Evaluating machine learning models for credit risk prediction across retail segments of South Africa

Willem D. Pieters

orcid.org/0009-0005-2373-7229

Mini-dissertation accepted in partial fulfilment of the requirements for the degree **Master of Business Administration (MBA)** at the North-West University

Supervisor: Prof JD van Romburgh

Graduation: June 2026

DECLARATION

I certify that this research project, titled “Evaluating machine learning models for credit risk prediction across retail segments of South Africa”, is my own and is being presented in partial fulfilment of the Master of Business Administration degree requirements at the North-West University. It has never been submitted for a degree or examination at any other university.

Additionally, I certify that I have obtained all necessary authorisations and consent to conduct this study.

DEDICATION

In loving memory to my late daughter E.

To my daughter A. and my wife H.

ACKNOWLEDGEMENTS

The author is grateful to the secondary data provider, who wishes to remain anonymous.

The author is grateful for the funder of the cloud computing cost, who wish to remain anonymous.

The author gives thanks to the North-West University Business School and its lecturers.

The author is grateful to Mr. Andries Blaauw, for the guiding inputs and discussions on the research quantitative approach.

ABSTRACT

The current study evaluated the relevance of product segmentation in credit application scorecards within the retail credit industry of South Africa. It specifically tested whether developing distinct models for different product populations yields a better predictive performance on a single model trained on a combined dataset.

A quantitative experimental design was employed utilising the AutoGluon automated machine learning (AutoML) framework to train and evaluate competing models (including neural networks and gradient boosting ensembles). The study compared the Area Under the Curve (AUC) performance of the models trained on the segmented product data against the single generalised model.

The results indicated that segmenting by credit product types, using the same target definition, over the same cross sectional time frame did not improve the overall model performance. Contrary to industry norms, the single generalised model achieved a higher AUC to that of the segmented models across all three product categories.

The study concludes that training models on combined datasets resulted in superior risk differentiation compared to segmented datasets. This suggests that modern AutoML frameworks leverages increased data volume more effectively than traditional segmentation in the retail credit environment.

Keywords: **Application scorecards, AutoGluon, AutoML framework, Credit risk, Ensemble models, Probability of default, Quantitative risk management, Risk assessment, Segmentation, Supervised machine learning.**

LIST OF ABBREVIATIONS

Abbreviation	Description
AI	artificial intelligence
AutoML	automated machine learning
AUC	area under the curve
ECL	expected credit loss
ERM	enterprise-wide risk management
GBDT	gradient boosting decision tree
KNN	k-nearest neighbours
LightGBM	light gradient boosting machine
ML	machine learning
NCA	National Credit Act 34 of 2005
NCR	National Credit Regulator
NCT	National Consumer Tribunal
NWU	North West University
POPIA	Protection of Personal Information Act 4 of 2023
RF	random forest
ROC	receiver operating characteristic
RP	research purpose
RQ	research question
SACRRA	South African Risk and Reporting Association
XGBoost	extreme gradient boosting

DEFINITIONS

Concept	Description
Application scorecards	Statistical credit risk assessment tool assessing the creditworthiness of credit consumers (Kritzinger & van Vuuren, 2021:270).
Data wrangling	The process of manipulating or transforming data into a desired format that is utilised for desired analysis (Petricek *et al.*, 2023).
Area under the ROC curve (AUC)	Is a performance indicator to assess and compare the efficiency of a binary classifier model (Gasmi *et al.*, 2025:12).
Gradient boosting	Modified boosting technique where multiple decisions trees are trained where the loss function is optimised (Chopra & Bhilare, 2018:3).
Homogenous	Of the same kind, low variation within a sample of data but different to another sample (Bryman *et al.*, 2021:229).
Loansharks	Informal and unregistered lenders, sometimes exceeding the prescribed interest rate charges on loans (Siyongwana, 2004:851).
Lobola	Also referred to as ‘bride price’, this practice is not unique to South Africa but also occurs in other parts of the world. It is a payment to a bride’s family at marriage by the groom (Ashraf *et al.*, 2020:617).
Machine learning	A subset of artificial intelligence, guided by algorithms and statistics as part of a computer program (Singh *et al.*, 2023:62).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION AND BACKGROUND

1.1 Introduction to credit in South Africa

Retail credit for individuals, unlike corporate credit, has unique challenges including high transaction volume and smaller loan sizes (Allen *et al.*, 2004:727). As such, in the event of a single retail credit consumer defaulting on their credit repayment, it is unlikely to have severe financial effects on the lender (Allen *et al.*, 2004:728). On the contrary, corporate structured institutions that are borrowers of large loans require more comprehensive risk analyses, as applying the same approaches needed for smaller, high volume loans will not be economically feasible (Allen *et al.*, 2004:728). These comprehensive risk analyses on wholesale consumers of large loans, can be outsourced to credit agencies (e.g., Moody's Investor Services, Standard & Poor, and Fitch Ratings) to supplement the lender's internal assessments (Allen *et al.*, 2004:735).

Loans issued to individual borrowers (i.e., not businesses) under specific repayment terms and conditions are referred to as the debtors book of a company (Allen *et al.*, 2004:733). Such loans may be term-based, revolving, secured by an underlying asset, or unsecured. These various concepts of lending are not a new phenomenon, as banks issuing loans have been documented since the mediaeval times (Brigham *et al.*, 2019:111). While the reasons for borrowing may vary among governments, corporates, and individuals, a common factor is the desire to access funds in advance. Reasons for borrowing money in South Africa can be broadly categorised as follow:

- Essential life changing events, which include subcategorises such as marriage or bride wealth (i.e., *Lobola*), funerals, and higher education (James, 2012:20);
- Consumption and aspirations, which cover items that reflect a high social status such as valuable goods, designer clothes, luxury cars, household furniture, and home appliances (James, 2012:29);
- Financial needs and coping mechanisms, which refer to short-term borrowing needs, such as loans due in the next 30 days, funds for gambling, or acquiring basic necessities (James, 2012:25); and

- Businesses and entrepreneurial activities. The Basel Committee on Banking Supervision (2005:49) describes that business-orientated, specialised borrowings are often used for building projects, physical assets (i.e., fleets), commodities (i.e., raw materials), and commercial property development.

In South Africa, the National Credit Act (NCA) 34 of 2005 ("the act") defines the legal framework by which all credit providers must comply. Section 12 of the act establishes the National Credit Regulator (NCR) as the enforcing body, which oversees fairness to consumers and standardised frameworks between credit providers (Schmulow, 2017:242). Furthermore, Section 26 of the act establishes the National Consumer Tribunal (NCT), which serves as the judicial body with investigative authority to bring matters related to credit before the court for prosecution (Schmulow, 2017:243). The act also sets out boundaries in which investigations and witnesses are summoned. In case offences should occur, the NCT may impose administrative penalties of up to ZAR 100 000, or 10% of the respondents turnover (Schmulow, 2017:245). In an effort towards public transparency, the NCR releases public circulars of offenders and corresponding imposed fines on their website (for example, findings on debt counsellors were reported in March 2025 Issue 1). For repeated offences, the NCT may suspend or cancel the respondent's registration leading to the eventual loss of its license (Brits, 2018:12). Prior to the implementation of the NCA, lending practices were considered irresponsible and contributed to over-indebtedness of South Africans (James, 2012). Furthermore, the purpose of the NCA is outlined in Section 3(b) of the act, where it states that the NCA works to "…*ensuring consistent treatment of different credit products and different credit providers*…", and in Section 3(c), "…*promoting responsibility in the credit market.*" This legislation, therefore, regulates different types of credit providers including banks, vehicle financiers, and clothing and furniture retailers (James, 2012:21).

Elaborating on Section 3(c)(i) of the NCA, the onus rests on credit providers to remain responsible when granting credit. As such, credit providers must avoid granting credit that would lead to over-indebted consumers. For this reason, credit providers are required to assess consumers' available cash flow beforehand, commonly referred to as an affordability assessment. The purpose of an affordability assessment involves determining if the credit consumer can meet the monthly instalments of the new loan (Schmulow, 2017:236). Notwithstanding, Section 3(c)(ii) of the NCA addresses the issue

of reckless lending. This refers to the practice of when credit is granted while failing to take the consumer's credit history into account (Schmulow, 2017:224). On this backdrop, credit providers use tools to assess credit default risk prior to issuing a credit facility - even if the customer has the affordability meeting the instalments. If the court finds the credit provider acted reckless, the agreement may be set aside leaving the credit provider with the losses and reputational damage (Brits, 2018:2).

James (2012:21) further explains that the NCA regulates all credit practices ‘at arm’s length’, including both the well-known formal market and the controversial practice of informal lending at exorbitant interest rates known as ‘loansharks’, the latter which is illegal in South Africa. Section 4 of the NCA is applicable to every credit agreement between parties dealing at arm’s length. Specifically, Section 4(2)(b) exclude loans between shareholders, familial relationships, dependants, and co-dependants.

Section 13(c) of the NCA sets out the monitoring responsibilities of the NCR to report to the Minister of Trade and Industry. These reports must include available credit, market conditions, pricing, trends, and stakeholder conduct (*National Credit Act*, 2005). Moreover, these reports on credit availability is monitored across various credit types for example term loans, unsecured loans, short terms loans, etc; as well as into different credit industries, for example banks and retailers. Lastly, the number of loan application received and the proportions that was rejected are also included in the reports to the Minister.

1.1.1 Credit landscape of South Africa

The workforce in South Africa consists of approximately 25.1 million people with 17.1 million people being employed as of the fourth quarter (Q4) in 2024 (Stats SA, 2025a:1). In the same period, 28.6 million active credit consumers were reported (National Credit Regulator, 2025a:1). From these data it can be calculated that 60% of people in South Africa have access to credit or a service agreement whilst being formally employed. The remaining 40% consist of natural persons without formal employment.

The fees and interest rates of credit consumers are determined by the Minister of Trade and Industry. Conversely, the maximum interest rate is linked to the repo rate of South Africa as determined by the South African Reserve Bank (Department of Trade and Industry (South Africa), 2015). In South Africa, total value of loans granted for Q3 ending in December 2024 was ZAR 158.7 billion, averaging approximately ZAR 52.9 billion per month (National Credit Regulator, 2025b:3). The value of granted loans per industry consists of 78.8% accounted for banks, 7.4% for non-banked vehicle financiers, 7.7% for retailers, and 6.1% for the remaining industries combined (National Credit Regulator, 2025b:4). Although retailers contributed a relatively small amount of the overall loan values across numerous industries, the value is comparable to non-banked motor vehicle financiers. A depiction of new loans granted by credit types over time is shown in **Figure 1**. New loans granted across the different credit types remained relatively consistent in Q1 and Q2 of 2024. In Q4 of 2024, both secured loans and unsecured loans had a noticeable increase in value. This trend was likely driven by seasonal factors such as higher consumer demand during the festive period.

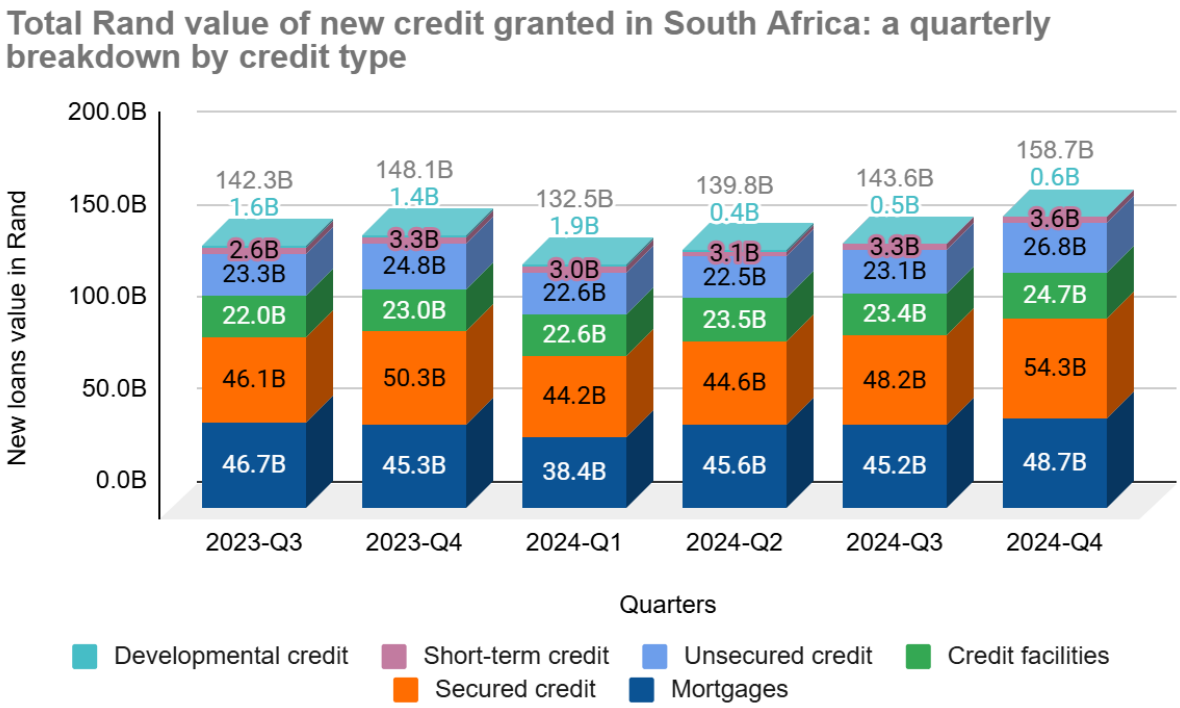

Figure 1.1: South African credit granted by credit type (Quarterly, ZAR; Source: National Credit Regulator, 2025a:4)

Despite new credit loan values in different credit types not yielding a definite trend, the retail credit industry still showed growth. This is illustrated in **Figure 1.2**, which shows a gradual increase in both new loans and the total book value of credit granted from Q4 2023 to Q4 2024. The average growth rate for new loans was 16.6% per quarter, with the highest being nearly 30% and the lowest 6%. Furthermore, **Figure 1.2**. also shows that the total book value had a 5.2% average growth rate per quarter, which indicates strong growth in the retail credit industry in South Africa.

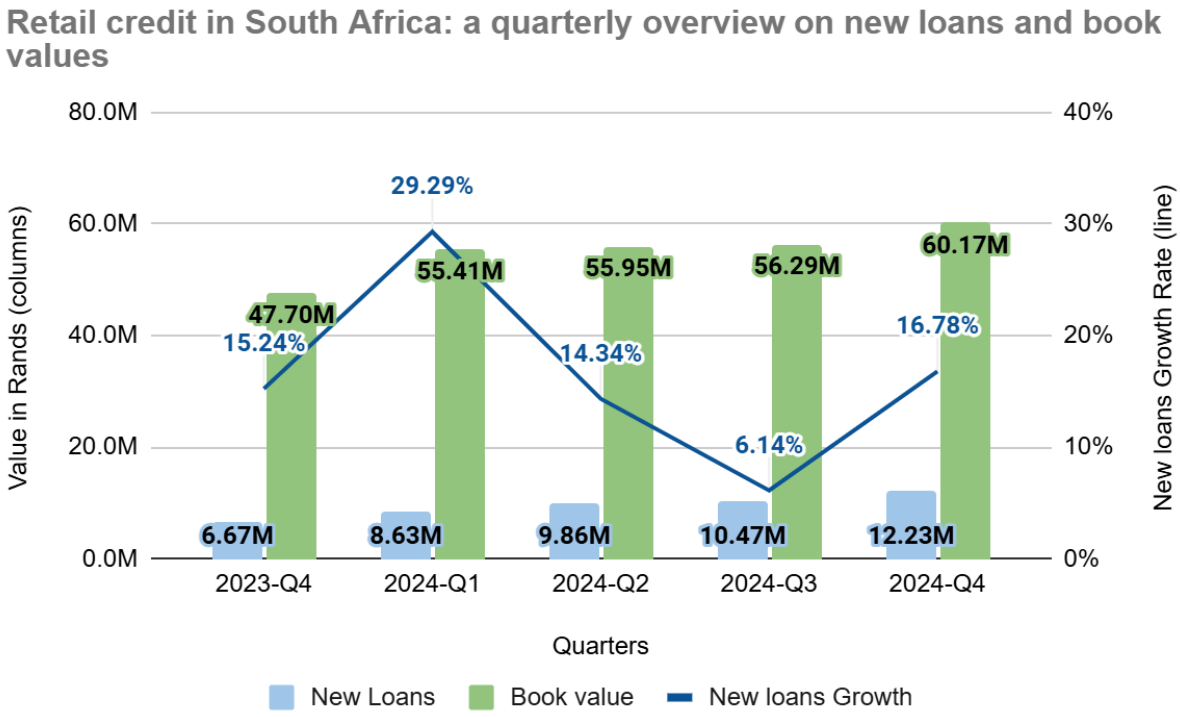

Figure 1.2: New loans and book value in the South African retail credit industry (Quarterly, ZAR; Source: National Credit Regulator, 2025a:4)

The depiction above reported on the nominal prices (unadjusted), the price of the item. Real prices measure the price movements adjusted for inflation, whereas nominal prices are the monetary value of goods sold at a point in time (Stats SA, 2025b:15). The difference between the two is the impact of the Consumer Price Index. Real prices have been observed as cyclical and on a downward trend since 2022 (**Figure 1.3**). Overall

nominal retail sales value grew by a mean 6.3% (Stats SA, 2025b:9), whilst credit sales grew by a mean 16.6% (National Credit Regulator, 2025b:4). This indicates strong demand for credit in the growing South African retail industry. The credit growth rate is normally higher than the nominal retail sales growth rate.

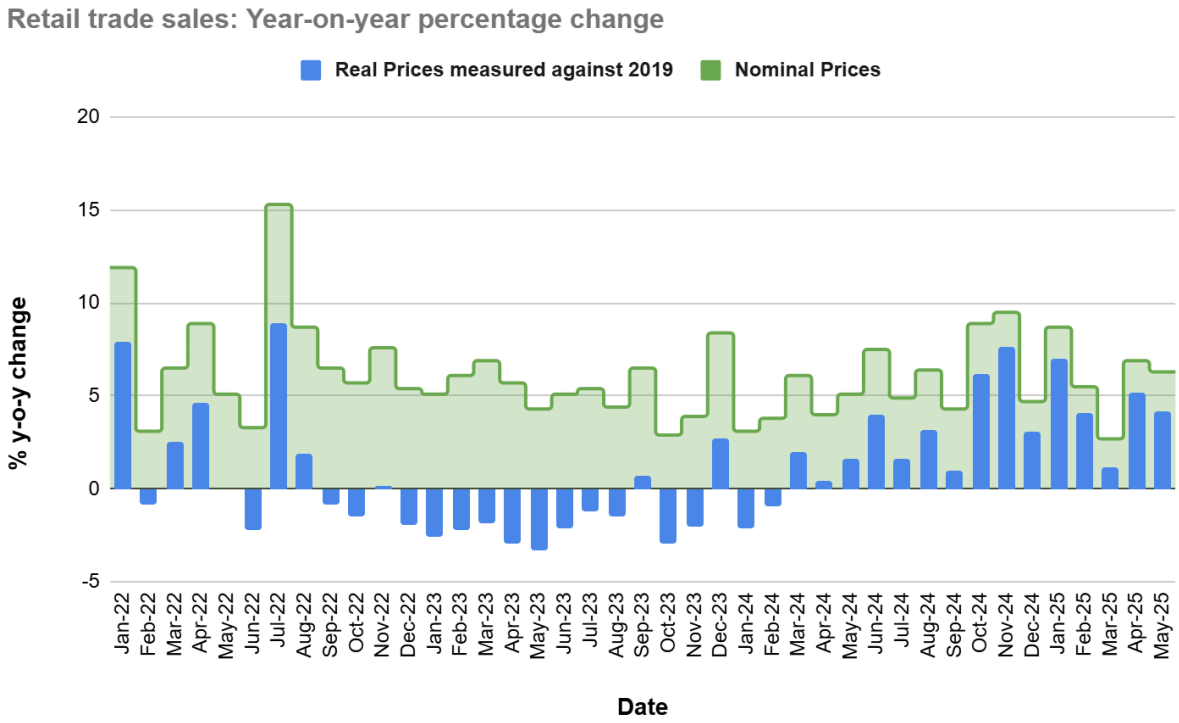

Figure 1.3: Retail trade sales prices: Year-on-year percentage change (Source: Stats SA, 2025b:6,9)

1.1.2 Regulatory reporting mandates for credit providers

All credit providers in South Africa must regularly report loan status, balances, and payment history to the NCR (National Credit Regulator, 2024:24). This process is based on NCR's guidelines and directed by the South African Credit and Risk Reporting Association (SACRRA), who is the only prescribing association in South Africa that standardise credit data submissions to credit bureaus (National Credit Regulator, 2024:2). Regular submissions may only take place through a secure portal called the

Data Transmission Hub. Ownership of the Data Transmission Hub is shared between SACRRA and the Credit Bureau Association (National Credit Regulator, 2024:2), signalling independence from the South African Government and from the private sector. As part of SACRRA's guidelines, minimum timeframes for reports to be completed are also stipulated. For example, all new and closed credit agreements are to be reported within 48 hours and monthly payment information is to be reported within 5 business days of the billing cycle (National Credit Regulator, 2024:6). These reporting guidelines ensure that credit assessments are accurate with recent information.

When applying for credit, personal identifiable information of the credit consumer must be provided in order to compile a complete credit profile for each individual. Information from credit providers (e.g., a retailer) is also required, such as the credit agreement number and type of credit account (National Credit Regulator, 2024:8). This allows registered credit bureaus to access standardised information from all credit providers and subsequently utilise the information to create credit profiles of consumers (National Credit Regulator, 2025a:1). Section 43 of the NCA describes how credit bureaus are registered, it also regulates that credit bureaus may not control shareholders of credit providers, debt collectors or any person (*National Credit Act*, 2005). Moreover, credit bureaus differentiate themselves by how they utilise these data and provide value added services back to credit providers, ensuring that improved credit risk decisions are made. One such example is the development of a credit risk assessment tool called application scorecards within the companies' risk management framework.

The limitation in the banking industry is that regulation does not allow for ungrouped risk assessment by individual credit products (Basel Committee on Banking Supervision, 2005:87).

1.1.3 Strategic risk management

Companies need to actively manage all types of risk, this is typically contained in an enterprise-wide risk management (ERM) framework (Sweeting, 2011:2). An effective ERM has several benefits including reduction in business cycle volatilities while targeting strategic risk adjusted returns (Sweeting, 2011:5), and can be used when introducing new

credit products, ensuring a profitable contribution to the company (Sweeting, 2011:5). Sweeting (2011:2) describes ERM within a given company as a process that understands the operating context, has the ability to identify associated risks, measures and compares contextual risks with its predetermined risk appetite, decides on appropriate management action, and regularly reports and assesses the effectiveness of management's actions. The central risk department, headed by the chief risk officer, traditionally manages this process (Sweeting, 2011:3). Credit providers utilise a similar approach with their own risk management frameworks. Moreover, they need to maintain strict risk appetite levels as it can significantly impact their risk and return strategy (Sweeting, 2011:3).

The role of credit application scorecards in the ERM is to leverage the ability of maintaining a desired risk appetite to ensure the success of strategic objectives of a credit provider. A second argument for segmentation is targeted risk management in the case of increasing credit limits (Seitshiro & Govender, 2024:31), where it is assumed that the credit consumer already has successfully passed the assessment criteria.

1.1.4 Loan application scorecards

Application scorecards are formal statistical tools used by retail credit providers to classify consumers into 'good' or 'bad' risk categories (Hand & Henley, 1997:523; Martin, 2013:6340). Scorecards assigns these two independent variables through predictions based on historical performances of the individual (Bijak & Thomas, 2012:2433). When the scorecard categorise customers as 'bad' status, this indicates a higher risk of loan default (Hand & Henley, 1997:523). Historically, the classification as 'bad' status was assessed by subjective human judgement and experiences (Myers & Forgy, 1963:800). However, in the retail industry where individual credit consumers are served, the human judgement approach is not cost-effective (Allen *et al.*, 2004:727). The solution is, therefore, process-driven credit scoring systems that can handle the large volumes of data and applications, while applying the credit risk predictions consistently (Abdou & Pointon, 2011:61).

Formal statistical scoring systems utilise independent and dependent variables. The independent variables consist of payment profile information that credit providers typically

have to reported to SACCRA. Important dependent variables have been identified as early as 1940 by Chapman *et al.*, (1940:110), where it was concluded that financial characteristics, past payment records, and demographic information had significant value when used in combination (Chapman *et al.*, 1940:138). Therefore, application scorecards are vital tools in the ERM framework of a credit providers as it allows for effective risk-based returns and profitability.

1.1.5 Population segmentation

Just as individual credit consumers differ in their credit risk, so can subpopulations also differ in risk outcome (Bijak & Thomas, 2012:2434). In credit scoring, segmentation is used to group consumers so that the subgroups are homogenous. Defined as "*a grouping of sufficiently homogenous exposures*" (Basel Committee on Banking Supervision, 2005:87), segmentation means that a large population can have distinct subpopulations where a single model may be less effective for the variability of the entire population (Siddiqi, 2017:103). As such, segmentation has been used to enhance the predictive performance of credit scorecards, as it allows for developing a focussed view on single subpopulations (Siddiqi, 2017:104). These subpopulations can be defined through factors such as consumer demographics, credit product type, employment type, the credit product portfolio, and the availability of data at the point of scoring (Siddiqi, 2017:105). Moreover, segmentation can be performed using either supervised or unsupervised techniques (Breed *et al.*, 2017:85). Despite the long history of models predicting credit default risk, Bijak & Thomas, (2012:2433) provide evidence that segmentation does not always provide better risk separation measured by the Gini coefficient.

Although the current research does not evaluate the segmentation methods, it focusses on comparing machine learning (ML) model's performance when segmenting the population by credit product types. While segmentation is widely applied to behaviour scorecards and loss prevention (Allen *et al.*, 2004:733), the focus of this study is to investigate its utility in the context of application models (Allen *et al.*, 2004:736).

1.2 Problem statement

The main argument for segmentation of subpopulations is to ensure more accurate risk assessment (Bijak & Thomas, 2012:2435). However, research within the South African retail credit industry remains sparse, specifically regarding the effects of segmentation across multiple credit products. While literature reviews such as Bhatore *et al.*, (2020:115, 120) affirm than Machine Learning (ML) is a viable tool for credit scoring, these studies focusses primarily on the banking sector and lack data from African sources (Bhatore *et al.*, 2020:130). Furthermore, studies on scorecard development incorporating segmentation success that conclusions are often case-specific and geographically limited (Aniceto *et al.*, 2020:13; Dastile *et al.*, 2020:18; Markov *et al.*, 2022:193).

Consequently, a gap exists in determining whether it is preferable to pool data into one generalised model or to maintain distinct credit product application scorecards. This uncertainty is added by the lack of direct studies examining the performance metrics between generalised customer application models and product specific scorecards for new loan originations in this region. This process is supported by a systematic literature review search, the criteria for this search is in Appendix A: Search criteria

Therefore, the research problem is to determine whether traditional product based segmentation strategies remain necessary to improve model prediction accuracy utilising advanced AutoML algorithms on South African retail credit data.

1.3 Research questions

To address the problem statement regarding the trade-off between data volume and population homogeneity, the following research questions (RQs) are posed:

RQ 1: To what extent does and AutoML framework offer superior predictive performance compared to individual AutoML framework models for retail credit scoring?

RQ2: Does product-based segmentation yield statistically superior predictive accuracy compared to the consolidated population baseline?

RQ3: Does a single generalised model, leveraging maximum data volume, outperform segment specific models when applied to distinct product sub-populations?

1.4 Research purpose

Blaikie & Priest (2019:21) refers to research objectives as research purposes. These purposes aim to minimise confusion by focusing on the type of knowledge desired while conducting specific research, which in this case, evaluated machine learning models across product segmentation in South Africa. Specifically, through using credit bureau data in determining if the segmentation approach on credit products results in increased prediction accuracy compared with not segmenting sub-populations.

To meet the primary research purpose, the following secondary purposes were established:

Research purpose (RP) 1: To evaluate the predictive efficacy of diverse machine learning architectures within the AutoGluon framework, identifying whether inherent techniques are fit for credit scoring.

RP2: To assess the necessity of product-based segmentation by testing whether homogeneous segment-specific models yield superior predictive accuracy compared to a consolidated population baseline.

RP3: To determine whether a single generalised model, leveraging maximal data volume, outperforms segment-specific models when applied to distinct sub-populations.

1.5 Scope and hypothesis

The study evaluated two primary hypotheses to address the research purposes, specifically testing the trade-off between population homogeneity (segmentation) and data volume (generalisation).

Regarding RP2 (Segmentation efficacy)

H_0 (Null): there is no statistically significant difference in the predictive performance (AUC) between segmented-specific ML models and a single generalised ML model.

H_1 (Alternative): Segment-specific ML models yield statistically superior predictive performance compared to a generalised ML model, indicating that distinct risk drivers exist across retail products the generalised model fails to capture.

Regarding RP3 (generalisation power)

H_0 (Null): The application of AutoML frameworks on a consolidated dataset does not yield a statistically significant improvement in predictive accuracy compared to segmented models.

H_1 (Alternative): The generalised ML model demonstrates statistically superior predictive performance compared to segment-specific models, indicating that the benefits of larger training datasets in AutoML outweigh the benefits of population segmentation.

Given that credit scorecards depend exclusively on supervised learning foundations (Dastile *et al.*, 2020:6), unsupervised, semi-supervised, and regression modelling were not included as part of the scope of this study.

1.5.1 Field of research

This study falls under corporate governance, technology, economics, and finances within the North-West University (NWU) Business School, with specific focus on applied science of risk management.

1.5.2 Economic sector under investigation

This study focussed on a retail credit provider with a direct-to-credit-consumer operating model through a network of stores consisting of several brands. The secondary data were sourced within the South African retail furniture industry, providing real-world context.

1.5.3 Geolocation

The secondary data utilised for this study were collected centrally from retail stores across South Africa. The data were enriched by a credit bureau service provider registered under the NCR.

1.5.4 Limitations

This study compared credit application scorecards built with ML frameworks based on a binary classification of ‘good’ or ‘bad’ customers. Due to the confidentiality agreement with the data provider, the exact criteria of ‘good’ and ‘bad’ credit consumers cannot be disclosed. It is acknowledged that other credit providers in the retail industry may use different definitions. In addition, for the sake of consistency and industry relevance, this study adopted the commonly used standard of evaluating customer performance after twelve months as determined by the systematic literature review. The sample data only contained a population who had active loans. It was considered to incorporate reject inference, but was decided against as it would have introduced synthetic data to the study, not critical in answering the research questions. Lastly, AutoML frameworks contain many parameters that are configurable, including hyperparameters and foundation model selections. However, finding the most optimal parameter configuration was out of scope for this study.

1.6 Research methodology

Business and management research philosophy is guided by the assumptions of ontology, epistemology, and axiology as outlined by Saunders *et al.* (2019:133). Research philosophy is defined as a system of beliefs and assumptions on how knowledge was, is, and going to be developed (Saunders *et al.*, 2019:130). Moreover, ontology is defined as if a belief is a true reflection of reality (Saunders *et al.*, 2019:133). For example, one could ask “what is the world like in 2025” (Saunders *et al.*, 2019:135). Epistemology has the viewpoint that relies on knowledge such as theories, proven methods, and even fictional literature (Saunders *et al.*, 2019:130). For example, one could

ask “what kinds of contribution to knowledge can still be made” (Saunders *et al.*, 2019:135). Lastly, axiology assumes the importance of values and ethics, meaning the ranking of choices based on personal values (Saunders *et al.*, 2019:134). For example, one could ask “how should we deal with the values of research participants” (Saunders *et al.*, 2019:135).

This study adopts the philosophical assumption of epistemology in its approach for solving the RQs, while the RPs are positioned to utilise statistical and data-driven techniques in the context of credit scoring. Furthermore, epistemology allows for both objective and subjective views to be incorporated into research methodology. In the current study, an objective stance is adopted to align with the empirical and quantitative methodologies required. As such, this study adopts a post-positivism philosophy. The rational for this choice lies in that the research will not rely on human assessment evaluating the 5Cs principle, which include character, capital, collateral, capacity and condition (Allen *et al.*, 2004:734). The study relies on statistical scientific principles grounded in the empirical observation of credit payment behaviour.

The field of natural sciences already incorporates established statistical principles, making inductive reasoning less appropriate for this study. A deductive approach was therefore appropriate to be applied on the research design, as it allows the testing of an existing theory through the method of hypothesis testing. Blaikie & Priest (2019:31) explains that deductive reasoning aims to answer the why-question. In the current study, the simplified why-question is why are businesses still using product segmentation in the time of advanced ML models? To be able to apply deductive reasoning here, Blaikie & Priest, (2019:102) provide the required steps as follow:

1. Start with the idea or hypothesis that shape a theory.
2. Find previously accepted academic studies that used a similar hypothesis in their testing, as well as the conditions and assumptions around it.
3. Understand the nuances of the conclusion and the reasoning supporting it.
4. Once the initial idea aligns with published reasoning, test previous conclusions by conducting an experiment.
5. Test the hypothesis using the appropriate measurement and determine if it shall be rejected or accepted, and

6. If the results are consistent with the previous conclusions, the theory is temporarily supported.

Based on these considerations, the current research makes use of a single method, quantitative design, evaluated on a cross-sectional dataset from the retail credit industry. The overall research methodology that depicts the philosophical stance, approach, strategy time horizon, and techniques is summarised in **Figure 1.4**.

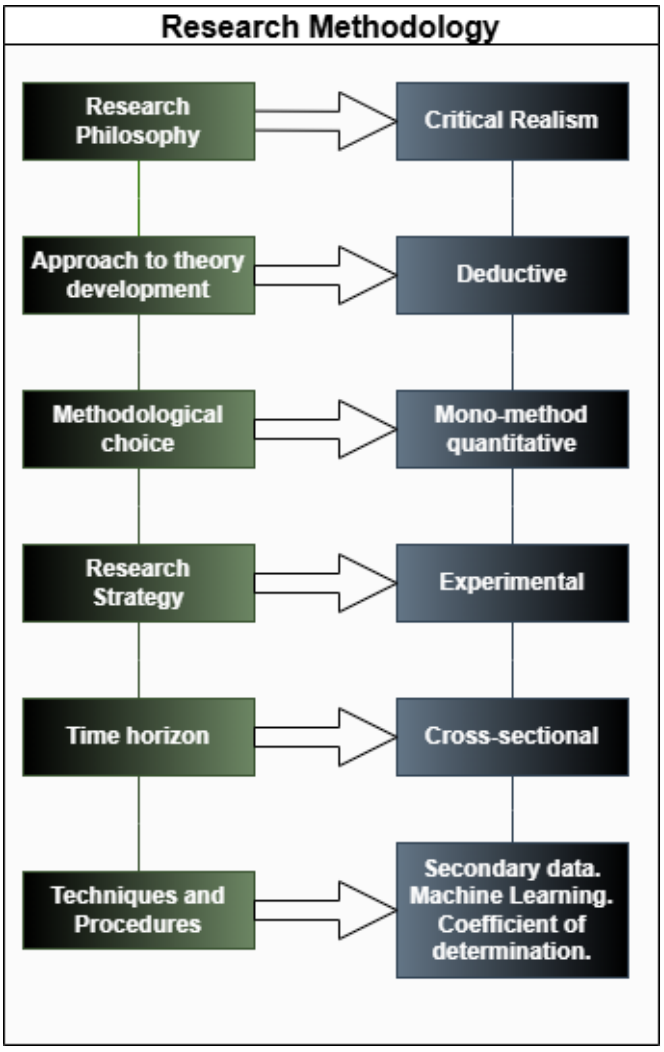

Figure 1.4: Research methodology (Source: Saunders et al., 2019)

1.7 Ethical considerations

Permission to access the secondary data used in this study was obtained from the designated gatekeeper of a single organisation. In addition to the gatekeeper application, a business case proposal containing the research value to the participating company was submitted. The participating organisation requested to remain anonymous. A non-circumvention, non-disclosure, and confidentially agreement was signed, which outlined conditions including, but not limited to, the following:

- The protection of confidential company information, including business strategies, products, services, trade secrets, customer data, internal systems, and operational practices.
- Strict adherence to the Protection of Personal Information Act 4 of 2013 (POPIA) regarding the processing and storage of personal or sensitive information of the customer.

In alignment with the protection of personal information relating to data subjects, the researcher was committed to full adherence with the provisions of the POPIA. Personal information was strictly utilised on a need-to-know basis during the data wrangling phase and for no other purpose. All data processing was conducted within the secure software infrastructure of the organisation.

Data integrity in the context of quantitative research covers a wide spectrum from data collection processing, storage, and aggregation for analytical insights (Cote, 2021). Cote further notes that data integrity is most likely compromised due to human error. For this research, data were processed and wrangled through data manipulation software, with the actions initiated through code. In adherence to high data quality standards, the researchers implement the following measures:

- Data was collected directly from the source databases.
- When data records from multiple sources were combined, they were matched using a unique identifier.

- The number of records at the start of data manipulation was verified against the final dataset to prevent accidental duplication or loss.

In compliance the North West University's research protocols, ethics clearance for the current research were obtained with reference number NWU-00699-25-A4.

1.8 Significance of the study

This study aimed to explore the relevance and impact of segmentation of credit products on credit application scorecards within the South African retail credit industry in 2025. The research findings may be valuable for companies using a generic industry score from credit bureaus. Moreover, they also may support companies considering the development of internal scoring models across multiple credit products or potentially offering a new credit product. Therefore, this study endeavoured to contribute to the retail credit industry knowledge base outside of the popular banking industry.

The research may also be valuable for credit providers with relatively small data science teams responsible for developing, maintaining, and deploying application scorecards. As Bijak & Thomas (2012:2442) noted, increasing the number of scorecards leads to an increase in the number of resources responsible for scorecards. Statistical model development is highly technical in its core, open source AutoML frameworks may allow for faster development times with higher accuracy.

1.9 Study layout

The remainder of this mini-dissertation is structured as follows:

- Chapter 2 contains a literature review of AutoML frameworks and how to assess credit application scorecards.
- Chapter 3 describes the technical methodology for the quantitative study.
- Chapter 4 contains the study results and interpretation.

- Chapter 5 concludes the mini-dissertation with additional future recommendations for further research.

CHAPTER 2 LITERATURE REVIEW

2.1 Introduction

This literature review will report on ML foundation models of the AutoGluon suite, a specific AutoML framework, in an effort to addressing RP1. To date, AutoGluon has consistently performed well amongst AutoML framework benchmarking by coming out as the top achiever in Kaggle competitions (Erickson *et al.*, 2020:8; Jurado *et al.*, 2025:3). This chapter further explores the strengths and weaknesses of various ML models before examining the concept of population segmentation.

2.2 Credit scorecard models

The search for indicators, or factors, that differentiate the various risks of credit consumers dates back to at least 1940, with early work published in a book by Chapman *et al.* (1940). At that time, the authors used the chi-square measure to assess the statistical difference between customers who have had a bad loan experience compared to those who have had a good loan experience. In their analysis, the authors used a sample of 1,294 records and determined that certain factors exist that are significant in separating good-loan from bad-loan experiences (Chapman *et al.*, 1940:137). Research have concluded that improving risk assessment is limited by utilising a low number of variables, typically in the range of 5 to 15 (Bijak & Thomas, 2012:2438; Scitovski & Šarlija, 2014:243).

Credit scoring is defined as using established quantitative statistical methods on historical data to classify credit consumers into ‘good’ or ‘bad’ categories of repayment behaviour (Hand & Henley, 1997; Montevechi *et al.*, 2024:1; Siddiqi, 2017:9). In order to achieve this, credit providers use various statistical models throughout the credit lifecycle of the consumer. Credit scorecards are further grouped under two categories; application scorecards for new loans and behaviour scorecards after a loan has been granted (Bijak & Thomas, 2012:2433; Hand & Henley, 1997:524). Application scorecards are trained on historical information to predict the future payment behaviour of new credit consumers (Bijak & Thomas, 2012:2433). Various types of models form the foundation of these

scorecards ranging from linear to more complex, non-linear models (Breed & Verster, 2017:2). In the context of the latter, a model refers to the statistical relationship between a dependent (i.e., 'good' or 'bad') variable together with one or more independent variables, as well as an error term (Banasik *et al.*, 2003:823). Behaviour scorecards are vital ongoing financial management tools (Bosker *et al.*, 2025:2) as the risk of credit consumers to default on repayment is always present, even after the new loan has been granted (Bosker *et al.*, 2025:3). More importantly, however, is that behaviour models are a regulatory requirement used to estimate the expected credit loss for companies (Basel Committee on Banking Supervision, 2005:48). As such, the expected credit loss is essentially an accounting realisation of the expected losses that could occur over a one-year period, which is the reported in the financial statements of a company (Basel Committee on Banking Supervision, 2005:82; Bosker *et al.*, 2025:3).

2.2.1 Linear models

Banasik *et al.* (2003:823) expresses a simple linear model as:

$$\boldsymbol{Y = \beta X + C + \epsilon} \qquad \textbf{Equation (1)}$$

Where:

- Y is the dependent variable,
- X represents one or more independent variables,
- β is the coefficient(s) of the independent variables,
- C is a constant, and
- ϵ is the error term normally distributed.

While linear models offer easier interpretability of results, available literature indicates that they are not appropriate for credit scoring. The prediction outcome is not constrained between 0 or 1, which is a key limitation as predictions in credit risk models make use of probabilistic outputs. Moreover, linear models are bound by their assumption that the

relationships between the dependant and independent variables are necessarily also linear. A comparison of arguments for compared with arguments against linear models are provided in **Table 2.1**.

Table 2.1: Advantages and disadvantages of linear models

Arguments for (i.e., pros)	Arguments against (i.e., cons)
• Conceptually easy to understand. • Results are transparently explained. • Simple to develop without specialised software. Does not require high computational resources.	• The dependent variable (i.e., results) has no boundaries. • Complex non-linear relationships cannot be rendered or calculated with these models.

(Sources: Breed & Verster, 2017:2; Hand & Henley, 1997:531; Montevechi *et al.*, 2024:14; Seitshiro & Govender, 2024:7; Zhang & Yu, 2024:8, 11.)

2.2.2 Logistic regression models

Logistic regression models are commonly used and highly researched in the practices of credit scoring (Allen *et al.*, 2004:736; Bijak & Thomas, 2012:859; Breed & Verster, 2017; Galindo & Tamayo, 2000:111; Hand & Henley, 1997:524; Hand & Kelly, 2001:989; Seitshiro & Govender, 2024:2). Although the logistic regression model is based on the linear model in **Equation (1**) the output is a non-linear transformation. Bijak & Thomas (2012:2435) expresses the model as:

$$Y = \mathbf{F}(\boldsymbol{\beta}\mathbf{X}) = \mathbf{1}/\left(\mathbf{1} + e^{-(\beta X + C + \epsilon)}\right) \qquad \textbf{Equation (2)}$$

The factors are the same as in **Equation (1**), with the addition of e being the base of the natural logarithm.

Logistic regression models are acceptable for credit scoring as they are well known and trusted by credit regulators. Moreover, the results they generate are transparent and easy to understand as the relationships between variables are linear. However, the assumption of linearity is restrictive, meaning that more complicated ML models may generally have higher predictive results. Another limitation is that logistic regression models struggle with high number of independent variables. A comparison of arguments for compared with arguments against logistic regression models are provided in **Table 2.2**.

Table 2.2: Advantages and disadvantages of logistic regressions models

Arguments for (i.e., pros)	Arguments against (i.e., cons)
• Results are easily interpretable and highly transparent. • Simple to develop. • Low computational resource intensity required. • Generally accepted by the industry and regulatory bodies for credit scoring.	• Assumptions of linearity are restrictive as they do not compensate for non-linear relationships. • Sensitive to multicollinearity. • Outperformed by more sophisticated ML algorithms. • Difficulty with large and high number of independent variables.

(Sources: Aniceto *et al.*, 2020:12; Bekhet & Eletter, 2014:23; Bhatore *et al.*, 2020:129; Dastile *et al.*, 2020:6, 11; Galindo & Tamayo, 2000:140; Halim & Humira, 2014:17; Hand & Henley, 1997:533; Kumar *et al.*, 2021:5; Montevechi *et al.*, 2024:1, 14, 16; Seitshiro & Govender, 2024:4.)

2.2.3 Non-linear models

Non-linear models, as the name suggest, do not assume a linear relationship between the dependent and independent variables of a function (Breed & Verster, 2017:1). As such, there is a growing interest in the use of non-linear application scorecards, particularly in the field of ML (Bhatore *et al.*, 2020:130). The AutoGluon framework that was used in the current study includes non-linear models such as random forests (RFs), extreme gradient boosting (XGBoost), CatBoost, light gradient boosting machine (LightGBM), neural networks, and *k*-nearest neighbours (KNN) (Erickson *et al.*, 2020:3). These models are explored in subsequent sub-sections in the context of credit scoring.

2.2.3.1 Random forests

The principle of RFs as a whole relies on a combination of decision trees, where each decision tree is derived from a random selection of variables and samples (Breiman, 2001:6). Each decision tree is independent from the previous variable, while maintaining the same distribution through allowing sample replacement (Breiman, 2001:6). Moreover, RFs utilise the concept of bootstrapping (i.e., bagging). Additional utilities and limitations of RFs are described in **Table 2.3**.

Table 2.3: Advantages and disadvantages of random forest models

Arguments for (i.e., pros)	Arguments against (i.e., cons)
• Often used in credit scoring. • Efficient with large datasets and large numbers of variables. • Improved predictive power compared with linear models with the ability to identify non-linear trends. • RFs do not use all the variables during training, resulting in uncorrelated trees and reductions in overfitting.	• Difficult to explain the decision or 'vote' for the most popular prediction result.

(Sources: Chopra & Bhilare, 2018:9; Dastile *et al.*, 2020:10; Mokheleli & Museba, 2023:508; Mushava & Murray, 2018:46; Rogojan *et al.*, 2023:1622; S. I. Serengil *et al.*, 2021:5; Zhang & Yu, 2024:12.)

RFs were suitable for the current research project, particularly due to their high popularity already seen in credit scoring practices, their efficiency with large datasets and numbers of variables, and their ability to identify non-linear trends and reduce overfitting by sub-setting variables. However, their complex, ensemble-based decision process makes reasoning the outcome challenging.

2.2.3.2 Extreme gradient boosting models

XGBoost is another scalable tree-based algorithm that have been reported to deliver state-of-the-art results (Chen & Guestrin, 2016:785). These models make up part of the gradient boosting decision tree (GBDT) framework of prediction models (Kumar *et al.*, 2021:12). XGBoost, as the name suggest, boosts model prediction by learning from the previous decision trees built in parallel (Dastile *et al.*, 2020:8). To date, the XGBoost model has already been employed in practice, leading to a reported 15% decrease in default rates (Chang *et al.*, 2024:30). Few systemic literature reviews exist where XGBoost is the subject of focus. However, the model may be applicable to credit risk scoring as it has increased predictive accuracy compared with logistic regression models and is efficient with computational resources when training. A comparison of arguments for compared with arguments against XGBoost models are provided in **Table 2.4**.

Table 2.4: Advantages and disadvantages of XGBoost models

Arguments for (i.e., pros)	Arguments against (i.e., cons)
• High predictive accuracy compared with logistic regression models. • Uses minimal computational resources and finds results faster across multiple factors. • Prevents overfitting. • Has the ability to handle missing values during training. • Can deal with imbalanced and/or sparse datasets. • Enabled graphics processing unit accelerated parallel computing. • Ability to predict credit risk.	• Difficulty explaining the predicted results. • Parameter tuning increase model complexity and computational resources. • Less used in credit scoring compared to logistic regression and RFs.

(Sources: Chang *et al.*, 2024:31, 2018:916, 919; Chen & Guestrin, 2016:785, 794; Dastile *et al.*, 2020:8; Mokheleli & Museba, 2023:504; S. I. Serengil *et al.*, 2021:129; Seitshiro & Govender, 2024:20; Zhang & Yu, 2024:12.)

2.2.3.3 CatBoost models

CatBoost models, which are also based on a GBDT framework, hold an advantage over models like XGBoost through its processing of categorical features (Prokhorenkova *et al.*, 2018:1). Categorical features, or variables, containing discreet information that are not comparable to each other (Prokhorenkova *et al.*, 2018:2). CatBoost models are trained using ordered boosting (Prokhorenkova *et al.*, 2018:1). The main difference between CatBoost and XGBoost, however, is that categorical variables are one-hot encoded in XGBoost (Prokhorenkova *et al.*, 2018:3). Moreover, CatBoost calculates a target score where, for example, in the current research a target score is the 'bad' rate. For a given categorical variable, the 'bad' rate is derived, and the categories are ordered by the 'bad' rate before the next tree is formed. A comparison of arguments for compared with arguments against CatBoost models are provided in **Table 2.5**.

Table 2.5: Advantages and disadvantages of CatBoost models

Arguments for (i.e., pros)	Arguments against (i.e., cons)
• Exceptionally high predictive power. • Uses less in-memory processing for large datasets. • Prevents target leakage through ordered boosting. • Minimal information loss when grouping categorical variables. • Prevents overfitting. • Applicable for credit scoring.	• Difficulty explaining the predicted results. • High complexity due to the boosting principle, where the next tree is dependable on the previous step.

(Sources: Chang *et al.*, 2024:7; Chopra & Bhilare, 2018:131; Dastile *et al.*, 2020:1; Mokheleli & Museba, 2023:500; Prokhorenkova *et al.*, 2018:1, 3; Zhang *et al.*, 2020:3, 13; Zhang & Yu, 2024:8, 12.)

CatBoost is applicable in credit risk scoring and can efficiently be trained on categorical variables with minimal loss of information loss.

2.2.3.4 Light gradient boosting machine

LightGBM is also part of the GBDT frameworks similar to XGBoost and CatBoost (Chang *et al.*, 2024:20; Ke *et al.*, 2017:2). However, it differentiates itself from XGBoost and CatBoost by growing depth-wise or leaf-wise (Chang *et al.*, 2024:20). The LightGBM model is applicable to credit scoring, while also being comparable to logistic regression and XGBoost outcomes in certain scenarios. As with CatBoost, its training algorithm also prevents overfitting. This model is sensitive to the initial hyperparameter selection and although the prediction accuracy is high, the interpretation of the prediction results may be complex to explain. A comparison of arguments for compared with arguments against LightGBM models are provided in **Table 2.6**.

Table 2.6: Advantages and disadvantages of LightGBM models

Arguments for (i.e., pros)	Arguments against (i.e., cons)
• Higher speed and calculation efficiency compared with XGBoost. • Optimal memory usage even with large datasets. • High predictive accuracy compared with logistic regression and XGBoost. • Effective in handling imbalanced datasets. • Guards against overfitting even with fewer independent variables. • Applicable for credit scoring.	• Difficult explaining the prediction results. • Hyperparameter sensitivity and tuning complexity is dependable on the initial parameters.

(Sources: Chang et al., 2024:5, 6; Galindo & Tamayo, 2000:109; Ke et al., 2017:8; Markov et al., 2022:180; Montevechi et al., 2024:14; S. I. Serengil et al., 2021:5; Zhang & Yu, 2024:18.)

2.2.3.5 Neural networks

Neural networks, also sometimes referred to as artificial neural networks or connectionist models, are another set of statistical prediction models (Abdou & Pointon, 2011:72; Hand & Henley, 1997:534) that take inspiration from the workings of the human brain by solving problems through trial and error (Abdou & Pointon, 2011:72). They consist of a set of interconnected processing units ordered in distinct layers (Rogojan *et al.*, 2023:1619) differentiated into of an input layer, several hidden layers, and an output layer (Chang *et al.*, 2024:18). Neural networks have been tested for credit scoring by Hand & Henley (1997:12). Some arguments for using them include the ability to learn complex non-linear relationships. Neural networks are usable on large datasets, and they have higher predictive capabilities over logistic regression (see **Table 2.7**). During the model's training process, overfitting should be assessed as this has been highlighted as the main counterargument for not using neural networks. Based on the literature, many complex models have difficulty in explaining its prediction results, with neural networks not being an exception (**Table 2.7**).

Table 2.7: Advantages and disadvantages of neural network models

Arguments for (i.e., pros)	Arguments against (i.e., cons)
• Ability to model complex and non-linear relationships. • High predictive accuracy over logistic regression. • Usable for large datasets. • Applicable for credit scoring. • Limited understanding of the data structure is needed.	• Difficulty in explaining the predicted results. • High demand on computational resources. • Sensitive to hyperparameter tuning through large numbers of layers and neurons. • Vulnerable to overfitting.

(Sources: Abdou & Pointon, 2011:69; Bhatore *et al.*, 2020:20; Chang *et al.*, 2024:6; Chen *et al.*, 2016:5, 6; Gasmi *et al.*, 2025:12; Hand & Henley, 1997:12, 13, 14; Kumar *et al.*, 2021:12; Mokheleli & Museba, 2023:499; Rogojan *et al.*, 2023:1620; Zhang & Yu, 2024:8.)

2.2.3.6 *k*-nearest neighbours model

This model is based on the principle of KNN and works through grouping data into similar clusters, with the clusters remaining homogeneous. The *k* in the name refers to the number of clusters that will be assessed before making a prediction, where a small number of clusters is more flexible but has increased prediction variances (Chen *et al.*, 2016:6; Rogojan *et al.*, 2023:1621). Moreover, this method retains training data, as the data serves as the model itself (Galindo & Tamayo, 2000:131). As such, KNN models are applicable for use in credit scoring. They use a distance measurement making predicted results problematic when variables have missing values. However, the model is robust against population drift specifically because of the distance measurement, while training is enhanced as only the important variables are present during training. KNN has a high demand on resources during training and is often less reliable. Additional utilities and limitations of KNN models are described in **Table 2.8**.

Table 2.8: Advantages and disadvantages of KNN models

Arguments for (i.e., pros)	Arguments against (i.e., cons)
• Non-parametric, the model does not make assumptions about the variables or data distribution. • Simple and interpretable as the nearest neighbour with the majority vote can be explained. This can imply that similar credit consumers behave in a similar way. • Evidence of robustness towards population drift. • Works well with low dimensionality (i.e., few independent variables). • Applicable for credit scoring.	• Vulnerable to overfitting where *k* is very large as the model does not predict well on unseen data. • Sensitive to imbalanced datasets where the 'good' and 'bad' populations are unequal. • Sensitive to irrelevant independent variables, by including features that add no predictive strength. • Missing values are problematic as the distance metric cannot be derived. • High demand on computational resources.

(Sources: Bhatore *et al.*, 2020:131; Dastile *et al.*, 2020:11; Erickson *et al.*, 2020:3; Galindo & Tamayo, 2000:131; Gasmi *et al.*, 2025:16; Hand & Henley, 1997:535, 536; Rogojan *et al.*, 2023:1621; Seitshiro & Govender, 2024:5.)

2.2.3.7 Ensemble models

Ensemble models utilise the prediction capabilities from a variety of models, for example linear models, logistic regression models, RFs, decision trees, LightGBM, XGBoost, CatBoost, neural networks, and KNN (Erickson *et al.*, 2020:4). Ensemble models have two benefits over individual models: they can combine a diverse number of individual prediction models (Chen *et al.*, 2016:8; Mian *et al.*, 2024:6) and they generally comparable to or outperform individual models (Chopra & Bhilare, 2018:132). Ensemble modelling also brings all the benefits of individual models together and will at minimum perform the same as the best individual model included in the group of models. By using this approach, it yields superior predictive accuracy and navigating complex data relationships in credit scoring. As with most ML models, its interpretability of the prediction results is difficult to explain. Additional advantages and disadvantages of ensemble models are provided in **Table 2.9**.

Table 2.9: Advantages and disadvantages of ensemble models

Arguments for (i.e., pros)	Arguments against (i.e., cons)
• Superior predictive accuracy and performance compared with individual models. • Effective in handling complex data relationships needed in credit scoring.	• The lack of interpretability from the single models are transferred over to the ensemble models and is the most significant drawback. • Requires large datasets for optimal performance that allows for stacking, weighting, and/or bagging techniques. • Need more than one base model.

(Sources: Aniceto *et al.*, 2020:4; Bhatore *et al.*, 2020:131; Chang *et al.*, 2024:8; Chopra & Bhilare, 2018:129; Dastile *et al.*, 2020:18; Erickson *et al.*, 2020:4; Galindo & Tamayo, 2000:115; Mian *et al.*, 2024:20; Mokheleli & Museba, 2023:497; Montevechi *et al.*, 2024:1, 7; Rogojan *et al.*, 2023:1622; Zhang & Yu, 2024:8.)

2.2.4 Model evaluation conclusion

The literature suggests that individual models are constantly improving by using popular boosting, bagging, and stacking techniques. Moreover, multiple individual models can be combined to form a single ensemble model with high predictive accuracy over traditional logistic regression models. Most of the ML models that have been listed in this section are tried and tested in the application of credit risk prediction. Linear models are deemed unusable, because its prediction results are not bound between 0 and 1. The reviewed literature confirmed the successful use of ML models, such as RFs, XGBoost, and neural networks in credit risk predictions. Therefore, the AutoGluon framework is appropriate for credit risk scoring, while the literature provided an affirmative answer to RQ1.

2.3 Scorecard performance assessments

Previously conducted systemic literature reviews show that many researchers of ML models measured performance with components of the confusion matrix and the area under the receiver operating characteristic curve (AUC-ROC) Abdou & Pointon, 2011:18; Montevechi *et al.*, 2024:13). The AUC is derived from plotting true positive rates and false positive rates (Montevechi *et al.*, 2024:13), it is a probability curve measuring the degree of separation (Chang *et al.*, 2024:26). Reasons for selecting AUC include, ranking of the predictions compared with the actual perdition values, and the utilisation in the absence of a classification threshold (Chang *et al.*, 2024:26). These popular performance measurements are particular to binary classification problems such as ‘good’ or ‘bad’ categorised credit consumers. However, prior studies have suggested not using the common accuracy, derived from the confusion matrix, in the context of credit scoring (Montevechi *et al.*, 2024:14; S. I. Serengil *et al.*, 2021:328). AutoGluon has the option to choose between a wide variety of evaluation metrics, all of which can be selected during the training process (Erickson *et al.*, 2020:2).

The AUC values ranges between 0 and 1, whereby a value of 0 represents a model with total incorrect predictions, a value of 0.5 a model cannot differentiate class predictions, and a value of 1 represents a perfectly predicting model (Chang *et al.*, 2024:26).

2.4 Conclusion

Literature suggests that the underlying models that make up the AutoGluon framework, an AutoML framework, is suitable for credit risk prediction. In terms of segmentation, there is a preference to segment data into subpopulations, but the results do not always support this. Regarding measuring ML model performance, the literature suggests using the AUC as derived from the confusion matrix. The next chapter will cover the technical research methodology in more detail.

CHAPTER 3 RESEARCH METHODOLOGY

3.1 Introduction

This chapter details the quantitative research methodology employed by the researcher to answer the remaining RPs. Following on from the literature review in the previous chapter, which established the theoretical foundation and relevance of ML models in credit scoring (answering RP1), this chapter outlines the practical and empirical framework of the study. Specifically, the methodology was designed to address the following:

RP2: Empirically develop and compare the predictive performance of models trained by the AutoGluon framework, in the context of segmentation by different credit product types.

RP3: Evaluate the performance of a single generalised model on the credit product segments.

This chapter will cover the research paradigm, the research approach and design, the dataset used, the detailed experimental procedure, and the specific evaluation metrics for assessing model performance.

3.2 Research paradigm and approach

The current research endeavour adopted a positivist research paradigm, which is concerned with observable, empirical, and measurable evidence (Saunders *et al.*, 2019:144). It involved a structured experiment to test hypotheses set out in the initial scope:

For **RP2** the null hypothesis assumed that segmentation on credit products did not improve the predictive separation between ‘good’ and ‘bad’ credit customers, measured by the AUC (i.e., a transformed version of the Gini coefficient). The alternative hypothesis assumed that segmentation did improve predictive separation.

For **RP3** the null hypothesis assumed that the generalised model did not perform differently on credit product segments. The alternative hypothesis is that there was a difference.

The research followed a deductive approach, where theories and propositions from a literature review were used to formulate a specific, testable experiment (Saunders *et al.*, 2019:144). The results of the experiment were then used to confirm or challenge these existing theories in the context of the AutoGluon framework.

3.3 Research design

A quantitative, experimental research design was employed for this study. The core of this design involves comparing a control group (i.e., a model trained on the entire dataset) with experimental groups (i.e., models trained on segmented data) to measure the differences or effects that the experimental group may exhibit compared with the control group. The research design did not include a reject inference.

Experiments were conducted in two main phases:

1. Generalised modelling phase: A single set of models was trained using the AutoGluon framework on a complete, unsegmented dataset. The performance and outcomes of the best model arising from the training served as the baseline. Importantly, the product classification variables were not allowed to be used in this process.
2. Segmented modelling phase: The dataset was subsequently divided into distinct subpopulations based on credit product type. A separate instance of AutoGluon was then trained on each segment.

The performance of the models used in each phase were systematically compared to determine the impact of segmentation using the AUC-ROC (Gijsbers *et al.*, 2024:16; Jurado *et al.*, 2025:4).

3.4 Techniques, and procedures

The technical methodology was developed by the researcher, drawing on more than a decade of data science experience in consultation with an industry expert with extensive experience in applied artificial intelligence (AI) and ML. The researcher has been actively involved in credit risk modelling in the banking and retail sectors. In addition, the researcher has experience in end-to-end model development, from data wrangling, model development, strategy design, and model implementation in the Cloud context. The researcher is familiar with best practices in the industry including application and behaviour modelling. The researcher is qualified in the field of quantitative risk management with an M.Sc. degree.

The current study utilised secondary data of a financial service provider in the retail credit industry of South Africa. The computational tasks were performed within AutoGluon, using Python as a programming language (**Figure 3.1**).

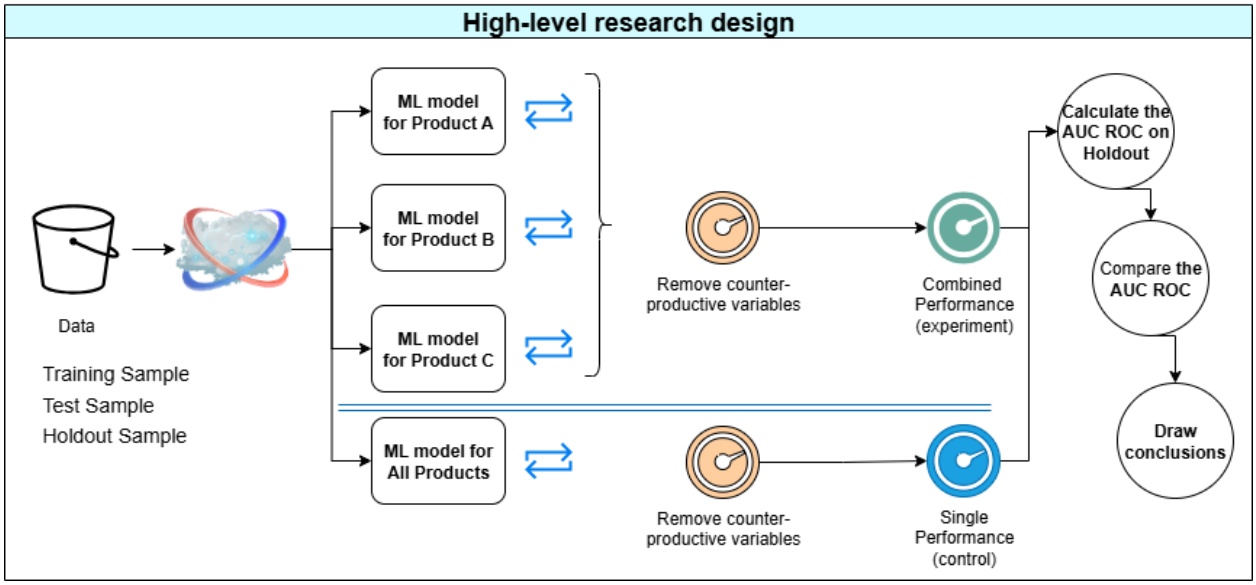

Figure 3.1: Research design overview

3.4.1 AutoML framework selection

This research made use of the open source AutoML framework, AutoGluon. AutoML frameworks have a wide area of applications, including image recognition, text prediction, regression and classification, etc. (Gijsbers *et al.*, 2024:1; Jurado *et al.*, 2025:1; Oliveira

et al., 2024:1). AutoGluon has consistently been acknowledged for reaching the highest ranks in benchmarking tests (Gijsbers *et al.*, 2024:31). In addition, AutoGluon-Tabular automatically solves for an end-to-end pipeline, from feature engineering to model deployments (Erickson *et al.*, 2020:8).

Breiman (2001:23–24) proposed a method to test variable importance so that when a variable is removed and the misclassification rate increases, the variable adds value to the model. Should the misclassification rate decrease, the variable is counterproductive or harmful to the model. Moreover, Fisher *et al.* (2019:8) also noted that removing variables with a negative contribution improves model prediction. In the current work, the researcher addressed the issues of these counterproductive variables. During the ML development process, the approach was to remove counterproductive variables with every model training iteration as follows:

1. Train ML models on 199 variables for a specific segment on ‘good quality’.
2. Test for variable importance and remove counterproductive variables.
3. Train ML models on reduced variables.
4. Repeat steps 2 and 3 until there are no more variables with negative importance.
5. Retrain the ML models on the remaining variables using ‘high quality’ presets.

The difference between a ‘good quality’ and a ‘high quality’ configuration was in their application. The former was for increased training time and ideal for prototyping, while the latter was designed for higher prediction results but took more time to train. The four models used for the hypothesis testing described in **Figure 3.** were the results of the ‘high quality’ presets from AutoGluon-Tabular.

3.4.2 Data structure and variable selection

Due to a confidentiality agreement entered into with the data provider, the bad rates were indexed at 100, which is the population bad rate. The cross-sectional sample covered the loans granted between the beginning of 2023 and the beginning of 2024. The target variable was the binary classification target of either ‘good’ or ‘bad’. For this research, all the credit products had the same target classification. **Table 3.1** describes the research

sample selection and key details of the master data. This structure ensured that there was no data leakage, limited overfitting, and accurate interpretation of the results.

Table 3.1: Secondary data breakdown by product segment

Models	Description	Bad index	Pop %	Counts	Random sample			Variables
					Training	Test	Holdout	External
Root	All Products	100%	100%	79 130	43 764	11 773	23 593	199
P_A	Product A	114.1%	6.9%	5 466	3 254	537	1 675	199
P_B	Product B	76.6%	14.0%	11 080	6 639	1 130	3 311	199
P_C	Product C	102.9%	79.1%	62 584	34 670	9 307	18 607	199

Product A was the smallest group at 6.9% of the total population but had the highest bad rate at 14.1% above the average. Product B was a small group at 14% and performed the best at 23.4% less risky than the average bad index. Product C was the dominant group making up 79.1% of the total population. Its bad rate performance was close to the overall average, having a 2.9% higher risk.

On average the training population (55.8%) was used to build the ML models, of which 14.4% of data was used to evaluate the model's performance during development and the remaining 29.8% of data was used for hypothesis testing. Only external variables of credit products from a credit bureau service provider in South Africa were utilised. The training, test, and holdout samples were determined through random uniform selection with no stratification. The external variables selected for the research were grouped as depicted in **Table 3.2**, where it showed a wide spread across various groups.

Table 3.2: External variable classification and count

Classification	Count
Long-term credit	42
Revolving credit	30
Unsecured lending	29
Customer aggregated	26
Retail	10
Personal loans	9

Credit card	9
Open accounts	8
Instalment	7
Banking	4
Furniture	4
Insurance	4
Telecoms	3
Retail apparel	3
Short term credit	2
Enquiries	2
Demographic	2
One-month loans	2
Short-term loans	1
Directorships	1
Geographic	1
Total	199

The master data covered a wide range of externally sourced independent variables, the largest categories were attributed to long term credit (n=42), revolving credit (n=30), and unsecured lending (n=29). It contained variations relating to a loan's account status, age, percentages and ratios, totals, and values.

3.4.3 Experimental procedure

The experiment was executed using Python and the AutoGluon software development kit. The following steps were taken to address the RPs.

Step 1: Data preparation

The dataset was loaded and randomly divided into training, testing, and holdout sets (**Table 3.1**). The test set was a requirement of AutoGluon to minimise overfitting. The holdout set remained unseen during training and was used for the final performance evaluation of the models during experimental Phases 1 and 2. Data processing, including missing value treatment, was handled automatically by the AutoGluon framework's internal pipeline.

Step 2: Generalised model and segmented model training (RP2)

The TabularPredictor.fit() in AutoGluon was initialised. With the following presets:

"framework_version"==1.3.0
"eval_metric": "roc_auc"
"train_data": X_train
"tuning_data": X_tune
"use_bag_holdout": True
"time_limit": 60*60*4 # (4 hours)
"presets": [high_quality, 'optimize_for_deployment']
"set_best_to_refit_full": True
"keep_only_best": True
"instance_type" = "ml.g5.4xlarge" # (GPU-based instance)
"instance_count" = 1

For the feature importance testing the presets were:

"data" = X_train
"num_shuffle_sets"= 5
"subsample_size"= 700

The 4-hour training time was consistent with benchmarking of AutoML frameworks (Gijsbers *et al.*, 2024:32). The research approach addressed the limitation mentioned by Gijsbers *et al.* (2024:32) and include graphics processing unit-based instances.

Step 3: Model evaluation (RP2)

Model evaluation during Step 3 occurred in two phases. For the generalised evaluation the best-performing model from the baseline training (Step 2) was evaluated on the full, unseen holdout set. Subsequently for segmented evaluation, the corresponding unseen holdout data for each product segment were passed to its specialised model. For example, the Product A holdout data were evaluated by the Product A model, and the same for Products B and C. All the data were combined and the AUC subsequently derived from the total holdout dataset.

Step 4: Model back testing (answering RP3)

Evaluated the generalised model on the product segmented data and calculated the AUC.

3.5 Conclusion

The quantitative research methodology was in accordance with best practices derived from past experiences of the researcher and in consultation with an industry expert in the field of AI and ML. Bureau variables from a wide range of categories were selected that represented all three credit products in the study. The data were trusted and the definition of the target (i.e., ‘good’ or ‘bad’) was consistent across the credit products. The next chapter contains the results from this research methodology.

CHAPTER 4 RESULTS

4.1 Introduction

This chapter presents the empirical findings obtained from the research conducted as described in Chapter 3. The results are structured to directly address RP2 and RP3. As such, this chapter begins by reporting the feature selection process before moving to the comparative analysis of the modelling approaches. The performance of the single generalised model is then reported, followed by an assessment of its generalisability across different product segments.

4.2 Feature selection and model optimisation

The iterative feature selection process was performed to remove counterproductive variables and optimise the final models as summarised in **Table 4.1**.

Table 4.1: Summary of variable reduction

Model	Removed variable count	Final variable count
Generalised model	-1	198
Model Product A	-64	135
Model Product B	0	199
Model Product C	0	199

Product A had the most features removed though the iterative process. Moreover, it consisted of the smallest segment of the population at 6.9% along with the highest bad rate at 14.1% above the total bad rate. Products B and C had no variables removed during this process. The model training results are reported in **Table 4.2**.

Table 4.2: Summary of model training results

Model	Model selection	Number of models	Training AUC	Test AUC
Generalised model	WeightedEnsemble_L3_FULL	76	0.9980	0.7852
Model Product A	WeightedEnsemble_L3_FULL	19	0.9924	0.7417
Model Product B	WeightedEnsemble_L2_FULL	9	0.9606	0.7836
Model Product C	WeightedEnsemble_L3_FULL	65	0.9800	0.7722

All models showed significantly high AUCs during the training process, which means the models had exceptional high predictive capability on the training sets. A high AUC meant that the models have learned the training data to such an extent that they could differentiate between ‘good’ and ‘bad’ credit risk nearly perfectly. It is not unexpected to see these results for the training set; however, conclusions should not be made on the training set but rather on unseen data represented by the holdout set.

4.3 Comparative performance: generalised versus segmented models

The AUC for the holdout sample was 0.7852 with the generalised model and 0.7718 with the segment specific model. The values depicted in **Figure 4.1** were the Gini coefficient values derived from the corresponding AUCs, which were 57.04% and 54.36% for each model, respectively. The Gini coefficient was 4.9% higher for the generalised model compared with the segment specific model. The difference in the AUC is on the mid- to upper range between the false positives and true positives. Thus the hypothesis test conclusion is that product segmentation did not improve model performance measured by the AUC.

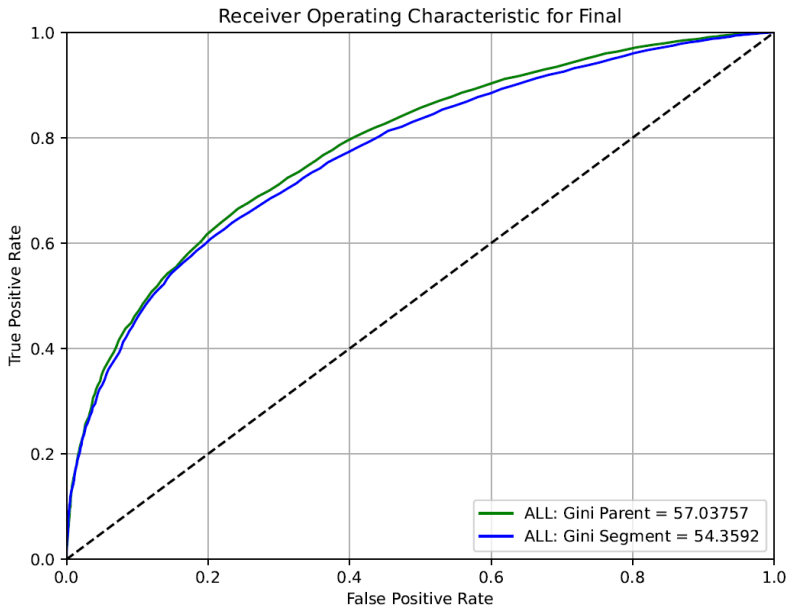

Figure 4.1: Model performance summary after variable reduction on the holdout sample

4.4 Generalisability of the single model across product segments

The main RP of this study was achieved through proving that predictive performance was not improved through product segmentation when using ML models. This section reports the results of the final RP, where the generalisability of the single model across product segments were tested (**Table 4.3**).

Table 4.3: Generalisability of the single model across product segments

Model	Holdout AUC of the segment model	Holdout AUC of the generalised model
Model Product A	0.7423	0.8206
Model Product B	0.7828	0.8157
Model Product C	0.7722	0.7766

Results reported from the analysis of generalisability clearly indicate that a single generalised model performs better than the models of the selected credit products of this

sample. The most notable performance improvements are seen on Product A and Product B. This suggests that the generalised model learns more robust patterns from the entire dataset that are applicable even to distinct subpopulations.

4.5 Conclusion

This chapter presented the empirical results of the mini-dissertation. The key findings show that a single generalised model, optimised through iterative feature selection, outperformed a segmented modelling approach. Specifically, the generalised model achieved a higher AUC of 0.7852 to 0.7718 of the segment specific model's on the holdout data (i.e., RP2). Finally, the generalised model proved to be highly robust, demonstrating superior performance on the total population and on each individual product segment when compared with specialised models (i.e., RP3). These findings lead into the final chapter, where their implications will be discussed.

CHAPTER 5 CONCLUSIONS AND RECOMMENDATIONS

5.1 Introduction

The final chapter will conclude with discussions of the implications of the findings from this study by connecting them back to the literature and the initial RPs. In addition, the chapter presents a series of recommendations for both industry practitioners and future academic researchers, while also acknowledging the limitations of the current study.

5.2 Summary of key findings

In the current study where different ML models were evaluated for credit risk prediction across retail segments of South Africa, a definitive answer was achieved that generalised models performed better in predicting credit risk than segmented models. Therefore, segmentation on credit products should not be perused. Firstly, the regulatory landscape of South Africa under the NCA with its enforcing body, the NCR, were discussed. The NCR comprehensively reports on the credit industry by subcategorising the industry every quarter, which revealed strong growth in the motor vehicle and retail industries over the 2023/2024 financial year. Then, following review of the literature, multiple credit products were observed to be standardised across different credit providers.

Under the NCA, credit providers are required to perform the necessary affordability and risk assessments prior to granting credit, whilst previously, consumers were not well protected. Currently, credit bureaus receive and maintain consumer payment behaviours, creating a credit consumer profile of historical payments. This information is utilised in the industry for assessing approvals for future credit agreements.

Credit providers have and maintain strategic ERM frameworks specific to their business model. Within the ERM, risk appetite levels limit exposure, which is determined by the chief risk officer. Limiting exposure levels result in stable returns during business cycles. For this reason, predictive models make part of assessment toolkit for risk appetite, called scorecards.

Loan application scorecards have been used for a long time as it is essentially predictive models derived from historical statistics. As time progressed, however, the use of scorecards increased in both use and complexity. The most basic and well-known model is the logistic regression model. Logistic regression models are limited by their intrinsic assumption of linearity. The limitations of linearity can be overcome by advanced ML models. Studies have suggested that segmenting the population into subcategories of a homogenous group could improve the overall predictive performance. In addition, the increase in predictive performance is often case specific or does not improve the model performance significantly.

ML models have been widely studied, although many on the same data repository from the University of California Irvine. One could argue that saturation has been reached on the data but it still serves the purpose of benchmarking different algorithms. ML models are complex to understand, that is why a variety of AutoML frameworks exist to make the underlying process of model development easier. One of the top performing frameworks, AutoGluon was selected for this study.

The primary purpose of this study was to determine if segmentation of credit products, as determined by the NCR, is the best approach for building scorecards. Credit providers often have unique scorecards for every credit product, resulting in a large business administrative process. Operationally, every model would have to be frequently re-assessed, monitored, validated, etc. Therefore, a single model would be beneficial in a business sense, if it could predict risk better.

The current research followed industry development standards, except for reject inference. Reject inference was deliberately excluded from the research as it would have introduced synthesised data to the study. The granted credit population in the industry could be as low 10%–20% of total applications. This means that the reject population would significantly outweigh the granted population (80% vs approximately 10%), notwithstanding the increase in data cost associated with this study.

On this backdrop, the study set out to develop four models: three models pertaining to three specific products and a fourth model for the entire population. During the training process all four models fitted the data exceptionally well with AUCs above 0.96, which translated to a Gini coefficient of 92%.

For conclusion of the current study, three findings, each relating to a specific RP, should be highlighted. The first finding was related to RP1 where the literature was reviewed to investigate which ML foundation models were appropriate for application scorecards. The literature review confirmed that ML models contained within the AutoGluon framework, including ensemble methods, have successfully been applied in credit risk prediction and are theoretically appropriate for credit scoring applications. The second finding was related to RP2 on developing and comparing the predictive performance of models trained by the in the context of segmenting by different credit product types compared with a generalised model. The primary empirical finding was that a single, generalised model had an AUC of 0.7852 (Gini coefficient of 57.04%) and outperformed the segmented modelling approach with an AUC of 0.7718 (Gini coefficient of 54.36%). The hypothesis that segmenting by credit product does not improve predictive performance was therefore confirmed. The third and final finding was related to RP3 where the performance of a single generalised model on the credit product segments was evaluated and the differences determined. The generalised model performed significantly better on segments for Product A and Product B. The null hypothesis was therefore rejected, meaning a single generalised model performs better on segmented individual models.

5.3 Business relevance

On the backdrop of these research findings, credit providers can confidently motivate for a single ML model to originate all credit loans. However, credit product managers would need to scale or align the generalised model output score according to their strategy and risk appetite mandate. In addition, credit providers using SAS software (from the SAS Institute), would need to utilise a micro-analytic service to invoke the AutoGluon ML framework model through an application programming interface. At the time of writing, it remains unknown to the researcher if AutoGluon can be used in conjuncture with a SAS Model Risk Management product.

Should the time arrive where incorporation of a single generalised model could be implemented into credit provider enterprises, the introduction should be phased to constantly assess real-world impact. This should be compared with the expected impact, on a 5%–10% sample allocation. A phased implementation plan will be needed to avoid

shocks and disruptions to normal business operations. When the comparative impact is then accepted by all product managers, the model percentage allocation to active use could be increased to 50% and then finally 100%. Subsequently, such a progressive implication will lead to the previous model being deprecated and fully replaced by the new model.

5.4 Future research

A literature review has highlighted the lack of diverse public datasets for credit risk modelling research, which limits future research and validation of current findings (Bhatore *et al.*, 2020:130). Public data from the University of California Irvine have featured in 77 peer reviewed papers, of which the most popular sets were from Australia, Germany, Japan, and Taiwan (Montevechi *et al.*, 2024:8). This research on ML models in the retail industry of South Africa adds a unique dimension on the outcome, although the underlying data are unfortunately not available for public use.

Reject inference is common in application credit risk prediction models. The current research did not include any reject interference, and it is recommended that future research should include it. However, careful planning would then be required to avoid target leakage. The magnitude of data required would also need substantial resource allocations. Therefore, it is proposed to explore several reject inference techniques. Licensed software like SAS Model Risk Management could streamline the reject inference process and control target leakage. The researcher did not have access to SAS Risk Modelling at the time of the study. However, there are acceptable open source AutoML frameworks available.

5.5 Conclusion

The study successfully answered all RQs aligned to the problem statement. From a quantitative modelling perspective it is not technically better to segment on credit products, in the retail credit sector of South Africa. It revealed that training ML models on more data resulted in greater risk separation, compared to training ML models on subpopulations.

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.

Allen, L., DeLong, G. & Saunders, A. 2004. Issues in the credit risk modeling of retail markets. *Retail Credit Risk Management and Measurement*. 28(4):727–752.

Aniceto, M.C., Barboza, F. & Kimura, H. 2020. Machine learning predictivity applied to consumer creditworthiness. *Future Business Journal*. 6(1):37.

Ashraf, N., Bau, N., Nunn, N. & Voena, A. 2020. Bride Price and Female Education. *Journal of Political Economy*. 128(2):591–641.

Banasik, J., Crook, J. & Thomas, L. 2003. Sample selection bias in credit scoring models. *Journal of the Operational Research Society*. 54(8):822–832.

Basel Committee on Banking Supervision. 2005. *International Convergence of Capital Measurement and Capital Standards: A Revised Framework*. Switzerland: Bank for International Settlements Press & Communications.

Bekhet, H.A. & Eletter, S.F.K. 2014. Credit risk assessment model for Jordanian commercial banks : neural scoring approach. *Review of Development Finance*. 4(1):20–28.

Bhatore, S., Mohan, L. & Reddy, Y.R. 2020. Machine learning techniques for credit risk evaluation: a systematic literature review. *Journal of Banking and Financial Technology*. 4(1):111–138.

Bijak, K. & Thomas, L.C. 2012. Does segmentation always improve model performance in credit scoring? *Expert Systems with Applications*. 39(3):2433–2442.

Blaikie, N. & Priest, J. 2019. *Designing Social Research : The Logic of Anticipation*. Newark, United Kingdom: Polity Press. http://ebookcentral.proquest.com/lib/northwu-ebooks/detail.action?docID=5638724.

Bosker, J., Gürtler, M. & Zöllner, M. 2025. Machine learning-based variable selection for clustered credit risk modeling. *Journal of Business Economics*. 95(4):617–652.

Breed, D.G. & Verster, T. 2017. The benefits of segmentation: Evidence from a South African bank and other studies. *South African Journal of Science*. 113(9/10):7.

Breed, D., Verster, T. & Terblanche, S. 2017. A semi-supervised segmentation algorithm as applied to k-means using information value. *ORiON*. 33(2):85.

Breiman, L. 2001. Random Forests. *Machine Learning*. 45(1):5–32.

Brigham, E.F., Ehrardt, M.C. & Fox, R. 2019. *Financial Management EMEA: Theory and Practice*. 2nd edn. Andover: Cengage.

Brits, R. 2018. The National Credit Act's remedies for reckless credit in the mortgage context. *Potchefstroom Electronic Law Journal*. 21:1–34.

Bryman, A., Bel, E., Hirschsohn, P., Dos Santos, A., Du Toit, J., … Wagner, C. 2021. *Research Methodology: Business and Management Contexts*. 2nd edn. Cape Town: Oxford.

Chang, V., Sivakulasingam, S., Wang, H., Wong, S.T., Ganatra, M.A. & Luo, J. 2024. Credit Risk Prediction Using Machine Learning and Deep Learning: A Study on Credit Card Customers. *Risks*. 12(11).

Chang, Y.-C., Chang, K.-H. & Wu, G.-J. 2018. Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Applied Soft Computing*. 73:914–920.

Chapman, J.M., Saulnier, R.J., Durand, D., Alexander, S.S., Davis, IsabelL., … Bettina, S. 1940. *Commercial Banks and Consumer Instalment Credit*. New York: National Bureau of Economic Research.

Chen, T. & Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. In: (KDD '16). *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery. pp. 785–794.

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.

Chopra, A. & Bhilare, P. 2018. Application of Ensemble Models in Credit Scoring Models. *Business Perspectives and Research*. 6(2):129–141.

Cote, C. 2021. *What is data integrity and why does it matter? Harvard Business School Online*. https://online.hbs.edu/blog/post/what-is-data-integrity Date of access: 13 Apr. 2025.

Dastile, X., Celik, T. & Potsane, M. 2020. Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing*. 91:106263.

Department of Trade and Industry (South Africa). 2015. National Credit Act, 2005 (Act No. 34 of 2005): review of limitations on fees and interest rates regulations. *Government Gazette*. 39379:107, 6 Nov.

Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., … Smola, A. 2020.

Fisher, A., Rudin, C. & Dominici, F. 2019. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research*. 20(177):1–81.

Galindo, J. & Tamayo, P. 2000. Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications. *Computational Economics*. 15(1/2):107–143.

Gasmi, I., Neji, S., Mansouri, N. & Soui, M. 2025. Bank credit risk prediction using machine learning model. *Neural Computing and Applications*. 1–18.

Gijsbers, P., Bueno, M.L.P., Coors, S., LeDell, E., Poirier, S., … Vanschoren, J. 2024. AMLB: an AutoML Benchmark. *Journal of Machine Learning Research*. 25(101):1–65.

Halim, S. & Humira, Y.V. 2014. Credit Scoring Modeling. *Jurnal Teknik Industri*. 16(1).

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.

Hand, D.J. & Kelly, M.G. 2001. Lookahead Scorecards for New Fixed Term Credit Products. *The Journal of the Operational Research Society*. 52(9):989–996.

James, D. 2012. Money-go-round: Personal Economics of Wealth, Aspiration and Indebtedness. *Journal of the International African Institute*. 82(1):20–40.

Jurado, I.C., Gijsbers, P. & Vanschoren, J. 2025.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., … Liu, T.-Y. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in neural information processing systems*. 30. https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

Kritzinger, N. & van Vuuren, G.W. 2021. Non-capital calibration of bureau scorecards. *Quarterly Review of Economics and Finance*. 79:260–271.

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

Markov, A., Seleznyova, Z. & Lapshin, V. 2022. Credit scoring methods: Latest trends and points to consider. *The Journal of Finance and Data Science*. 8:180–201.

Martin, N. 2013. Assessing scorecard performance: A literature review and classification. *Expert Systems with Applications*. 40(16):6340–6350.

Mian, Z., Deng, X., Dong, X., Tian, Y., Cao, T., … Jaber, T.A. 2024. A literature review of fault diagnosis based on ensemble learning. *Engineering Applications of Artificial Intelligence*. 127.

Mokheleli, T. & Museba, T. 2023. Machine Learning Approach for Credit Score Predictions. *Journal of Information Systems and Informatics*. 5(2):497–517.

Montevechi, A.A., Miranda, R. de 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*. 137:109082.

Mushava, J. & Murray, M. 2018. An experimental comparison of classification techniques in debt recoveries scoring: Evidence from South Africa's unsecured lending market. *Big Data Analytics for Business Intelligence*. 111:35–50.

Myers, J.H. & Forgy, E.W. 1963. The Development of Numerical Credit Evaluation Systems. *Journal of the American Statistical Association*. 58(303):799–806.

National Credit Act. 2005.

National Credit Regulator. 2024. *Amendment to the Guideline for the Submission of Credit Information in Terms Of Regulation 19(13) Of The National Credit Act, 34 Of 2005, as Amended*. https://sacrra.org.za/wp-content/uploads/2024/04/4_GUIDELINE-2-2024-MARCH-AMENDMENT.pdf Date of access: 21 Apr. 2025.

National Credit Regulator. 2025a. *Credit Bureau Monitor Report: Fourth Quarter December 2024*. https://www.ncr.org.za/documents/CBM/CBM%20Q4%202024.pdf Date of access: 24 July 2025.

National Credit Regulator. 2025b. *Consumer Credit Market Report: Fourth Quarter December 2024*. https://www.ncr.org.za/documents/CCMR/CCMR_Q4%202024.pdf Date of access: 24 July 2025.

Oliveira, S.D., Topsakal, O. & Toker, O. 2024. Benchmarking Automated Machine Learning (AutoML) Frameworks for Object Detection. *Information*. 15(1):63.

Petricek, T., van den Burg, G.J.J., Nazabal, A., Ceritli, T., Jimenez-Ruiz, E. & Williams, C.K.I. 2023. AI Assistants: A Framework for Semi-Automated Data Wrangling. *IEEE Transactions on Knowledge & Data Engineering*. 35(9):9295–9306.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V. & Gulin, A. 2018. CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*. 31.

Rogojan, L.C., Croicu, A.E. & Iancu, L.A. 2023. Modern Approaches in Credit Risk Modeling: A Literature Review. *Proceedings of the International Conference on Business Excellence*. 17(1):1617–1627.

S. I. Serengil, S. Imece, U. G. Tosun, E. B. Buyukbas, & B. Koroglu. 2021. A Comparative Study of Machine Learning Approaches for Non Performing Loan Prediction. In: *2021 6th International Conference on Computer Science and Engineering (UBMK)*. pp. 326–331.

Saunders, M., Lewis, P. & Thornhill, A. 2019. *Research Methods for Business Students*. 8th edn. Harlow, United Kindgom: Pearson Education Limited. http://ebookcentral.proquest.com/lib/northwu-ebooks/detail.action?docID=5774742.

Schmulow, A. 2017. Curbing Reckless and Predatory Lending: A Statutory Analysis of South Africa's National Credit Act. *Competition and Consumer Law Journal*. 24(3):220–247.

Seitshiro, M.B. & Govender, S. 2024. Credit risk prediction with and without weights of evidence using quantitative learning models. *Cogent Economics & Finance*. 12(1):2338971.

Siddiqi, N. 2017. *Intelligent Credit Scoring: Building and Implementing Better Credit Risk Scorecards*. John Wiley & Sons.

Singh, S.K., Tiwari, A.K. & Paliwal, H.K. 2023. A state-of-the-art review on the utilization of machine learning in nanofluids, solar energy generation, and the prognosis of solar power. *Engineering Analysis with Boundary Elements*. 155:62–86.

Siyongwana, P. 2004. Informal moneylenders in the Limpopo, Gauteng and Eastern Cape provinces of South Africa. *Development Southern Africa*. 21(5):851–866.

Stats SA. 2025a. *Quarterly Labour Force Survey: Quarter 4: 2024*. (Statistical Release P0211). Pretoria: Statistics South Africa. https://www.statssa.gov.za/publications/P0211/P02114thQuarter2024.pdf Date of access: 24 July 2025.

Stats SA. 2025b. *Retail Trade Sales: May 2025*. (Statistical Release P6242.1). Pretoria: Statistics South Africa. https://www.statssa.gov.za/publications/P62421/P62421May2025.pdf Date of access: 07 Aug. 2025.

Sweeting, P. 2011. *Financial enterprise risk management*. Reprinted with corr., 4. print edn. (International series on actuarial science). Cambridge: University Press.

Zhang, X. & Yu, L. 2024. Consumer credit risk assessment: A review from the state-of-the-art classification algorithms, data traits, and learning methods. *Expert Systems with Applications*. 237:121484.

Zhang, H., Zeng, R., Chen, L. & Zhang, S. 2020. Research on personal credit scoring model based on multi-source data. *Journal of Physics: Conference Series*. 1437(1):012053.

APPENDIX A: SEARCH CRITERIA

'ti' refers to the tile and 'kw' refers to keywords:

- ti:("Africa") AND ti:("literature" OR "review" OR "Study" OR "Research") AND kw:("Credit scoring" OR "Application scorecard" OR "Credit risk model" OR "Credit assessment" OR "Consumer credit risk" OR "Credit granting" OR "Retail credit risk")
- ti:("Africa") AND ti:("literature" OR "review") AND kw:(Credit) AND kw:("Machine learning" OR "Artificial intelligence" OR "ML" OR "AI" OR "Neural network" OR "Logistic regression" OR "Random forest" OR "Gradient boosting" OR "Support vector machine" OR "Deep learning" OR "XGBoost" OR "LightGBM")
- ti:("Africa") AND ti:("literature" OR "review") AND kw:(Credit) kw:("Segmentation" OR "Customer segmentation" OR "Market segmentation" OR "Cluster analysis" OR "Clustering" OR "K-means" OR "Latent class" OR "Classification" OR "Profiling")