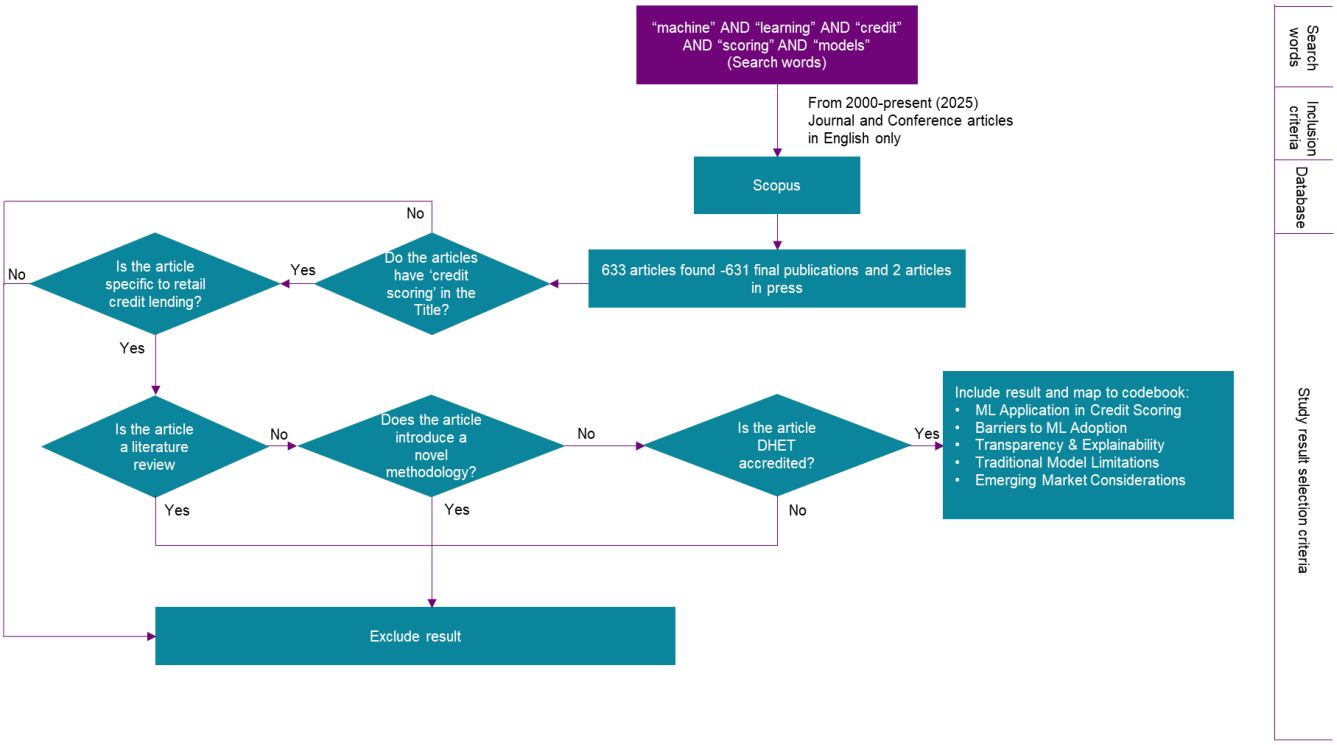

Figure 1: Study selection procedure
Source: Own compilation (2025)

Data extraction

All papers were extracted from the NWU Library and studied for analysis. To aid in visualising and understanding the information, a structured data extraction form was developed to systematically collect information from each study (refer **Appendix B**). The following variables of interest were extracted:

- ***Primary Outcomes (model performance and other metrics):*** the papers were searched for quantitative metrics that relate to predictive accuracy. Measures of predictive accuracy are vast; therefore, this search included common metrics/terms such as:

- *Area Under the receiver operating Curve (AUC)* – measures how well a model distinguishes between good and bad borrowers, with higher values showing better separation – it summarises a classifier's rank discrimination across all thresholds (Kraus, 2014).
- *Gini Coefficient* – is a simple linear transformation of AUC (Gini = 2·AUC − 1) and is widely reported in the credit-risk practice as an intuitive measure of discriminatory power (Guégan & Hassani, 2018).
- *Accuracy* – the proportion of correct predictions (both good and bad borrowers) out of all predictions made (Fang & Chen, 2019).
- *Precision, Recall and the F1-score* – describe trade-offs between false positives and false negatives for a selected operating threshold and are particularly useful when class distributions are imbalanced or when the business cost of different error types is asymmetric (Montevechi et al., 2024).
- *Kolmogorov-Smirnov (KS) statistic* – measures the maximum separation between the cumulative score distributions of 'goods' and 'bads' and is commonly used by practitioners for population separation checks (Fang & Chen, 2019).

- ***Secondary Outcomes (model accuracy and transparency/ XAI):*** In line with the research, papers that included information that relates to model explainability were sought. This included the name of the interpretability technique used (e.g. SHAP, LIME) as well as discussions that relate to the trade-off between accuracy and transparency.
- ***Other variables:*** Other variables were extracted from the papers as well as the tracking template (refer **Appendix A**), including (a) the author(s) and publication year; (b) the journal and article title; (c) on which data the study was based (geographical context); (d) the type of data (public/ private datasets); (e) the ML models tested; (f) in the case of benchmark studies, the benchmark model; and (g) the data preparation techniques. An 'Author's Notes' column was added to the table for simplified context.

Where a study did not report on a specific variable, the variable was recorded as 'Not Applicable'.

Addressing bias and certainty

The risk of reporting bias (outcome reporting bias), where studies with findings that align with the research aim are more likely to be considered, is acknowledged as an inherent limitation of the literature review; this is considered when drawing final conclusions. In terms of certainty in evidence, conclusions that are supported by multiple studies are presented with a higher degree of certainty. Lastly, the study was submitted to the Economic and Management Sciences Research Ethics Committee (EMS-REC) and was classified as a "no risk" study.

Results and Discussion

Study selection

The searches initially resulted in a total of 633 references, with 1 duplicate that was eliminated. After screening the titles and abstracts, 520 articles were excluded as they did not meet either one, or all, of the inclusion criteria. The full texts of the remaining 30 articles were then retrieved. Papers that were issued by the International Financial Stability Board and the European Commission were added manually for regulatory breadth, given that no position papers have been issued by South African regulatory bodies on the use ML in credit scoring. **Figure 2** presents the details of the selection process in a PRISMA-style flow diagram.

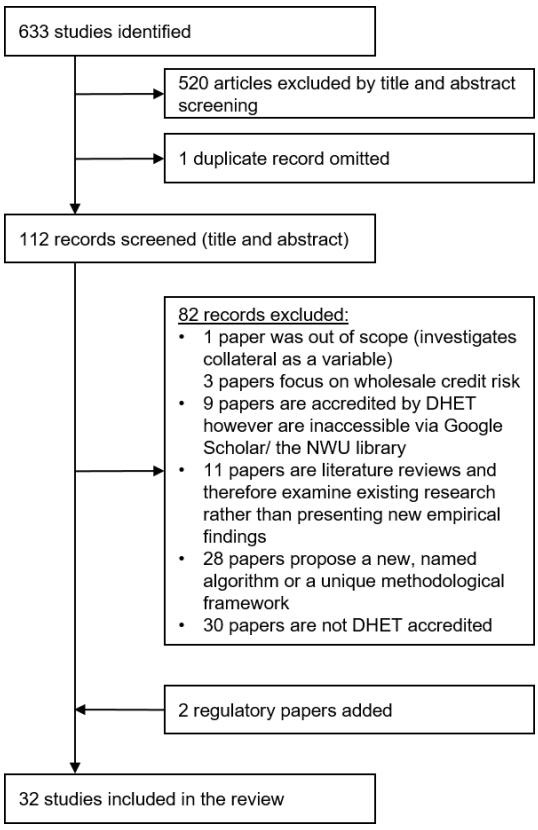

Figure 2: Flowchart for selecting studies
Source: Own compilation (2025)

Notes on coding and synthesis

The three stages of thematic synthesis, as described by Thomas and Harden (2008), were adopted for the remaining 32 studies: (1) free, line-by-line coding of the extracted summaries, (2) grouping codes into descriptive themes and (3) developing analytical themes that explain patterns across the literature. Coding was performed against the credit model development lifecycle phases to which each study primarily contributed (for example, data/feature selection, model building and model validation/deployment). Papers that contributed to more than one phase were tagged accordingly (refer **Appendix B**). From this, diverse approaches were identified to construct descriptive themes.

This approach aided in not “hard-coding” the analytical themes, which would cement rigidity; they were broad enough to accommodate the varying approaches that were studied in the literature.

The final themes and sub-themes (codes) identified are detailed in **Table 2**:

Table 2: Codebook themes

#	Theme	Code	Definition
1	Traditional model limitations	1.1 General limitations of traditional models	Identifies common limitations in traditional assessment scoring models.
2	Emerging market considerations	2.1 Application with alternative data	Studies that apply ML models to non-traditional data sources to score individuals, particularly those with thin credit files.
3	Machine Learning application in credit scoring	3.1: Comparative performance analysis	**1.1a Tree-Based Ensemble Methods (Bagging and Boosting):** Studies where the main comparison involves different types of tree-based ensembles (e.g. Random Forest, AdaBoost, Gradient Boosting models like XGBoost, LightGBM, CatBoost).
			1.1b Neural Networks and Deep Learning: Studies that specifically evaluate or compare the performance of Artificial Neural Networks (ANN), Multilayer Perceptrons (MLP) or more complex Deep Learning architectures (like CNNs or RNNs).
			1.1c Kernel-Based Methods (e.g. Support Vector Machines): Studies where Support Vector Machines (SVM) are a central part of the comparison.
			1.1d Broader Multi-Category Comparisons: Studies that conduct a wide-ranging "bake-off" across multiple algorithm families (e.g. comparing an ensemble, a neural network, an SVM and logistic regression all in one paper).
		3.2: Data handling and feature engineering techniques	Studies that focus on techniques used in conjunction with ML models, such as handling imbalanced data, feature selection or feature engineering.
4	Barriers to ML adoption	4.1 Regulatory resistance	Studies that detail obstacles to ML use in South Africa or comparable contexts (i.e. emerging markets).
5	Transparency & explainability	5.1: Model interpretability technique used	Studies that address how transparency is achieved in ML models.
		5.2: Accuracy and transparency trade-off	Captures whether and how the study addresses both accuracy, predictability and explainability in tandem, such as reporting on the predictive performance of transparent models, using interpretable algorithms without significant accuracy loss or discussing concessions or enhancements made to preserve both qualities.

Instances from the literature that spoke to each theme were extracted from the Data Extraction form (refer **Appendix B**) and included in the codebook; an overall summary of the paper (mainly extracted from the abstracts and results sections of the articles) was also extracted and included in the codebook (refer **Appendix C**).

The impetus for innovation: traditional model limitations (theme 1)

Traditional statistical approaches are widely used in banking; methodologies such as decision trees, linear discriminant analysis and logistic regression (LR) for credit scoring have been in use for over 30 years (Bücker et al., 2022; Du Toit et al., 2024; Laborda & Ryoo, 2021), and their appeal stems from their familiarity with regulators and auditors who understand these models due to their linear nature, thus making them highly interpretable (Bücker et al., 2022; Hlongwane et al., 2024). However, despite their established role, the limitations of these models have been documented extensively in academia, including their low classification accuracy due to the restrictive assumptions made by the models (Chopra & Bhilare, 2018; Cubiles-De-La-Vega et al., 2013; Liu et al., 2024). This rigidity makes the models less resilient when confronted with large, high-dimensional datasets that characterise modern banking as the model assumptions are often violated in the real world (Chopra & Bhilare, 2018; Hussin et al., 2022; Thuy et al., 2025).

Intrinsically, these models often exhibit a performance ceiling. Research shows that ML models consistently outperform traditional models and achieve higher predictive accuracy (Bono et al., 2021; Chopra & Bhilare, 2018; Hlongwane et al., 2024; Laborda & Ryoo, 2021), and the accuracy of traditional methods can decline when non-linear interactions are present in the data, thereby limiting their ability to differentiate between high and low-risk borrowers (Thuy et al., 2025); this is especially problematic when assessing applicants with no formal credit history (termed ‘thin-file’ customers), which is a common characteristic in emerging markets, including South Africa (Thuy et al., 2025).

Emerging market consideration: financial inclusion (theme 2)

To address the challenge in scoring thin-file customers, ML offers a solution: using alternative sources such as social media activity, e-commerce data and mobile device usage to determine the credit score (Djeundje et al., 2021; Li et al., 2024). Applying ML to these alternate sources allows banks the opportunity to gain a more nuanced view of an applicant’s creditworthiness by assessing qualitative factors like financial discipline and willingness to pay (Schindler et al., 2017; Thuy et al., 2025); thus, it is evident how ML impacts the financial inclusion mandate by including the unbanked (Tigges et al., 2024).

Enhancing predictive power: comparative performance analysis (theme 3)

Thus, the limitations of traditional models are the driving force for the extensive research into the application of ML in credit scoring, which has uncovered a world of powerful algorithms capable of delivering predictive accuracy that far exceeds traditional models (Chopra & Bhilare, 2018; Laborda & Ryoo, 2021):

Dominance of tree-based ensemble methods (theme 3.1a)

Ensemble methods leverage the idea that combining predictions from multiple models typically yields better results than a single model alone (Chopra & Bhilare, 2018; Kozak, 2019, p. 107; Tripathi et al., 2021). Decision trees are algorithms that can be used for both classification (the decision tree predicts a categorical label) and regression (the decision tree predicts a continuous value) tasks (Kozak, 2019, p. 1), and, across the literature, tree-based ensemble methods emerge as the most consistently high-performing class of algorithms in credit scoring (Ben-David & Frank, 2009). Studies such as those by Chopra and Bhilare (2018), Li and Chen (2020) and those referenced by Laborda and Ryoo (2021) repeatedly find that ensemble methods outperform individual classifiers in terms of predictive accuracy. Across the literature assessed, Random Forest is lauded for its strong performance where in comparative performance assessments, Random Forest returns the best performance across five metrics (accuracy, area under the curve (AUC), Kolmogorov–Smirnov statistic (KS), Brier score (BS) and model operating time) when compared to other ensemble methods (Li & Chen, 2020; Thuy et al., 2025).

Gradient Boosting Machines (GBMs): another ensemble ML approach, including modern implementations like XGBoost, LightGBM and CatBoost, are also recognised for their exceptional predictive power (Boughaci & Alkhawaldeh, 2020; Liu et al., 2024; Tran, 2021), where XGBoost and LightGBM came in as close contenders to Random Forests in the study by Li and Chen (2020). The computational efficiency of these modern GBMs, especially when dealing with large datasets, is also noted (Li & Chen, 2020; Liu et al., 2024) where XGBoost is often identified as the most accurate model, even in direct comparisons by using Vietnamese data, which holds emerging market similarities with South Africa (Sugianto et al., 2024) along with CatBoost, which scores high in recall and AUC (Liu et al., 2024; Thuy et al., 2025). Therefore, the consensus is that ensemble methods greatly improve single decision trees and traditional models by reducing variance and bias, thereby leading to more reliable identification of potential defaulters (Chopra & Bhilare, 2018; Hussin et al., 2022).

Neural networks and deep learning (theme 3.1b)

Neural networks (NNs) and deep learning (DL) architectures are another powerful frontier in credit scoring (Muslimin, 2022; Zhirov et al., 2021); these models excel at automatically capturing complex, non-linear relationships and can process diverse data types, including both structured and unstructured inputs (Bequé & Lessmann, 2017; Hussin et al., 2022), making them well-suited for modelling temporal patterns in borrower behaviour or incorporating alternative data sources (Thuy et al., 2025; Zhirov et al., 2021). Research by Zhirov et al. (2021) confirmed the effectiveness of NNs in credit scoring where the system demonstrated 95 per cent correctness in determining borrower creditworthiness.

However, NNs and DL lack transparency, which complicates validation and regulatory compliance (Cubiles-De-La-Vega et al., 2013; Du Toit et al., 2024). Another drawback is that they are often computationally intensive and sensitive to data imbalances, which may distort default predictions (Dessain et al., 2023; Thuy et al., 2025).

Kernel-based and other methods (theme 3.1c and 3.1d)

Support Vector Machines (SVMs) are frequently evaluated in credit scoring literature as they are considered competitive classifiers (Sugianto et al., 2024) with some studies finding their performance comparable to NNs (Hussin et al., 2022; Laborda & Ryoo, 2021; Thuy et al., 2025), yet, like NNs, they lack interpretability while comparative studies show that modern ensemble methods perform better in terms of accuracy (Chopra & Bhilare, 2018; Thuy et al., 2025). Other "bake-off" (comparative) studies that compare multiple algorithms, such as LR, NNs, SVMs and ensembles, reinforce the general conclusion that ensemble methods often provide the optimal balance between predictive power and stability (Hussin et al., 2022; Laborda & Ryoo, 2021; Li & Chen, 2020).

<u>Data handling and feature engineering (theme 3.2)</u>

The success of any model is dependent on the quality of the underlying data (Yi & Yuwen, 2021); literature highlights two critical data pre-processing steps: imbalanced data management and feature selection.

1. **The "imbalanced data" problem:** Credit datasets are known to be imbalanced in that there are far more non-defaulters than defaulters (Hussin et al., 2022; Thuy et al., 2025), which effectively biases a model towards the majority and, thereby, reduces its ability to identify actual risk. Techniques to mitigate against such biases are essential, and methods such as random oversampling and SMOTE (Synthetic Minority Oversampling Technique) are widely used to artificially create new minority class instances, thus balancing the dataset and improving classifier importance (He et al., 2022; Hussin et al., 2022).

2. **Feature selection and engineering:** To avoid overfitting, and to reduce model complexity, it is imperative to select the most informative predictors from large datasets (Koutanaei et al., 2015; Laborda & Ryoo, 2021). An array of techniques exists that range from simple statistical (and transparent) filter methods like the chi-squared test and correlation coefficients to more sophisticated ML-based approaches (Hussin et al., 2022; Liu & Schumann, 2005; Trivedi, 2020; Yi & Yuwen, 2021). Prominently featured are wrapper methods, which use a specific ML algorithm to iteratively evaluate subsets of features, with forward and backward stepwise selection being noted for their effectiveness (Laborda & Ryoo, 2021; Liu & Schumann, 2005). More advanced approaches include embedded methods where feature selection is an intrinsic part of the model's training process (Talaat et al., 2023) and even uses complex algorithms like gradient boosting decision trees explicitly for feature engineering before selecting the features (Laborda & Ryoo, 2021). It is evident in the literature that these feature selection techniques are value-adding in that they directly enhance the predictive accuracy of credit scoring models (Laborda & Ryoo, 2021; Liu & Schumann, 2005).

Bridging the trust gap: interpreting ML (themes 5, 6 and 7)

As such, it is clear in the literature that ML offers superior accuracy over traditional models across the board; however, central to ML adoption in banking is addressing their complexity and transparency, which is a major barrier to adoption in South Africa (High-Level Expert Group on Artificial Intelligence, 2019; Schindler et al., 2017). The South African Reserve Bank's (SARB) definition of model risk – as it pertains to regulatory models – takes two forms: the first is the incorrect valuation methodology and the second is unobservable (and possibly incorrect) calibration parameters (Du Toit et al., 2024; South African Reserve Bank, 2015). Therefore, model parameters must be justifiable, and to overcome the black-box barrier in adopting ML, is Explainable AI (XAI), which provides tools to interpret complex models without compromising on predictive power (Du Toit et al., 2024; Thuy et al., 2025).

Studies show several methods for achieving model transparency (e.g. Local Interpretable Model-agnostic Explanation (LIME), iBreakdown approach) with SHAP (Shapley Additive exPlanations or Shapley values) emerging as the most widespread method for explaining the output of ML models (Dessain et al., 2023; Du Toit et al., 2024; Hlongwane et al., 2024; Talaat et al., 2023). The SHAP model provides insights by quantifying the contribution of each feature to a specific prediction, which allows for both global and local interpretability: understanding which features are most important for the model's prediction overall and explaining why an individual applicant received their specific scorecard outcome (Liu et al., 2024; Talaat et al., 2023; Thuy et al., 2025). The advantage of the SHAP model lies in its ability to generate explanations for even the most complex models, thereby

translating a black-box decision into an evidence-based reason code that aligns with industry practice (Bücker et al., 2022; Hlongwane et al., 2024).

Transparency or accuracy? Or both? (themes 6 and 7)

Hlongwane et al. (2024) show that transparency and accuracy are not necessarily mutually exclusive; by building-in XAI, it is possible to leverage the high accuracy of ML models like XGBoost and random forest, and that credit scores derived from SHAP can align closely with those from traditional LR models. Studies have also begun quantifying the "cost of explainability" where the financial return of the best performing black-box ML model and the best performing inherently explainable ML model are compared, and Dessain et al. (2023) find that there is, in fact, a cost linked to explainability; however, the study also found that traditional models performed poorly in both statistical and financial terms.

Discussion

This review set out to examine how ML techniques are being applied to credit scoring, how they compare to traditional statistical models and what barriers remain to adoption in the South African context. The literature shows that traditional statistical scorecards (logistic regression and linear discriminant analysis) remain widely used because they are familiar to auditors and regulators and are straightforward to interpret (Bücker et al., 2022). However, these models rest on linearity and other restrictive assumptions that increasingly limit their discriminatory power as banking datasets become larger, higher-dimensional and more non-linear in structure. In emerging market contexts with many thin file borrowers, these limitations are especially acute because simple models struggle to extract useful signals from sparse or non-traditional data (Li et al., 2024).

Across the studies coded, tree-based ensemble methods were the most consistently high-performing family of algorithms. Random Forest and gradient boosted implementations (XGBoost, LightGBM, CatBoost) and repeatedly delivered superior discrimination (AUC/Gini/KS) and strong performance across datasets and metrics in comparative studies (Zou et al., 2025). The ensembles' superior performance is explained by their architecture: they combine multiple base learners to reduce variance (bagging) and bias (boosting), thereby producing more stable and accurate predictors than single classifiers (Lessmann et al., 2015). This consistency across the reviewed papers makes tree-based ensembles a pragmatic first choice when rank ordering bad from good customers (discrimination) is the primary objective in credit scoring.

Neural networks and other DL models can also achieve high predictive performance and have particular strengths in modelling complex, non-linear relationships and unstructured or temporal inputs; however, they raise two concerns: transparency (difficulty of explanation and validation) and sensitivity to class imbalance and limited data, which complicates reliable deployment in regulated settings (Zou et al., 2025). Support Vector Machines are competitive in many studies but likewise suffer from limited interpretability in production environments (Bono et al., 2021). In short, while many algorithms can reach strong statistical performance on given samples, their operational suitability depends on governance, explainability and data readiness as much as on accuracy.

Explainable AI (XAI) techniques, especially SHAP, are now widely reported as the practical route to reconciling ensembles' predictive power with auditability requirements. The SHAP model provides both global and local explanations by attributing marginal contributions of features to predictions, enabling production-grade "reason codes" that are intelligible to stakeholders and auditors (Du Toit et al., 2024). Studies in the codebook show SHAP explanations for tree ensembles produce score level reason codes that align closely with traditional scorecard outputs while preserving the

ensemble's superior discrimination, showing us that accuracy and transparency need not be mutually exclusive (Hlongwane et al., 2024; Liu et al., 2024). Nonetheless, recent work has begun to quantify the possible "cost of explainability", and that cost appears context-dependent and is generally modest in published examples. Banks must, therefore, weigh this potential cost against the regulatory and operational benefits of explainability within their own portfolios.

The literature emphasises that much of ML's practical value lies upstream of model fitting: careful data preparation, principled feature selection and imbalance handling materially determine downstream success. Credit datasets are typically highly imbalanced, and authors routinely recommend methods such as random oversampling and SMOTE to rebalance classes and improve minority class detection (Hussin et al., 2022). Feature selection strategies – from transparent filters (chi-square, correlations) to wrapper and embedded methods – consistently improve predictive accuracy and reduce overfitting; many studies show that feature engineering and selection are as important, if not more so, than the chosen classifier (Laborda & Ryoo, 2021). Because these pre-processing steps can themselves be algorithmic, governance teams should document and justify them with the same rigour as model selection and calibration.

Alternative data and financial inclusion appear repeatedly as both an opportunity and a risk. Studies demonstrate that non-traditional signals (mobile phone metadata, payments and digital footprints) can let ML score thin-file applicants and extend credit to previously unscorable populations, enhancing financial inclusion when implemented responsibly (Li et al., 2024; Óskarsdóttir et al., 2019). At the same time, the literature warns that alternative data can encode historical and structural biases; therefore, fairness testing, transparency, and explicit consent/privacy considerations must accompany any alternative data pilot (Tigges et al., 2024). This is especially important for South Africa, given historical inequalities.

Regulatory and governance readiness is a recurrent barrier to ML adoption in banking. Practical model risk frameworks for ML have been proposed that integrate Shapley-based interpretability tests and model validation procedures, but published guidance specific to South African supervisory practice is limited in the corpus that was reviewed in this study; consequently, South African banks should build hybrid governance that maps international best practice to local regulatory expectations and engage supervisors early in pilot programmes (Du Toit et al., 2024). In other words, a defensible adoption roadmap is staged and cautious: use ML for augmentation (feature engineering, segmentation) while maintaining interpretable decision layers or SHAP-based reason codes for the production decision, accompanied by documented validation, monitoring and recourse mechanisms (Bücker et al., 2022).

Recommendations

Therefore, in summary, I present practical recommendations for South African banks:

- **Adopt a staged, hybrid approach:** an implementation strategy that separates ML-driven data work from the final decision layer; in other words, banks should explore the use of ML for data preparation (feature engineering, signal discovery and segment identification) and retain an interpretable decision model (or an auditable explanation layer) for production credit decisions so that ML's data benefits are captured while preserving transparent, regulator-friendly outputs (Laborda & Ryoo, 2021; Liu & Schumann, 2005).
- **Prioritise tree-based ensemble models for score derivation** where discrimination is the objective, and embed explainability tools from project inception. Random Forest and modern gradient boosting consistently deliver top discrimination and stability across studies, making them sensible first choices for production modelling (Li & Chen, 2020). However, they require integrated XAI (e.g. SHAP) in the development pipeline so that every modelled decision has an auditable local and global explanation (Khan et al., 2025).
- **Explore alternative data points** and treat alternative data as a strategic inclusion pilot, not a wholesale replacement. Alternative data allows for expansion into underserved markets, thereby impacting financial inclusion (Schindler et al., 2017; Thuy et al., 2025); therefore, run controlled pilots (with privacy and fairness safeguards) to evaluate whether adding mobile, utility or behavioural signals improves both approval rates and realised defaults for thin-file applicants, and always accompany these pilots with bias and privacy impact assessments (Li et al., 2024; Tigges et al., 2024).
- **Engage supervisors early** and document model risk extensions. Because South African supervisory guidance is less prescriptive for business scorecards than for high-impact regulatory models, banks should map international best practice to local requirements, document Shapley-based validation tests and seek early supervisory dialogue for novel ML use cases to reduce regulatory friction (Schindler et al., 2017).
- **Testing reports should include both technical and economic outcomes:** any ML pilot (challenger models) should track business metrics (approval volumes, take up, default rates and other portfolio-specific metrics) in addition to statistical model monitoring metrics (AUC, calibration etc.) to allow for investment decisions that reflect real economic impact rather than only classifier gains (Dessain et al., 2023).

Appendix D provides a visual dashboard that summarises the practical implications of this research for the South African banking sector.

Conclusion

This review has shown a clear case for the use of ML in South Africa. In a fast-paced, ever-changing global environment, it is the bank's business to explore ways to remain competitive while retaining regulatory compliance. Traditional models, while interpretable, are increasingly outmatched by ML algorithms – tree-based ensembles in particular. Further, the long-standing barrier to adoption (model transparency) is being effectively dismantled by a new generation of XAI techniques, with Shapley values providing a regulator-friendly method for explaining ML decisions. Thus, for South African banks, the path is clear: adopting a hybrid programme that exploits ML where it adds the most value by leveraging the non-prescriptive nature of the SARB towards business models, implementing tree-based ensemble models and cautiously exploring alternative data, banks can greatly enhance their risk assessment capabilities. If implemented with the recommended tests, validations and regulatory engagement, ML becomes a present-day, implementable opportunity to improve lending inclusion and risk management.

References

Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intelligent Systems in Accounting, Finance and Management*, *18*(2-3), 59-88. https://doi.org/10.1002/isaf.325

Addo, P. M., Guegan, D., & Hassani, B. (2018). Credit Risk Analysis Using Machine and Deep Learning Models. *Risks*, *6*(2), 38. https://www.mdpi.com/2227-9091/6/2/38

Agung Dharmawan, B. (2020). The effect of credit risk and capital adequacy on financial distress in rural banks. *Accounting*. http://dx.doi.org/10.5267/j.ac.2020.7.023

Allen, L., DeLong, G., & Saunders, A. (2004). Issues in the credit risk modeling of retail markets. *Journal of Banking and Finance*, *28*(4), 727-752. https://doi.org/10.1016/j.jbankfin.2003.10.004

Ashofteh, A., & Bravo, J. M. (2021). A conservative approach for online credit scoring. *Expert systems with applications*, *176*. https://doi.org/10.1016/j.eswa.2021.114835

Baesens, B., Roesch, D., & Scheule, H. (2016). *Credit risk analytics: measurement techniques, applications, and examples in SAS*. Wiley. https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=1357222

Basel Committee on Banking Supervision. (2004). *International convergence of capital measurement and capital standards: a revised framework*. B. f. I. Settlements. https://www.bis.org/publ/bcbs107.pdf

Basel Committee on Banking Supervision. (2013). *Principles for effective risk data aggregation and risk reporting'*. B. f. I. Settlements. https://www.bis.org/publ/bcbs239.pdf

Ben-David, A., & Frank, E. (2009). Accuracy of machine learning models versus "hand crafted" expert systems - A credit scoring case study. *Expert systems with applications*, *36*(3), 5264-5271. https://doi.org/10.1016/j.eswa.2008.06.071

Bequé, A., & Lessmann, S. (2017). Extreme learning machines for credit scoring: An empirical evaluation. *Expert systems with applications*, *86*, 42-53. https://doi.org/10.1016/j.eswa.2017.05.050

Bijak, K., & Thomas, L. C. (2012). Does segmentation always improve model performance in credit scoring? *Expert systems with applications*, *39*(3), 2433-2442. https://doi.org/10.1016/j.eswa.2011.08.093

Blöchlinger, A., & Leippold, M. (2006). Economic benefit of powerful credit scoring. *Journal of Banking and Finance*, *30*(3), 851-873. https://doi.org/10.1016/j.jbankfin.2005.07.014

Bono, T., Croxson, K., & Giles, A. (2021). Algorithmic fairness in credit scoring. *Oxford Review of Economic Policy*, *37*(3), 585-617. https://doi.org/10.1093/oxrep/grab020

Boughaci, D., & Alkhawaldeh, A. A. K. (2020). Appropriate machine learning techniques for credit scoring and bankruptcy prediction in banking and finance: a comparative study. *Risk and*

Decision Analysis 8(1-2), 15-24. https://north-on-worldcat-org.nwulib.idm.oclc.org/atoztitles/link/?sid=Elsevier:Scopus

Boughaci, D., Alkhawaldeh, A. A. K., Jaber, J. J., & Hamadneh, N. (2020). Classification with segmentation for credit scoring and bankruptcy prediction. *Empirical Economics : Journal of the Institute for Advanced Studies, Vienna, Austria*, *61*(3), 1281-1309. https://doi.org/10.1007/s00181-020-01901-8

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

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

Bücker, M., Szepannek, G., Gosiewska, A., & Biecek, P. (2022). Transparency, auditability, and explainability of machine learning models in credit scoring. *Journal of the Operational Research Society*, *73*(1), 70-90. https://doi.org/10.1080/01605682.2021.1922098

Çallı, B. A., & Coşkun, E. (2021). A Longitudinal Systematic Review of Credit Risk Assessment and Credit Default Predictors. *SAGE Open*, *11*(4). https://doi.org/10.1177/21582440211061333

Chen, N., Ribeiro, B., & Chen, A. (2016). Financial credit risk assessment: a recent review. *Artificial Intelligence Review : An International Science and Engineering Journal*, *45*(1), 1-23. https://doi.org/10.1007/s10462-015-9434-x

Chopra, A., & Bhilare, P. (2018). Application of ensemble models in credit scoring models. *Business Perspectives and Research*, *6*(2), 129-141. https://doi.org/10.1177/2278533718765531

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

Cubiles-De-La-Vega, M.-D., Blanco-Oliver, A., Pino-Mejías, R., & Lara-Rubio, J. (2013). Improving the management of microfinance institutions by using credit scoring models based on Statistical Learning techniques. *Expert systems with applications*, *40*(17), 6910-6917. https://doi.org/10.1016/j.eswa.2013.06.031

Dastile, X., Celik, T., & Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing Journal*, *91*. https://doi.org/10.1016/j.asoc.2020.106263

Dessain, J., Bentaleb, N., & Vinas, F. (2023). Cost of Explainability in AI: An Example with Credit Scoring Models. *Communications in Computer and Information Science*, *1901 CCIS*, 498-516. https://north-on-worldcat-org.nwulib.idm.oclc.org/atoztitles/link/?sid=Elsevier:Scopus

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

du Toit, H. A., Schutte, W. D., & Raubenheimer, H. (2024). Integrating traditional and non-traditional model risk frameworks in credit scoring. *South African Journal of Economic and Management Sciences*, *27*(1). https://doi.org/10.4102/sajems.v27i1.5786

Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European journal of operational research*, *297*(3), 1178-1192. https://doi.org/10.1016/j.ejor.2021.06.053

Edelberg, W. (2006). Risk-based pricing of interest rates for consumer loans. *Journal of Monetary Economics*, *53*(8), 2283-2298. https://doi.org/10.1016/j.jmoneco.2005.09.001

Fang, F., & Chen, Y. (2019). A new approach for credit scoring by directly maximizing the Kolmogorov-Smirnov statistic. *Computational Statistics and Data Analysis*, *133*, 180-194. https://doi.org/10.1016/j.csda.2018.10.004

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics and Data Analysis*, *38*(4), 367-378. https://doi.org/10.1016/S0167-9473(01)00065-2

Gomber, P., Koch, J.-A., & Siering, M. (2017). Digital Finance and FinTech: current research and future research directions. *Journal of Business Economics : Zeitschrift für Betriebswirtschaft*, *87*(5), 537-580. https://doi.org/10.1007/s11573-017-0852-x

Gross, M. (2020). *Expected Credit Loss Modeling from a Top-Down Stress Testing Perspective*. International Monetary Fund. http://elibrary.imf.org/view/IMF001/29147-9781513549088/29147-9781513549088/29147-9781513549088.xml

Guégan, D., & Hassani, B. (2018). Regulatory learning: How to supervise machine learning models? An application to credit scoring. *The Journal of Finance and Data Science*, *4*(3), 157-171. https://doi.org/10.1016/j.jfds.2018.04.001

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

Harris, T. (2013). Quantitative credit risk assessment using support vector machines: Broad versus Narrow default definitions. *Expert systems with applications*, *40*(11), 4404-4413. https://doi.org/10.1016/j.eswa.2013.01.044

He, F., Zhang, W., & Yan, Z. (2022). A novel multi-stage ensemble model for credit scoring based on synthetic sampling and feature transformation. *Journal of Intelligent & Fuzzy Systems*, *42*(3), 2127-2142. https://doi.org/10.3233/JIFS-211467

High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI*. E. Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Hlongwane, R., Ramabao, K., & Mongwe, W. (2024). 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

Hlophe, N. (2024). Artificial intelligence in the financial sector and emerging consumer protection concerns. 16. Retrieved 20 February 2025, from

http://fsca.co.za/Documents/Presentation%20on%20AI%20in%20Finance%20and%20Consumer%20Protection%20FSCA%20Conference%20Day%201.pdf

Houda, Z. A. E., Brik, B., & Khoukhi, L. (2022). “Why Should I Trust Your IDS?”: An Explainable Deep Learning Framework for Intrusion Detection Systems in Internet of Things Networks. *IEEE Open Journal of the Communications Society*, *3*. https://doi.org/10.1109/OJCOMS.2022.3188750

Huang, J., Wang, H., & Wang, W. (2018). Variable selection in classification model via quadratic programming. *Communications in Statistics - Simulation and Computation*, *47*(7), 1922-1939. https://doi.org/10.1080/03610918.2017.1332211

Hussin, A. K., Ahmed, A., & Bee, M. (2022). Machine learning models and data-balancing techniques for credit scoring: what is the best combination? *Risks*, *10*(9), 169. https://doi.org/10.3390/risks10090169

International Accounting Standards Board. (2014). *IFRS 9 Financial Instruments*. I. Foundation. https://www.ifrs.org/content/dam/ifrs/publications/pdf-standards/english/2021/issued/part-a/ifrs-9-financial-instruments.pdf

Khan, F. S., Mazhar, S. S., Mazhar, K., A. AlSaleh, D., & Mazhar, A. (2025). Model-agnostic explainable artificial intelligence methods in finance: a systematic review, recent developments, limitations, challenges and future directions. *Artificial Intelligence Review : An International Science and Engineering Journal*, *58*(8). https://doi.org/10.1007/s10462-025-11215-9

Kočenda, E., & Vojtek, M. (2011). Default Predictors in Retail Credit Scoring: Evidence from Czech Banking Data. *Emerging Markets Finance and Trade*, *47*(6), 80-98. https://doi.org/10.2753/REE1540-496X470605

Koutanaei, F. N., Sajedi, H., & Khanbabaei, M. (2015). A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring. *Journal of Retailing and Consumer Services*, *27*, 11-23. https://doi.org/10.1016/j.jretconser.2015.07.003

Kozak, J. (2019). *Decision tree and ensemble learning based on ant colony optimization*. Springer. https://doi.org/10.1007/978-3-319-93752-6

Kraus, A. (2014). *Recent Methods from Statistics and Machine Learning for Credit Scoring*. Cuvillier Verlag. https://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=5018539

Kumar, A., Sharma, S., & Mahdavi, M. (2021). Machine Learning (ML) Technologies for Digital Credit Scoring in Rural Finance: A Literature Review. *Risks*, *9*(11), 192. https://doi.org/10.3390/risks9110192

Laborda, J., & Ryoo, S. (2021). Feature selection in a credit scoring model. *Mathematics*, *9*(7), 746. https://doi.org/10.3390/math9070746

Lee, T.-S., Chiu, C.-C., Lu, C.-J., & Chen, I. F. (2002). Credit scoring using the hybrid neural discriminant technique. *Expert systems with applications*, *23*(3), 245-254. https://doi.org/10.1016/S0957-4174(02)00044-1

Leo, M., Sharma, S., & Maddulety, K. (2019). Machine Learning in Banking Risk Management: A Literature Review. *Risks*, *1*, 29. http://dx.doi.org/10.3390/risks7010029

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

Li, C., Wang, H., Jiang, S., & Gu, B. (2024). The Effect of AI-Enabled Credit Scoring on Financial Inclusion: Evidence from an Underserved Population of over One Million. *MIS Quarterly*, *48*(4), 1803-1834. https://doi.org/10.25300/MISQ/2024/18340

Li, Y., Bellotti, T., & Adams, N. (2019). Issues using logistic regression with class imbalance, with a case study from credit risk modelling. *Foundations of Data Science*, *1*(4), 389-417. https://doi.org/10.3934/fods.2019016

Li, Y., & Chen, W. (2020). A comparative performance assessment of ensemble learning for credit scoring. *Mathematics*, *8*(10), 1756. https://doi.org/10.3390/math8101756

Liu, Y., Huang, F., Ma, L., Zeng, Q., & Shi, J. (2024). Credit scoring prediction leveraging interpretable ensemble learning. *Journal of Forecasting*, *43*(2), 286-308. https://doi.org/10.1002/for.3033

Liu, Y., & Schumann, M. (2005). Data mining feature selection for credit scoring models. *Journal of the Operational Research Society*, *56*(9), 1099-1108. https://doi.org/10.1057/palgrave.jors.2601976

Lundberg, S. M., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions* https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Madeira, C. (2023). Adverse selection, loan access and default behavior in the Chilean consumer debt market. *Financial Innovation*, *9*(1). https://doi.org/10.1186/s40854-023-00458-6

Makuch, W. M. (2001). *The Basics of a Better Application Score* (E. Mays, Ed.). Glenlake Publishing Company.

Martens, D., Baesens, B., Van Gestel, T., & Vanthienen, J. (2007). Comprehensible credit scoring models using rule extraction from support vector machines. *European journal of operational research*, *183*(3), 1466-1476. https://doi.org/10.1016/j.ejor.2006.04.051

Martin, N. (2013). Assessing scorecard performance: A literature review and classification. *Expert systems with applications*, *40*(16), 6340-6350. https://doi.org/10.1016/j.eswa.2013.05.060

Mestiri, S. (2024). Credit scoring using machine learning and deep Learning-Based models. *Data Science in Finance and Economics*, *4*(2), 236-248. https://doi.org/10.3934/DSFE.2024009

Mohua, R. (2006). A Review of Bank Lending to Priority and Retail Sectors. *Economic and Political Weekly*, *41*(11), 1035-1040.

Montevechi, A. A., Miranda, R. d. C., Medeiros, A. L., & Montevechi, J. A. B. (2024). Advancing credit risk modelling with Machine Learning: A comprehensive review of the state-of-the-art. *Engineering Applications of Artificial Intelligence*, *137*. https://doi.org/10.1016/j.engappai.2024.109082

Muneri, K. P., & Kuhn, K. (2023). Note on household sector debt and

debt-service cost statistics. *Full Quarterly Bulletin – No. 309 – September 2023*, *309*(September 2023), 94-97. https://www.resbank.co.za/en/home/publications/publication-detail-pages/quarterly-bulletins/quarterly-bulletin-publications/2023/FullQuarterlyBulletinNo309December2023

Muslimin, F. (2022). Implementation of machine learning technology for consumer credit scoring in banking industry: study case of PT Bank BNI Syariah. *Journal of Theoretical and Applied Information Technology*, *100*(7), 1628-1642. https://north-on-worldcat-org.nwulib.idm.oclc.org/atoztitles/link/?sid=Elsevier:Scopus

Mutsonziwa, & Fanta. (2021). Small Business Performance: Is It Access to Formal or Informal Credit that Matters? *Journal of African Business*, *22*(4), 550-563. https://doi.org/10.1080/15228916.2020.1826854

Neal, R. S. (1996). Credit derivatives: new financial instruments for controlling credit risk. *Economic Review-Federal Reserve Bank of Kansas City*, *81*, 15-28.

Okeleke, P. A., Ajiga, D., Folorunsho, S. O., & Ezeigweneme, C. (2024). Predictive analytics for market trends using AI: a study in consumer behavior. *International Journal of Engineering Research Updates*. https://doi.org/https://doi.org/10.53430/ijeru.2024.7.1.0032

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

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Journal of Clinical Epidemiology*, *134*, 178-189. https://doi.org/10.1016/j.jclinepi.2021.03.001

Parums, D. V. (2021). Editorial: Review Articles, Systematic Reviews, Meta-Analysis, and the Updated Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 Guidelines. *Medical science monitor : international medical journal of experimental and clinical research*, *27*, e934475. https://doi.org/10.12659/MSM.934475

Poon, M. (2007). Scorecards as devices for consumer credit: the case of Fair, Isaac & Company Incorporated. *The Sociological Review*, *55*, 284-306. https://doi.org/10.1111/j.1467-954X.2007.00740.x

Ralf, W., Yvonne, T., Christian, S., Mareike, D., Dominique, S., & Tim, G. (2024). A global comparison of credit bureaus based on data utilization in credit scoring. *The International Journal of Banking and Finance*, *20*(1). http://dx.doi.org/10.32890/ijbf2025.20.2.2

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

Rosenberg, E., & Gleit, A. (1994). Quantitative Methods in Credit Management: A Survey. *Operations Research*, *42*(4), 589-613.

Schindler, J., de Souza Moraes, A., Li, S.-P., Bruno, G., Marcucci, J., Moscatelli, M., Yamaoka, H., Elizondo, A., ter Braak, M., Pang, D., Xuchun, L., Mueller, H., Seta, M., Thurner, S., Caviness, B., Landesman, J., Jacobs, E., Bandeh-Ahmadi, A., Leonova, I., . . . Chiong, R. (2017). *Artificial intelligence and machine learning in financial services: market developments and financial stability implications*. F. S. Board. https://www.fsb.org/uploads/P011117.pdf

Siddiqi, N. (2017). *Intelligent credit scoring : building and implementing better credit risk scorecards* (Second edition ed.). Wiley.

Simatele, M., & Maciko, L. (2022). Financial Inclusion in Rural South Africa: A Qualitative Approach. *Journal of Risk and Financial Management*, *15*(9), 376. https://doi.org/10.3390/jrfm15090376

National Credit Act 34 of 2005, (2005). https://www.justice.gov.za/mc/vnbp/act2005-034.pdf

South African Reserve Bank. (2015). *Amendments to the Regulations relating to Banks, and matters related thereto*. https://www.resbank.co.za/content/dam/sarb/publications/prudential-authority/pa-deposit-takers/banks-directives/2015/6664/D4-of-2015.pdf

Stats SA. (2025). *Survey of Employers and the Self-employed*. D. o. S. S. Africa. https://www.statssa.gov.za/publications/P0276/P02762023.pdf

Sugianto, Widyasari Y.D.L., & K.D.K., W. (2024). Modeling and Application of Credit Scoring Based on A Multi-Objective Approach to Debtor Data in PT. Bank Riau Kepri. *International Journal on Informatics Visualization*, *8*(1), 220-230. https://north-on-worldcat-org.nwulib.idm.oclc.org/atoztitles/link/?sid=Elsevier:Scopus

Talaat, F. M., Aljadani, A., Badawy, M., & Elhosseini, M. (2023). Toward interpretable credit scoring: integrating explainable artificial intelligence with deep learning for credit card default prediction. *Neural Computing and Applications*, *36*(9), 4847-4865. https://doi.org/10.1007/s00521-023-09232-2

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology*, *8*(1), 1-10. https://doi.org/10.1186/1471-2288-8-45

Thomas, L., Crook, J., & Edelman, D. (2017). *Credit scoring and its applications*. SIAM.

Thuy, N. T. H., Ha, N. T. V., Trung, N. N., Binh, V. T. T., Hang, N. T., & Binh, V. T. (2025). Comparing the effectiveness of machine learning and deep learning models in student credit scoring: a case study in Vietnam. *Risks*, *13*(5), 99. https://doi.org/10.3390/risks13050099

Tigges, M., Mestwerdt, S., Tschirner, S., & Mauer, R. (2024). Who gets the money? A qualitative analysis of fintech lending and credit scoring through the adoption of AI and alternative data. *Technological Forecasting & Social Change*, *205*. https://doi.org/10.1016/j.techfore.2024.123491

Tran, K. Q. (2021). Machine learning-based empirical investigation for credit scoring in Vietnam's banking. *Lecture Notes in Computer Science Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*, *12799 LNAI*, 564-574. https://north-on-worldcat-org.nwulib.idm.oclc.org/atoztitles/link/?sid=Elsevier:Scopus

Tripathi, D., Edla, D. R., Cheruku, R., & Kuppili, V. (2019). A novel hybrid credit scoring model based on ensemble feature selection and multilayer ensemble classification. *Computational Intelligence*, *35*(2), 371-394. https://doi.org/10.1111/coin.12200

Tripathi, D., Edla, D. R., Kuppili, V., Bablani, A., & Dharavath, R. (2018). Credit Scoring Model based on Weighted Voting and Cluster based Feature Selection. *Procedia Computer Science*, *132*, 22-31. https://doi.org/10.1016/j.procs.2018.05.055

Tripathi, D., Shukla, A. K., Reddy, B. R., Bopche, G. S., & Chandramohan, D. (2021). Credit scoring models using ensemble learning and classification approaches: a comprehensive survey. *Wireless Personal Communications: An International Journal*, *123*(1), 785-812. https://doi.org/10.1007/s11277-021-09158-9

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

Wang, T., & Lin, Q. (2021). Hybrid predictive models: When an interpretable model collaborates with a black-box model. *Journal of Machine Learning Research*, *22*(137), 1-38.

Wenbing, C., Yinglai, L., Yiyong, X., Xinglong, Y., Xingxing, X., Siyue, Z., & Shenghan, Z. (2019). A Machine-Learning-Based Prediction Method for Hypertension Outcomes Based on Medical Data. *Diagnostics*, *9*(4), 178. http://dx.doi.org/10.3390/diagnostics9040178

West, D. (2000). Neural network credit scoring models. *Computers and Operations Research*, *27*(11), 1131-1152. https://doi.org/10.1016/S0305-0548(99)00149-5

Whittaker, J., Somers, M., & Whitehead, C. (2007). A Dynamic Scorecard for Monitoring Baseline Performance with Application to Tracking a Mortgage Portfolio. *The Journal of the Operational Research Society*, *58*(7), 911-921.

Xia, Y., Liu, C., Li, 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

Yi, W., & Yuwen, P. (2021). Application analysis of credit scoring of financial institutions based on machine learning model. *Complexity*, *2021*. http://dx.doi.org/https://doi.org/10.1155/2021/9222617

Zhirov, V. K., Staroverova, N. A., Shustrova, M. L., & Tomilova, M. N. (2021). Neural network as a tool to solve the problem of credit scoring. *Journal of Physics: Conference Series*, *2032*(1). https://doi.org/10.1088/1742-6596/2032/1/012120

Zou, Y., Xia, M., & Lan, X. (2025). Interpretable credit scoring based on an additive extreme gradient boosting. *Chaos, Solitons and Fractals: the interdisciplinary journal of Nonlinear Science, and Nonequilibrium and Complex Phenomena*, *194*. https://doi.org/10.1016/j.chaos.2025.116216

REFLECTION

I compare this experience to sewing in that on paper the process requires you to measure twice and cut once, however, in practice, you will still find yourself ripping your stitches no matter how careful you were in preparing. And in the end, when you see how your vision has come to life, it's all worth it.

Planning, conducting, and writing this study has been both intellectually challenging and deeply rewarding. My professional background in credit risk initially led me to approach the topic purely from a technical perspective however, as the research progressed, I realised that the heart of the issue extends far beyond technical performance- it considers transparency, documentation, and regulatory trust. This shift in perspective became the defining insight of my project.

The study set out to examine how ML techniques are being applied to credit scoring, how they compare to traditional models, and what barriers remain to adoption within the South African banking context. By conducting a PRISMA-aligned systematic literature review, the study consolidated global findings and adapted them to the unique regulatory and socio-economic realities of South Africa. The research showed that ML models demonstrate clear superiority in predictive accuracy, however their adoption needs to be explainable and a hybrid programme is the best approach to ease stakeholders into the use of these models. The main contribution of the study lies in its contextualisation of global insights into a practical, locally relevant framework and the resulting recommendations provide a summary of how institutions can begin adopting ML responsibly. **Appendix D** provides a visual dashboard summarising the practical implications of this research for the South African banking sector. It illustrates how each major theme identified in the literature translates into actionable steps that banks, auditors, and regulators can implement to responsibly adopt machine learning in credit scoring.

Methodologically, the systematic review approach proved effective for synthesising fragmented research, but I acknowledge some limitations -the reliance on Scopus as the main database excluded potentially useful grey literature and practitioner papers. If I were to redo this study, I would broaden the search to include these papers to better capture local experimentation. A benchmarking exercise amongst South African banks would also contribute immensely to this body of work.

Personally, to do my Masters has been more than a dream come true as a woman that started her career unable to afford to complete her undergrad. This research journey has had a profound

personal impact. It not only expanded my technical understanding of ML models but also deepened my appreciation for the delicate balance between innovation and accountability in financial institutions. It also allowed for more detailed discussions with my colleagues in the model risk audit team who challenged the interpretations I made and provided practical insights on the operational realities of model governance. Consulting across my network enriched the study with every conversation I had as it allowed me to replay academic findings against what is currently in practice in the real world. In addition, the Model Development team hosts regular knowledge-sharing sessions on how they're exploring ML applications and interpretability techniques -this also allowed me to make the study relevant to the organisation. Interestingly, and unintentionally, some of the papers in the sample were written by our model experts in the organisation however, I could not engage them directly as I was leading an audit on them.

I came to see ML as a governance challenge as much as a technological one. I learned that the most advanced models hold little value without the governance structures that allow them to be used responsibly. The supposed benefits need to be quantifiable and not only in model performance terms- the financial benefits need to be worthwhile.

APPENDICES

Appendix A: Tracking Sheet

This tracking sheet is the core dataset used to populate the PRISMA flow diagram, particularly for the Identification, Screening, Eligibility, and Inclusion stages. This sheet includes all included and excluded papers for the literature review. Below is a snapshot of the tracking sheet, for full version please double-click on the icon to open in MS Excel:

Appendix A - Tracking sheet.xlsx

Appendix A: Tracking sheet

This tracking sheet is the core dataset used to populate the PRISMA flow diagram, particularly for the Identification, Screening, Eligibility, and Inclusion stages. This sheet includes all included and excluded papers for the literature review

#	Author(s)	Title	Year	Journal Title	Cited by	Abstract	Document Type	Publication Stage	Open Access	Source (Database)	Duplicate	Meets Inclusion Criteria?	Reason for Exclusion (if any)	Included in Final Review?	DHET Accredited?
1	Dastile X.; Ce	Making Deep	2021	IEEE Access	50	Credit scoring has become an important risk i	Article	Final	All Open Access; Go	Scopus	No	No	This paper focusses on propos	No	Accredited
2	Dumitrescu E.	Machine lear	2022	European Journal of	234	In the context of credit scoring, ensemble me	Article	Final	All Open Access; Bro	Scopus	No	No	This paper focusses on propos	No	Accredited
3	Laborda J.; R	Feature sele	2021	Mathematics	38	This paper proposes different classification al	Article	Final	All Open Access; Go	Scopus	No	Yes	Not Applicable	Yes	Accredited
4	Zhang D.; Zho	Vertical bag	2010	Expert Systems with	129	In recent years, more and more people, espe	Article	Final		Scopus	No	No	This paper focusses on propos	No	Accredited
5	Cubiles-De-La	Improving th	2013	Expert Systems with	43	A wide range of supervised classification algo	Article	Final	All Open Access; Gre	Scopus	No	Yes	Not Applicable	Yes	Accredited
6	Li Y.; Chen W	A comparati	2020	Mathematics	97	Extensive research has been performed by ov	Article	Final	All Open Access; Go	Scopus	No	Yes	Not Applicable	Yes	Accredited
7	Ben-David A.;	Accuracy of	2009	Expert Systems with	41	Relatively few publications compare machine	Article	Final		Scopus	No	Yes	Not Applicable	Yes	Accredited
8	Djeundje V.B.	Enhancing cr	2021	Expert Systems with	91	Hundreds of millions of people in low-income	Article	Final		Scopus	No	Yes	Not Applicable	Yes	Accredited
9	Kumar A.; Sh	Machine lear	2021	Risks	42	Rural credit is one of the most critical inputs f	Review	Final	All Open Access; Go	Scopus	No	No	This paper is a literature review	No	Accredited
10	Chopra A.; Bh	Application o	2018	Business Perspectiv	41	Loan default is a serious problem in banking i	Article	Final		Scopus	No	Yes	Not Applicable	Yes	Accredited
11	Bücker M.; Sz	Transparenc	2022	Journal of the Opera	92	A major requirement for credit scoring models	Article	Final	All Open Access; Gre	Scopus	No	Yes	Not Applicable	Yes	Accredited
12	Bequé A.; Les	Extreme lear	2017	Expert Systems with	125	Classification algorithms are used in many do	Article	Final		Scopus	No	Yes	Not Applicable	Yes	Accredited
13	Teles G.; Roc	Machine lear	2020	Neural Computing ar	38	Among the numerous alternatives used in the	Review	Final		Scopus	No	No	This paper is out of scope	No	Accredited
14	Dastile X.; Ce	Statistical an	2020	Applied Soft Compu	255	In practice, as a well-known statistical metho	Article	Final		Scopus	No	No	This paper is a literature review	No	Not Accredited
15	Liu Y.; Schum	Data mining	2005	Journal of the Opera	90	The features used may have an important effe	Article	Final		Scopus	No	Yes	Not Applicable	Yes	Accredited
16	Trivedi S.K.	A study on c	2020	Technology in Socie	93	A bit hurdle for financial institutions is to decid	Article	Final		Scopus	No	Yes	Not Applicable	Yes	Accredited
17	Tsai C.-F.; Hu	Modeling cre	2014	Kybernetes	36	Purpose – Credit scoring is important for finar	Article	Final		Scopus	No	No	This paper focusses on wholes	No	Accredited
18	Tripathi D.; Ed	Credit Scorir	2018	Procedia Computer	44	Credit scoring concerns with developing empi	Conference paper	Final	All Open Access; Go	Scopus	No	No	This paper is not from a journal	No	Not Accredited
19	Tripathi D.; Ed	A novel hybri	2019	Computational Intelli	63	Credit scoring focuses on the development of	Article	Final		Scopus	No	No	This paper focusses on propos	No	Accredited
20	Ashofteh A.; E	A conservati	2021	Expert Systems with	48	This research is aimed at the case of credit s	Article	Final	All Open Access; Gre	Scopus	No	No	This paper focusses on propos	No	Accredited
21	Ha V.S.; Nguy	Credit scorin	2016	MATEC Web of Cor	36	In financial risk, credit risk management is on	Conference paper	Final	All Open Access; Go	Scopus	No	No	This paper is not from a journal	No	Not Accredited
22	Xia Y.; Zhao	A novel tree-	2020	Expert Systems with	83	Ensemble models have been extensively appl	Article	Final		Scopus	No	No	This paper focusses on propos	No	Accredited
23	Tsai C.-F.; W	Using neural	2008	Expert Systems with	399	Bankruptcy prediction and credit scoring have	Article	Final		Scopus	No	Yes	Not Applicable	Yes	Accredited
24	Koutanaei F.N	A hybrid data	2015	Journal of Retailing a	146	Data mining techniques have numerous applic	Article	Final		Scopus	No	Yes	Not Applicable	Yes	Accredited
25	Guégan D.; H	Regulatory le	2018	Journal of Finance a	23	The arrival of Big Data strategies is threatenir	Article	Final	All Open Access; Go	Scopus	No	Yes	Not Applicable	Yes	Accredited
26	Talaat F.M.; A	Toward inter	2024	Neural Computing ar	26	In recent years, the increasing prevalence of	Article	Final		Scopus	No	Yes	Not Applicable	Yes	Accredited
27	Wu Y.; Pan Y	Application A	2021	Complexity	14	Credit score is the basis for financial institutio	Article	Final	All Open Access; Go	Scopus	No	Yes	Not Applicable	Yes	Accredited
28	Tigges M.; Mo	Who gets the	2024	Technological Forec	14	Credit scoring plays an important role in deter	Article	Final		Scopus	No	Yes	Not Applicable	Yes	Accredited
29	Singh P.	Comparative	2018	Proceedings of the I	16	Credit Scoring is the primary method for class	Conference paper	Final		Scopus	No	No	This paper is not from a journal	No	Not Accredited
30	Hayashi Y.	Emerging Tr	2022	Electronics (Switzerl	13	This systematic review aims to provide deep	Review	Final	All Open Access; Go	Scopus	No	No	This paper is a literature review	No	Accredited
31	Tripathi D.; Sh	Credit Scorir	2022	Wireless Personal C	30	Credit scoring models are developed to stren	Review	Final		Scopus	No	Yes	Not Applicable	Yes	Accredited
32	Boughaci D.; A	Appropriate	2020	Risk and Decision A	23	Machine learning techniques have been used	Article	Final		Scopus	No	Yes	Not Applicable	Yes	Accredited
33	Xia Y.; He L.;	A dynamic ci	2021	Technological and E	34	Credit scoring, which is typically transformed	Article	Final	All Open Access; Go	Scopus	No	No	This paper focusses on propos	No	Accredited
34	Siham A.; Sar	Feature sele	2021	2021 International C	13	Feature Selection (FS) is one of the power sc	Conference paper	Final		Scopus	No	No	This paper is not from a journal	No	Not Accredited
35	Bono T.; Crox	Algorithmic f	2021	Oxford Review of Ec	17	The use of machine learning as an input into c	Article	Final		Scopus	No	Yes	Not Applicable	Yes	Accredited
36	Luo C.	A compreher	2020	Industrial Manageme	21	Purpose: The purpose of this paper is to prov	Article	Final		Scopus	No	No	This paper focusses on propos	No	Accredited
37	Zhou H.; Wan	Application o	2013	Proceedings - 9th In	14	Along with the increase number of users for tl	Conference paper	Final		Scopus	No	No	This paper is not from a journal	No	Not Accredited
38	Hussin Adam	Machine Lea	2022	Risks	23	Forecasting the creditworthiness of customer	Article	Final	All Open Access; Go	Scopus	No	Yes	Not Applicable	Yes	Accredited
39	Van Sang H.;	A novel cred	2016	Indian Journal of Sci	26	Background/Objectives: This article presents	Article	Final	All Open Access; Go	Scopus	No	No	This paper is not from a journal	No	Not Accredited
40	Alblooshi M.;	Unlocking Tr	2024	International Confer	6	Credit score analysis is vital to modern bankir	Conference paper	Final		Scopus	No	No	This paper is not from a journal	No	Not Accredited
41	Safiya Parvin	An Ensemble	2020	Proceedings - 2020	7	Credit scoring is a way of analyzing statistical	Conference paper	Final		Scopus	No	No	This paper is not from a journal	No	Not Accredited
42	Liu Y.; Huang	Credit scorin	2024	Journal of Forecasti	9	Credit scoring models based on machine lear	Article	Final		Scopus	No	Yes	Not Applicable	Yes	Accredited

Appendix B: Data Extraction Form

The data extraction form aides in developing a systematic approach to gathering insights. Below is a snapshot of the data extraction form, for full version please double-click on the icon to open in MS Excel:

Appendix B - Data Extraction Form.xlsx

Appendix B: Data Extraction Form

The data extraction form aides in developing a systematic approach to gathering insights

ID	Author	Journal	Yr	Title	Geographical	Dataset Type	ML Models Tested	Benchmark Model	Data Preparation Techn	Primar	Primary Outco	XAI Technique	Findings on Accuracy	Findings on Transparency	Tree-B	Neural	Kernel	Broade	Data Hy	Regula	Model	Accura	Genera	Applica	Key Quotes	Auth
10	Chopra A.; Bl	Business Pe	2018	Application of Ensemble Models in Credit Scoring Models	Not specified - paper states that the methodology was used on Indian bank data, but the	Publicly available banking dataset	Decision Tree, Bagging, Random Forests, Ada-boosting, Gradient Boosting	Decision Tree (Base learner)	Data partitioning (70% training, 30% evaluation). Iterative tuning of Bagging parameters (sample ratio, iterations).	AUC|Gini not reported.	Accuracy: 81% Recall: 52% Precision: 67% (for Gradient Boosting)	Not Applicable. The paper focuses on performance comparison and variable importance provided by the models themselves, not XAI.	"The results of the empirical analysis suggest that the gradient boosting model outperforms the base decision tree learner, indicating that ensemble model works better than individual models".	"The study identifies the relative importance of predictor variables, noting that "debt to income ratio" was the most important parameter for the best-performing model (Gradient	1	0	0	0	0	0	0	0	1	0	"This article aims to conduct empirical analysis on publically available bank loan dataset to study banking loan default using decision tree as the base learner and comparing it with ensemble tree learning techniques such as bagging, boosting, and random forests"	Refer on th "reca How is tes
56	Tran K.Q.; Di	Lecture Note	2021	Machine Learning-Based Empirical Investigation for Credit Scoring in Vietnam's Banking	Vietnam	Private data from the "Kalapa Credit Score Challenge For Students," based on real customer	LightGBM, CatBoost, Random Forest, Support Vector Machine (SVM), Logistic Regression	Single-model algorithms (SVM, Logistic Regression)	Removal of low-impactlhigh-correlation attributes, imputation (mean|mode), normalization, K-means clustering, feature engineering, and feature selection via target	AUC of 0.81 (Random Forest); 0.82 (CatBoost). Gini of	F1-Score: 0.83 (Random Forest).	Not Applicable.	"The best algorithm is Random Forest with AUC, Recall, and F1-score is 0.81, 0.81, and 0.83, respectively... The CatBoost Classifier performs the best AUC performance with 0.82".	The analysis highlights that the model with the highest AUC score (CatBoost) may not be optimal in practice due to higher profit volatility and potential financial risk. The paper notes that Random Forest	1	0	0	0	0	0	0	0	0	0	"In this paper, we aim to develop novel and effective solutions for credit scoring in Vietnam with machine learning models... We conduct experiments with modern machine learning methods based on ensemble learning models: LightGBM, CatBoost, and Random Forest".	Refer great busi selec tech risk
105	Sugianto; Wi	International	2024	Modeling and Application of Credit Scoring Based on A Multi-Objective Approach to Debtor Data in PT. Bank Riau	Indonesia	Private bank data (PT. Bank Riau Kepri) combined with insurance data	Decision Tree, Random Forest, Support Vector Machine (SVM), k-Nearest Neighbour (k-NN), Gradient Boosting, XGBoost.	The models are compared against each other rather than a single benchmark. Decision Tree seems to be the initial point of comparison	Feature selection using Featurewiz. Data partitioning (80% training, 20% testing). K-Fold cross-validation (K=2, 4, 6, 8, 10).	AUC|Gini not reported.	Accuracy: The XGBoost algorithm achieved the highest accuracy of 93.27% after K-Fold validation (with	Not Applicable.	"The K-Fold validation shows that the XGBoost algorithm has the best accuracy, with 93.27% on a K-Fold equal to 9".	Transparency is not a direct focus of the study's findings. The paper focuses on predictive accuracy and the practical deployment of the best-performing model into a web application.	1	0	1	0	0	0	0	0	0	0	"This model tested the Decision Tree algorithm with an accuracy value of 92.49%, the Random Forest with an accuracy value of 91.72%, the Support Vector Machine (SVM) with an accuracy value of 91.25%, and K-Nearest Neighbor (K-NN) with an accuracy value. 90.58%, Gradient Boosting with an accuracy value of 90.63%, and XGBoost	Noth
42	Liu Y.; Huan	Journal of Fi	2024	Credit scoring prediction leveraging interpretable ensemble learning	Not specified however it uses public dataset from AliCloud Tianchi	Publicly available dataset (AliCloud Tianchi)	LR, DT, KNN, GaussianNB, RF, Bagging, AdaBoost, GDBT, XGBoost, CatBoost, LightGBM	Five single classification models (LR, DT, KNN, GaussianNB, RF) serve as the baseline.	Data encoding, Pearson correlation for feature removal, Recursive Feature Elimination (RFE), and an improved	AUC: 0.8920 (CatBoost)	Accuracy: Not reported directly, but Recall was 83.68% and	SHAP (SHapley Additive exPlanations)	"CatBoost stands out among the models, exhibiting superior performance in recall, AUC, and G_mean evaluation indicators". "Performance evaluation indicators of the	"Empirical results reveal that the SHAP method conforms well to CatBoost and delivers a global and local interpretation of the predictions"	1	0	0	0	0	0	1	0	1	0	"We also scrutinize representative five single classification models and six ensemble learning models for credit scoring prediction. We empirically test the superior performance of ensemble learning models and identify the	Very mod focu data
6	Li Y.; Chen W	Mathematics	2020	A comparative performance assessment of ensemble learning for credit scoring	Not Specified (uses public dataset from Lending Club, USA)	Publicly available dataset (Lending Club, USA)	Random Forest, AdaBoost, XGBoost, LightGBM, Stacking, Neural Network (NN), Decision Tree (DT), Logistic Regression (LR), Naive Bayes (NB), SVM	Five individual learners (NN, DT, LR, NB, SVM)	Manual feature selection (removing irrelevant and highly correlated variables). Data partitioning (80% training, 20% testing) with 5-fold cross-validation . Grid search for hyper-parameter tuning	AUC: 0.9580 (Random Forest)	Accuracy: 81.05% (Random Forest) KS Statistic: 0.5507 (Random Forest) Brier Score: 0.1260 (Random Forest)	Not Applicable (provides ROC curve analysis)	"Experimental findings reveal that the performance of ensemble learning is better than individual learners... In addition, random forest has the best performance in terms of five metrics, XGBoost and LightGBM are close challengers".	The ROC curve analysis visually demonstrates that the Random Forest curve "lies above all the other curves across all threshold values," providing a clear, comparative view of model performance .	1	0	0	0	0	0	0	0	0	0	"In this research, we provide a comparative performance evaluation of ensemble algorithms, i.e., random forest, AdaBoost, XGBoost, LightGBM and Stacking, in terms of accuracy (ACC), area under the curve (AUC), Kolmogorov–Smirnov statistic (KS), Brier score (BS), and model operating time in terms of credit scoring".	Prov rang oper thou conc well
31	Tripathi D.; S	Wireless Pe	2022	Credit Scoring Models Using Ensemble Learning and Classification Approaches: A Comprehensive Survey	Multi-country (utilising benchmark datasets from Taiwan, Germany, Japan, and Australia)	Public benchmark datasets from UCI Repository (Taiwan), Bank-marketing, German, Japanese, Australian)	The study tested five base classification models: Partial Decision Tree (PART), Radial Basis Function Neural Network (RBFN), Logistic Regression (LR), Naive Bayes Decision Tree (NBT), and Sequential Minimal Optimization (SMO). These were combined with nine different ensemble	Not explicitly a **single** benchmark model. The study compared the performance of individual classifiers against various ensemble configurations (single-layer, multi-layer, different aggregation methods) to see which combination performed best.	The primary data preparation techniques mentioned were the elimination of samples with missing values and the use of label encoding to transform categorical features into a numerical format for the models. 10-Fold-Cross-Validation .	AUC|Gini not reported.	Accuracy: 90.05% on Australian dataset (MultiBoost with Layered Weighted Voting) . 92.86% on Bank-marketing dataset (MultiBoost with Layered Weighted Voting) .	Not Applicable.	"From the experimental outcomes, it is observed that Multiboost and Dagging are best ensemble learning approaches... Overall, MultiBoost and Dagging with multilayer ensemble frame is the best approach for credit score classification".	Not Applicable. The paper does not address the topic of model transparency or explainability.	1	0	0	1	0	0	0	0	0	0	"So, this study presents a comprehensive comparative analysis of nine ensemble learning approaches such as Multiboost, Cross Validation Parameter, Random Subspace, Metacoast, etc. with five classification approaches such as Partial Decision Tree (PART), Radial Basis Function Neural Network (RBFN), Logistic Regression (LR), Naive Bayes Decision Tree (NBT) and Sequential Minimal Optimization (SMO) along with various ensemble classifiers frameworks arranged in single and multi layer with various aggregation approaches".	The tests ense archi The in a l key
85	Zhirov V.K.; S	Journal of P	2021	Neural network as a tool to solve the problem of credit scoring	Russia	Compiled Dataset from Literature	Artificial Neural Network (Feedforward)	Logistic Regression (conceptual comparison)	Binary Conversion	Not Reported	Accuracy: 95%	Not Applicable	"The neural model developed in this work for solving the problem of credit scoring has demonstrated the effectiveness of this methodology... Testing the system	Not Applicable	0	1	0	0	0	0	0	0	0	0	"The advantage of an artificial neural network over the logistic regression method is the ability to support nonlinear solutions, as well as the lack of processing of complex functions used in solving in the developed	This conc How limit
5	Cubiles-De-L	Expert Syste	2013	Improving the management of microfinance	Peru	Private data from a Peruvian Microfinance	• Linear Discriminant Analysis (LDA) • Quadratic Discriminant	Classical statistical models (LDA, QDA, Logistic Regression) [1-3]. Established classifiers including ANN, SVM, KNN, CART, J4.8, and LR-R	• Outlier Removal (top|bottom 1%) • Feature Selection (Removing	Best Model (MLP 14); AUC	Best Model (MLP 14); Accuracy: 88.23%	Not Applicable	"After examining the wide set of models that were fitted in this paper, a MLP model trained inside the bias-statistical	"A possible reluctance to use a multilayer perceptron in credit	0	1	0	0	0	0	0	0	0	0	"Microfinance institutions (hereafter, MFIs) offer saving services and small loans (namely microcredits) to those agents of the non-f ative with a very limited access to	Com mark Para
12	Bequé A.; Le	Expert Syste	2017	Extreme learning machines for credit scoring: An empirical evaluation	Multi-country (Public Datasets)	Public Benchmark Datasets (Australian Credit, German Credit,	• Extreme Learning Machine (ELM) • Artificial Neural Network (ANN)	Gradient Boosting (GB) and Support Vector Machine (SVM)	• Standardisation of numeric variables • Weight-of-Evidence (WOE) transformation for nominal	Best Model (Bagging LR-R on Australian	Best Model (Bagging LR-R on Australian Credit dataset):	Not Applicable	"Correct tuning of meta-parameters confirms that ELM might outperform industry standard (LR) and other all other involved classifiers."	"ELM inherit the black-box character of ANNs. Comprehending the relationship between input variables and ELM prediction is as	0	1	0	0	0	0	0	0	0	0	"This paper explores the potential of extreme learning machines (ELM), a recently proposed type of artificial neural network, for consumer credit risk management. ELM possess some interesting properties, which might enable	This Extre agai key
3	Laborda J.; F	Mathematics	2021	Feature Selection in a C	Taiwan	Public benchmark	Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Random Forest	The paper compares the ML models against each other, both with and without feature selection, rather than against a single benchmark model.	**Feature Selection:** • Filter Method (Chi-squared test, Correlation coefficients) • Wrapper Methods (Forward stepwise selection, Backward	Not Reported	Mean Absolute Error (MAE): The best-performing model (Random Forest with forward	Not Applicable	"The results suggest that forward stepwise selection yields superior performance in each one of the classification algorithms used". This method led to models with lower	Not Applicable.BUT the feature selection process makes models "simpler, making them more understandable". Models with fewer	0	0	1	0	1	0	0	0	0	0	"Different classification algorithms-logistic regression, support vector machine, K-nearest neighbors (KNN), and random forest-(are proposed) in order to identify which candidates are likely to default for a credit scoring model.	Filter meth featu (lowe
70	Dessain J.; B	Communica	2023	Cost of Explainability in AI: An Example with Credit Scoring Models	Europe	Private insurance data (Trade Credit Insurer)	A total of twelve models were tested: Logistic Regression (LR), Elastic Net (ELN), Naive Bayes (NB), Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting (GB), Light GBM (LGBM), XGBoost (XGB), Multi-Layer Perceptron (MLP), Explainable Boosting Machine (EBM), Gami-net	Inherently Explainable Models (e.g., Logistic Regression, GAMs) and a "No Model" scenario for financial baseline.	K-Nearest Neighbors (KNN) for missing data imputation, adding historical defaults to manage class imbalance, winsorisation for outlier management, removal of highly correlated variables for feature selection, and standardisation| normalisation of financial features.	AUROC: 77.11% (achieved by XGBoost model).	Financial Return on Investment: MLP was the best financial performer, achieving a return of 1.82% (and up to 2.06% with a higher risk appetite), outperforming models with higher statistical AUROC scores.	Isotonic Smoothing on GAMs: The study applied isotonicity constraints to GAMs (EBM, GAMI), forcing features to align with expert knowledge (e.g., ensuring higher debt-to-equity ratios lead to higher risk predictions).	The "cost of explainability" was quantified as a 15-20 basis point reduction in annual return on investment when using the best explainable model compared to the best black-box model. It also found that enhancing GAMs with expert-led isotonicity improves interpretability without sacrificing financial or statistical performance. "By introducing expert judgment to obtain isotonic GAMs, we improve the explainability of the GAM without compromising the financial or	0	0	0	1	0	1	0	1	0	0	"We examine twelve different models, which are detailed in [4–14]. Of these models, four are considered inherently "explainable": Logistic Regression (LR), Elastic Net applied to a logistic regression (ELN), Naive Bayes (NB), and Linear Discriminant Analysis (LDA) . Six models are classified as black-boxes interpretable ex-post: Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting (GB), Light GBM (LGBM), eXtreme gradient Boosting (XGB), and Multi-Layer Perceptron (MLP). In addition to these models, we also test two recent explainable Generalized Additive Models (GAMi..."	I che relev pract mea use adhe requ Altho really cons
7	Ben-David A.	Expert Syste	2009	Accuracy of machine learning models versus "hand crafted" expert systems - A credit scoring case study	Israel	Private Bank Data	The study compared a handcrafted expert system against dozens of ML models available in the Weka workbench, including NaiveBayes, RBFNetwork, various tree-	A "mind crafted" credit scoring expert system developed through knowledge engineering interviews with human experts.	The study treated the problem as both a classification and a regression task. All ML algorithms were used with their default parameter settings in Weka to simulate a naive user and avoid overfitting from	Not Reported. The primary evaluation metrics used were	Cohen's Kappa: NaiveBayes (0.3292), RBFNetwork (0.3148), Expert System (0.3116). Hit Ratio	Not Applicable. The study compared ML models to an inherently transparent benchmark (the expert system) rather than applying post-hoc XAI techniques to the ML models themselves.	"No machine learning model showed any statistically significant advantage over the expert sys-tem's accuracy when the same problem was treated as classification". However, seven ML models did show a statistically significant advantage in the regression experiment	The expert system was considered transparent because the experts "felt more comfortable with the rules they defined themselves" and had little trust or control over machine learning models. The choice of the accuracy metric (Kappa vs. Hit Ratio)	0	0	0	1	0	0	0	0	0	0	"The accuracy of a real world "mind crafted" credit scoring expert system is compared with dozens of machine learning models. ... Unlike the above reports, which compare the accuracy of a single expert system with one machine learning model, this research performs a comparison with dozens of such models, making it the most comprehensive, methodological, research published	Altho rema uniq syste how drast perfo

Appendix C: Codebook

The three stages of thematic synthesis as described by Thomas and Harden (2008) were adopted, that is, coding the text line by line (free coding), grouping these codes into related categories to form descriptive themes, then, generating analytical themes from them. Below is a snapshot of the data extraction form, for full version please double-click on the icon to open in MS Excel:

Appendix C-Codebook.xlsx

Appendix C: Codebook

The three stages of thematic synthesis as described by Thomas and Harden (2008) were adopted, that is, coding the text line by line (free coding), grouping these codes into related categories to form descriptive themes, then, generating analytical themes from them.

#	Theme	Code	Definition	Paper Title	Reference	Instance(s) from article	Summary of Article
1	Traditional Model Limitations	General limitations of traditional models	Identifies common limitations in traditional scoring models	Transparency, auditability, and explainability of machine learning models in credit scoring	Bücker et al., 2022	"Thus, in credit scoring, very simple predictive models such as logistic regression or decision trees are still widely used and the superior predictive power of modern machine learning algorithms cannot be fully leveraged. Significant potential is therefore missed, leading to higher reserves or more credit defaults."	This paper discusses the trade-off between the interpretability of traditional models like logistic regression and the superior predictive accuracy of modern machine learning algorithms. It argues that while traditional models are valued for their transparency, their limitations in predictive power and handling complex data can be financially detrimental. The authors propose a framework (TAX4CS) to make "black box" machine learning models more transparent and auditable, demonstrating that it is possible to achieve interpretability comparable to traditional models while benefiting from the enhanced performance of machine learning.
					Bücker et al., 2022	"Conversely, it is not possible to cover nonlinear high-order multidimensional dependencies in the data. This becomes a potential source of error resulting from the manual interference if the number of variables is large."	
				Application of Ensemble Models in Credit Scoring Models	Chopra & Bhilare, 2	"The traditional LDA and logistic regression have low classification accuracy in the credit scoring, as the relationship among variables is linear."	This article focuses on the shortcomings of traditional credit scoring models, such as logistic regression and linear discriminant analysis (LDA), within the modern competitive banking environment. It highlights that these models often have low classification accuracy because they assume linear relationships between variables and struggle with large datasets. The paper empirically demonstrates that ensemble machine learning techniques, particularly gradient boosting, significantly outperform traditional models, and it
					Chopra & Bhilare, 2	"However, these models are less resilient when it comes to large amounts of data input; therefore, some of the assumptions in the classical statistical analysis fail. This influences the accuracy of prediction and model generalizations."	
				Algorithmic fairness in credit scoring	Bono et al., 2021	"A related notion is that a simple, transparent model will lead to fairer outcomes. However, Kleinberg and Mullainathan (2018) find that simplistic models can breed unfairness and that, by increasing accuracy, more flexible algorithms can actually reduce demographic biases."	This study investigates whether switching from a traditional logit credit scoring model to more complex ensemble machine learning methods impacts algorithmic fairness. It acknowledges that while traditional models are more interpretable, this can come at the cost of predictive accuracy. The research confirms that machine learning models are more accurate and do not exacerbate, and may even slightly improve, statistical fairness issues present in traditional models. The paper concludes that simple, traditional models are not inherently fairer and that their limitations can perpetuate biases found in historical data
					Bono et al., 2021	"Linear models are popular in credit scoring agencies partly because they give more interpretable results than other machine learning methods, like ensemble models or neural nets. This may come at the expense of prediction power."	
				Comparing the Effectiveness of Machine Learning and Deep Learning Models in Student Credit Scoring: A Case Study in Vietnam	Thuy et al., 2025	"Numerous studies have demonstrated that ML-based models outperform traditional statistical methods in predicting loan defaults (Abbas and Hussein 2024; Golbayani et al. 2020; Schmitt 2022).".	This study compares the effectiveness of four machine learning and deep learning models (Random Forest, Gradient Boosting, Support Vector Machine, and a Deep Neural Network) for predicting student loan eligibility in Vietnam, an emerging market where students often lack formal credit histories. The research uses non-traditional data specific to the student experience, such as tuition fees, living costs, part-time employment status, and academic background, as proxies for financial stability and discipline. The findings indicate that Deep Learning achieved the highest classification accuracy, highlighting the value of applying advanced models to alternative data to assess "thin file" applicants.
				Credit scoring prediction leveraging interpretable ensemble learning	Liu et al., 2024	"Despite being able to explicate the causal relationship between borrowers' credit features and credit default behavior (Giudici et al., 2020; Kyeong & Shin, 2022), traditional statistical methods need to be more accurate (Barboza et al., 2017; Bueff et al., 2022)."	This article focuses on achieving a balance between accuracy and interpretability in credit scoring models. It achieves transparency by employing the SHAP (SHapley Additive exPlanations) method to interpret the predictions of a CatBoost ensemble learning model. The study demonstrates SHAP's ability to provide both global explanations for the model's overall behaviour and local explanations for individual predictions. This enhances transparency by making the "black-box" model's reasoning understandable, which is a critical requirement for financial institutions.
2	Emerging Market Considerations	Application with Alternative Data	Studies that apply ML models to non-traditional data sources to score individuals, particularly those with thin credit files.	The Effect of AI-Enabled Credit Scoring on Financial Inclusion: Evidence from an Underserved Population of over One Million	Li et al., 2024	"The enhancement in financial inclusion came from the improved prediction accuracy of the AI model. The use of both weak signals and sophisticated machine learning algorithms contributed to this accuracy improvement because they could help generate novel features that are predictive of creditworthiness and connect these features to creditworthiness in complex and novel ways".	This study provides empirical evidence from a major bank on the impact of adopting an AI-enabled credit scoring model for an underserved population of over one million people. It demonstrates how the AI model, by leveraging "weak signals" (alternative data such as social security fund data and in-app behaviours) and advanced machine learning algorithms, significantly enhances financial inclusion. The research found that this approach simultaneously increased the loan approval rate for the underserved population while reducing the default rate for all borrowers. The core mechanism is the AI's ability to improve prediction accuracy, thereby reducing reliance on traditional "strong signals" like credit history, which often marginalises individuals with thin files.
					Li et al., 2024	"Unlike strong signals, weak signals can cover a broader range of financial or nonfinancial data domains. Previous studies have discussed the following data domains of weak signals: electronic footprint and trajectory (Kim, 2020; Lu et al.,2023), social networks (Lin et al., 2013; Lu et al., 2012; Gao et al., 2022), educational background (Li & Hu, 2019), lenderborrower communication (Xu & Chau, 2018), mobile phone usage (Ma et al., 2018; Lu et al., 2023), facial information (Chen et al., 2023), and other soft information (Iyer et al.,2016, Hou et al., 2023).".	
				Enhancing credit scoring with alternative data	Djeundje et al., 20	"Hundreds of millions of people in low-income economies do not have a credit or bank account because they have insufficient credit history for a credit score to be ascribed to them".	This paper evaluates the predictive accuracy of credit scoring models that use alternative data—specifically psychometric variables and email usage characteristics -to assess the credit risk of individuals with insufficient or no credit history ("thin files"). It compares various statistical and machine learning classifiers, demonstrating that models incorporating these non-traditional data sources can achieve greater predictive accuracy than those relying solely on demographic data. The study concludes that using alternative data can
					Djeundje et al., 20	"This paper reports on experiments to assess the predictive accuracy of credit scoring models that use certain types of alternative data instead of, or as well as conventional predictors. The aim of this paper is to evaluate the predictive performance of using psychometric variables and/or characteristics of email usage to predict the	

Appendix D: Practical Contribution of the Study to the South African Banking Sector (Dashboard)

Below is a visual dashboard summarising the practical implications of this research for the South African banking sector:

Insight Area	Key Findings from the Study	Practical Implications for SA Banks	Recommended Actions	Value to SA Banks
1. Traditional Model Limitations	Traditional scorecards (e.g., logistic regression) are transparent but rely on linear assumptions that limit predictive accuracy, especially in complex or non-linear datasets.	Banks relying solely on linear models risk underestimating or misclassifying borrower risk, particularly for thin-file or informal economy clients.	Use ML for data exploration and feature engineering to uncover non-linear relationships; retain traditional methods for final decisioning where required for transparency.	Improves model robustness and enhances regulatory compliance by maintaining interpretability.
2. Machine Learning (ML) Application in Credit Scoring	Tree-based ensemble methods (Random Forest, XGBoost, LightGBM, CatBoost) consistently outperform traditional models across accuracy metrics (AUC, Gini, KS).	ML methods can strengthen portfolio discrimination and stability, improving default prediction accuracy and capital allocation.	Implement hybrid scoring frameworks that embed tree-based ensembles for predictive layers while retaining interpretability layers (e.g., SHAP-based explanations).	Supports improved credit risk segmentation and profitability without compromising governance standards.
3. Transparency & Explainability (XAI)	Explainable AI (especially SHAP) enables model interpretability, aligning ML outputs with audit and regulatory expectations.	Transparency barriers can be mitigated using XAI tools, allowing complex models to meet documentation and model risk requirements.	Embed SHAP explanations into the ML development pipeline; produce both global and local explanations for model monitoring and audit.	Strengthens regulatory compliance, auditability, and trust in AI-driven models.
4. Data Preparation & Feature Engineering	ML adds greatest value in pre-processing: feature selection, imbalance correction, and data enrichment significantly impact downstream accuracy.	Many “black box” risks arise not from model choice but from undocumented data pre-processing steps.	Apply ML for automated feature selection and class rebalancing; ensure full governance documentation for pre-processing pipelines.	Enhances data quality and model stability, reducing unanticipated model risk.
5. Emerging Market Considerations & Financial Inclusion	ML models using alternative data (e.g., mobile, utility, or behavioural data) can effectively score thin-file customers, promoting inclusion.	Responsible alternative data use can expand credit access, but carries fairness and bias risks.	Run pilot projects to test inclusion and bias impact of alternative data sources; accompany pilots with fairness, privacy, and ethics testing.	Expands financial inclusion while maintaining consumer protection and ethical standards.

6. Regulatory & Governance Readiness	South African regulation (e.g., SARB model risk guidance, NCA) is less prescriptive for business scorecards than for regulatory models.	Banks can responsibly innovate using ML within business (non-regulatory) models, provided transparency and documentation are maintained.	Develop hybrid governance frameworks that adapt international ML validation standards (e.g., explainability tests, challenger monitoring) to local regulatory requirements.	Positions banks for innovation while maintaining regulatory readiness and reducing supervisory friction.
7. Strategic Implementation Roadmap	ML's greatest immediate value lies in augmentation, not replacement, of traditional models.	Gradual, well-documented adoption allows innovation without breaching governance principles.	Adopt a staged hybrid strategy: ML for data prep, traditional model or interpretable layer for decisioning, XAI for explainability.	Provides a balanced path for innovation -enhancing accuracy, transparency, and compliance simultaneously.