Predicting business failure in a South African business bank

NL Motale

orcid.org 0000-0001-9707-824X

Dissertation submitted in fulfilment of the requirements for the degree

Masters of Science in Computer Science

at the North-West University

Supervisor: Prof PD Pretorius

Graduation ceremony: April 2019
Student number: 23276134
ABSTRACT

The aim of the study is to understand which factors cause business failure in a South African business bank and how can business banks successfully retain business banking customers with a probability of business failure by using a customer retention strategy and a predictive model. Business failure has been a topic for research projects across different industries such as hospitality, fishery, mining and mobile companies. Only a few studies have focused on business failure in South Africa relating to the failure of business banking customers and how can business banks effectively assist their customers by offering them services to help their business needs through the use of big data, customer value management, customer life cycle and analytical tools.

There have been discussions on how to use analytical tools and statistical methodologies to help predict and detect business failure across different industries in order to help retain businesses that show a probability of business failure. With the availability of big data and analytical tools, there is also the challenge of data quality, data integrity and data access. As business banks’ data generally are situated across different servers and warehouses, it requires the data to be merged from different warehouses and be put into a sensible format, which is a complex process.

A logistic regression model is used in the study to help predict business failure; it uses a methodology that has a dichotomous binary dependent variable that is recorded as either a zero or one, where one is true for business failure and zero is false for business failure.

In a South African business bank, whenever a business banking customer’s business fails, the loss of time, cost and effort in managing that customer is absorbed by the bank. This then affects the country’s GDP target and the National Development Plan, which is to develop entrepreneurs and to grow the economy. Business failure increases the unemployment rate of the country, as employees will be retrenched because the business would not be sustained.

Through a customer retention strategy, business banking customers will be provided with products that meet their business needs, be advised on their business’s financial
positioning and be given support through the bank’s entrepreneurship programme, which is generally given to business banking customers at no cost.

The study will show the effectiveness of the use of data analytics and statistical tools in solving banking problems and deriving solutions based on informed decisions through strategic data usage.

Contributions of the study are as follows:

- Some business failure factors could be determined using business banking customer data.
- A logistic regression model can be used to predict business failure.
- A customer retention strategy is proposed to help retain business customers that show signs of business failure.
- Text mining was used in the study to determine the industry the customer is in, as some of the business banks standard industry classification codes were incorrect, therefore, text mining was used to confirm the industry of the business customer using the business name.

**Key terms:** business bank, business failure, customer retention strategy, analytical tools
ACKNOWLEDGEMENTS

First of all, I would like to thank my Lord and Saviour, Jesus Christ, with whom all is possible.

To my dissertation supervisor, Professor Phillip Pretorius, of the Faculty of Natural and Agricultural Sciences at the North West University, thank you for your guidance and patience throughout my years of study. Whenever I had a question regarding my research, you were always willing to assist by steering me in the right direction and allowing this research to be my own work.

To my husband, Khothatso Motale, you have always provided me with unfailing support, understanding and continuous encouragement during my research; a lot of family time was sacrificed, but I have finally made it, the hard work has paid off. Thank you for your patience.

I must express my very profound gratitude to my parents, Modise Meshack Makale and Ditlhare Stephina Makale; you are my rock and foundation. To my little sister, Karabelo Makale, always dream big, anything is possible as long as you work hard and remain focused.

This achievement was feasible through your support.
# Table of contents

ABSTRACT ............................................................................................................................... ii

ACKNOWLEDGEMENTS ........................................................................................................ iv

LIST OF TABLES ...................................................................................................................... ix

LIST OF FIGURES ................................................................................................................... x

CHAPTER 1 ............................................................................................................................... 1

1. INTRODUCTION .................................................................................................................. 1

1.2 PROBLEM STATEMENT ..................................................................................................... 4

1.3 RESEARCH OBJECTIVES .................................................................................................. 5

1.3.1 Primary objectives ........................................................................................................ 5

1.3.2 Secondary objectives ................................................................................................... 5

1.3.3 Theoretical objectives .................................................................................................. 6

1.3.4 Empirical objectives .................................................................................................... 7

1.4 RESEARCH METHODOLOGY AND DESIGN .................................................................. 8

1.4.1 Literature review .......................................................................................................... 8

1.4.2 Empirical study ............................................................................................................ 8

1.4.2.1 Target population .................................................................................................... 8

1.4.2.2 Sampling frame ...................................................................................................... 8

1.4.2.3 Sampling methodology .......................................................................................... 9

1.4.2.4 Sample size ........................................................................................................... 9

1.4.2.5 Measuring instrument and data collection method .................................................... 9

1.4.3 Statistical analysis ....................................................................................................... 12

1.5 ETHICAL CONSIDERATIONS ......................................................................................... 12

1.6 LIMITATIONS OF THE STUDY ...................................................................................... 13

1.7 DEFINITION OF TERMS ................................................................................................. 13

1.8 CHAPTER CLASSIFICATION ............................................................................................. 14
CHAPTER 2: LITERATURE REVIEW ................................................................. 15

2.1 INTRODUCTION .................................................................................. 15

2.2 BUSINESS BANKING IN SOUTH AFRICA .............................................. 15

2.2.1 Introduction ....................................................................................... 15

2.2.2 South African overview of the banking sector .................................. 17

2.2.3 Banking market share ....................................................................... 18

2.2.4 Business banking segmentation ....................................................... 21

2.2.5 The role of the South African Reserve Bank (SARB) and its impact on banking systems ................................................................. 22

2.3 CUSTOMER RETENTION .................................................................... 23

2.3.1 The importance of customer retention .............................................. 25

2.3.2 Factors that influence customer retention ........................................ 26

2.4 BUSINESS FAILURE ......................................................................... 27

2.4.1 Causes of business failure ................................................................ 28

CHAPTER 3: RESEARCH METHODOLOGY AND DESIGN ............................. 30

3.1 INTRODUCTION .................................................................................. 30

3.2 RESEARCH PARADIGM AND METHODOLOGY .................................... 30

3.2.1 Research paradigm .......................................................................... 30

3.2.2 Statistical Software tools used .......................................................... 31

3.2.2.1 Microsoft Excel .......................................................................... 31

3.2.2.2 Microsoft Excel VBA ................................................................. 32

3.2.2.3 Microsoft Excel VBA benefits .................................................... 32

3.2.2.4 Microsoft Excel disadvantages .................................................. 33

3.2.2.5 SAS .......................................................................................... 33

3.2.2.6 SAS data access ......................................................................... 34

3.2.2.7 SAS software ............................................................................. 34

3.2.2.8 Enterprise miner package of SAS ............................................. 34

3.2.2.9 Enterprise miner package benefits ............................................. 35
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2.2.10 SAS enterprise miner more favourable over Microsoft Excel</td>
<td>36</td>
</tr>
<tr>
<td>3.3 STATISTICAL METHODOLOGY</td>
<td>36</td>
</tr>
<tr>
<td>3.3.1 SAS</td>
<td>36</td>
</tr>
<tr>
<td>3.3.2 SAMPLE</td>
<td>38</td>
</tr>
<tr>
<td>3.3.3 DATA COLLECTION AND DATA ANALYSIS</td>
<td>39</td>
</tr>
<tr>
<td>3.3.3.1 Data tables used</td>
<td>39</td>
</tr>
<tr>
<td>3.3.3.1.1 Demand deposit table</td>
<td>39</td>
</tr>
<tr>
<td>3.3.3.1.2 Customer information table</td>
<td>39</td>
</tr>
<tr>
<td>3.3.3.1.3 Reward point table</td>
<td>40</td>
</tr>
<tr>
<td>3.3.3.1.4 Product table</td>
<td>40</td>
</tr>
<tr>
<td>3.4 MODELLING ANALYSIS</td>
<td>40</td>
</tr>
<tr>
<td>3.5 WHICH MODEL WILL BE USED IN THE STUDY</td>
<td>43</td>
</tr>
<tr>
<td>3.5.1 Logistic regression model</td>
<td>43</td>
</tr>
<tr>
<td>3.5.2 Linear regression model</td>
<td>44</td>
</tr>
<tr>
<td>3.6 PURPOSE OF COMPARING TWO MODELS</td>
<td>46</td>
</tr>
<tr>
<td>CHAPTER 4: MODEL BUILDING, RESULTS AND FINDINGS</td>
<td>48</td>
</tr>
<tr>
<td>4.1 DATA SELECTION</td>
<td>48</td>
</tr>
<tr>
<td>4.2 DATA PARTITION</td>
<td>49</td>
</tr>
<tr>
<td>4.3 LINEAR &amp; LOGISTIC REGRESSION MODELLING</td>
<td>50</td>
</tr>
<tr>
<td>4.3.1 Interactive grouping for linear and logistic regression models</td>
<td>50</td>
</tr>
<tr>
<td>4.3.2 Gini coefficient for a linear regression and logistic regression model</td>
<td>51</td>
</tr>
<tr>
<td>4.3.3 Further analysis of the predictive variables using interactive grouping</td>
<td>54</td>
</tr>
<tr>
<td>4.3.4 Linear regression model statistics</td>
<td>57</td>
</tr>
<tr>
<td>4.3.5 Receiver operating characteristics chart for a linear regression model</td>
<td>61</td>
</tr>
<tr>
<td>4.3.6 Receiver operating characteristics chart for a logistic regression model</td>
<td>62</td>
</tr>
<tr>
<td>4.3.7 The cumulative lift chart</td>
<td>66</td>
</tr>
<tr>
<td>4.4 PILOT CAMPAIGN RESULTS</td>
<td>68</td>
</tr>
<tr>
<td>4.4.1 Campaign structure</td>
<td>69</td>
</tr>
</tbody>
</table>
4.4.2 Campaign messaging .............................................................. 70
4.4.3 Campaign results ................................................................. 71

CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS .........................77
5.1 INTRODUCTION ............................................................................ 77
5.2 CONCLUSIONS ........................................................................... 77
5.3 STUDY RECOMMENDATIONS ....................................................... 79

REFERENCES .................................................................................... 81
APPENDIX A ....................................................................................... 90
DECLARATION BY LANGUAGE EDITOR ........................................... 90
APPENDIX B ....................................................................................... 92
LIST OF SOUTH AFRICAN BANK INCLUDING MUTUAL BANKS, LOCALLY
CONTROLLED BANKS, FOREIGN CONTROLLED BANKS, BANKS IN
LIQUIDATION, BRANCHES OF FOREIGN BANKS AND FOREIGN BANK
REPRESENTATIVES ............................................................................. 92
APPENDIX C ....................................................................................... 96
PROC CORRELATION SIMPLE STATISTICS ...................................... 96
APPENDIX D ....................................................................................... 98
PEARSON CORRELATION COEFFICIENTS ..................................... 98
PROB > |R| UNDER H0: RHO=0 .......................................................... 98
NUMBER OF OBSERVATIONS ........................................................... 98
LIST OF TABLES

Table 1.1: Data collection and model building process ............................................. 11
Table 2.1: South African banking market 2016/17 ..................................................... 20
Table 3.1: DATA step logical flow .............................................................................. 37
Table 3.2: PROC SQL logical flow ............................................................................. 37
Table 3.3: PROC FREQ logical flow ......................................................................... 37
Table 4.1: Metadata summary ...................................................................................... 48
Table 4.2: Data partition summary .............................................................................. 49
Table 4.3: Summary statistics for class targets ............................................................ 50
Table 4.5: Interactive grouping on closing balance amount variable ....................... 55
Table 4.6: Interactive grouping on last month’s account risk category variable ....... 55
Table 4.7: Interactive grouping on account risk category variable ............................. 56
Table 4.8: Interactive grouping on yearly credit turnover variable ............................ 56
Table 4.9: Interactive grouping on customer total number of products variable ....... 57
Table 4.10: Proc correlation simple statistics .............................................................. 58
Table 4.11: Pearson correlation coefficients ............................................................... 59
Table 4.12: Linear regression Gini statistics of predictive variables ......................... 60
Table 4.13: Confusion matrix ....................................................................................... 64
LIST OF FIGURES

Figure 2.1: Number of banks in SA ................................................................. 18
Figure 2.2: Number of customers in each bank ............................................. 19
Figure 2.3: Market capitalisation comparison .................................................. 21
Figure 2.4: ABSA business banking segment .................................................. 22
Figure 2.5: Customer retention stages ............................................................ 25
Figure 3.2: Business customer’s type of business ............................................. 41
Figure 3.3: Yearly credit turnover of businesses that closed ............................ 41
Figure 3.4: Yearly credit turnover of business that closed ............................... 42
Figure 3.5: Business age analysis ................................................................. 43
Figure 3.6 Strong positive linear relationship between the $x$ and $y$ variable .......... 45
Figure 3.7 Strong negative relationship between two variables ....................... 45
Figure 3.8 A non-linear relationship among the $x$ and $y$ variable .................... 45
Figure 3.9 Between the $x$ and $y$ variable there exists a mix of positive and negative relationships due to wide variations among data points ......................... 46
Figure 4.1: Interactive grouping – event rate .................................................... 51
Figure 4.2: Information value interpretation ..................................................... 53
Figure 4.3: Linear regression ROC chart ......................................................... 62
Figure 4.4: Logistic regression ROC chart ....................................................... 63
Figure 4.5: Cumulative lift chart for training data and validation data ............... 67
Figure 4.6: Campaign structure ..................................................................... 69
Figure 4.7: The list of customers to be campaigned to using the initial SMS were distributed as follows: ........................................................................................................71

Figure 4.8: Holistic view of the initial SMS campaign ........................................72

Figure 4.9: Reminder SMS ...................................................................................73

Figure 4.10: Holistic view of the reminder SMS campaign .................................74

Figure 4.11: Follow-up phone call campaign .....................................................75
CHAPTER 1
Predicting business failure in a South African business bank

Keywords: business failure, business banking, attrition model, business banking customer

1. INTRODUCTION

The success or failure of a privately owned business contributes to the stability and growth of the economy of South Africa. According to Bhattacharjee *et al.* (2009), in South Africa about eight out of ten new businesses fail within three years of operating due to microeconomic and macroeconomic factors that affect businesses. Poor economic conditions and the business industry in which the business is conducted are common causes of business bankruptcies (*Avi-Yonah et al.*, 2017:15). Businesses that run the risk of bankruptcy may find an exit route, which will allow them to be acquired such that their assets may be redeployed by forming friendly merges that are not affected by distress.

The general idea of business failure is based on a lack of resources as it is often argued that a lack of resources affects a business’ growth. A general assumption would be that a business with more employees would succeed and a smaller business would not succeed due to lack of employees (*Watson*, 2006). According to Williams (2014), the environment in which a business is conducted makes the business vulnerable to digital change, economical shift and regulatory changes to name a few, which puts pressure on business management’s strategy, which, if not managed, could lead to business failure.

The global competitiveness index (GCI) is a report that takes into account all the fundamentals of an economy, such as current development, commodity price, the country’s currency and security issues. The GCI report is compiled by comparing 137 countries to one another in terms of their different fundamentals and each country is then given a rating.
Looking at the 2017/18 GCI report by the World Economic Forum (WEF), South Africa is seen to be one of the countries to look out for, especially when competing with other African countries as it is currently ranked 61 out of 137 countries. It dropped 14 positions from the overall rankings with respect to the following factors: institutional environment 76th out of 137, innovation 39th out of 137 and goods market efficiency 54th. South Africa is relatively good in the African market but it is rated weaker than last year. Some of the reasons that South Africa is rated weaker than last year are because of business factors such as bribery, criminal activity and a lack of stability in terms of the South African government (WEF 2017/18 Executive Opinion Survey).

The country has witnessed successful banks penetrating the market since 1991, which started with the Amalgamated Banks of South Africa (ABSA), through the merger of Volkskas, United, Sage, Allied and Bankorp (including Trustbank, Senbank and Bankfin) banks. In 2013, ABSA formed a merger with Barclays on 31 July 2013, where the group name changed from ABSA Group Limited to Barclays Africa Group Limited on 2 August 2013. In March 2018, the Barclays Africa Group announced that there will be a separation from the merger that was formed in 2013 and on 11 July 2018 ABSA Group Limited unveiled their new name, slogan and logo to the public. They also registered ABSA Group Limited on the Johannesburg Stock Exchange (JSE) as their new trading name (ABSA, 2018). Following ABSA in 1991 was Capitec, which was established in 2001 and listed on the JSE in 2002; Capitec is a bank that penetrated the banking industry successfully (Fin24, 2016). Prior to 1991, there were other banks such as uBank and Grindrod where customers were not depositing large amounts of money for safekeeping (Nhundu, 2016).

According to Fin24, banks such as Amalgamated Bank of South Africa (ABSA), First National Bank (FNB), Standard Bank, Nedbank and Investec collectively had a market share of 89.2 percent in 2017 compared to the 90.6 percent in 2015, while international banks grew to 7.3 percent from 5.3 percent, which was witnessed in 2014 (Fin24, 2017).

In the increasingly competitive business banking sector, banks are opposing high debt levels, competing with other banks and offering duplicate products to their business banking customers. South Africa has more than five banks that offer business banking
services to customers who either operate or manage a registered business and have a business bank account (Fin24, 2018).

Business banks can differentiate themselves through service fees, quality of service, incentives, customer value, business image, innovation and making the customer comfortable by offering them different servicing platforms where customers can transact, submit bank required documents and update their details such as branch, online banking, ATM, APP, cell phone banking as well as WhatsApp banking. WhatsApp banking is the first social media banking platform to be introduced in South Africa by ABSA Group Limited, previously launched in India in May 2018. WhatsApp banking platform allows ABSA Group Limited customers to check their bank balance, buy electricity, data, airtime as well as pay beneficiaries (BusinessTech, 2018).

According to the 2017 Global Bank Quality Benchmarking study by Lafferty Group, which is a study that gives 100 banks in 32 countries a rating out of five based on the banks’ sustainability, strategy, culture, management as well client services in both retail banking and business banking. South African banks got high ratings in 2016 and 2017 based on the information made available on public platforms about the South African banks and how they are viewed. Capitec was given a rating of five, while ABSA was awarded a rating of four and FNB, Standard Bank and Nedbank were rated a three. According to Fin24, the CEO of Lafferty Group, Michael Lafferty, mentioned that Capitec was given a rating of five as it is seen as a remarkable bank; it was one of a few banks that was rated high due to its financial ratios, culture, strategy, customer satisfaction and management. Capitec is seen as a customer-focused bank that focuses on ordinary South African customers and knows how to serve them well, which other banks should look at in order to move up the ratings (Fin24, 2017).

Banks play a critical role in economic growth through investment lending, offering loans, accepting deposits and ensuring that they adhere to the policies and procedures of the South African Reserve Bank (SARB), whose main function is to manage South African money and its banking system (Nhundu, 2016). It is crucial for banks to support businesses. When a business banking customer’s business fails, the loss of time, cost and effort in managing that customer is absorbed by the bank. This then affects the
country’s GDP target and the National Development Plan, which is to develop entrepreneurs and to grow the economy.

Retaining existing customers is more important and cost effective than to acquire new customers, as more costs are incurred at the beginning of the bank and customer relationship (Anderson et al., 2014; Gan et al., 2006). Product owners and bank management need to come up with new strategies to increase customer retention and to reduce customer attrition by understanding what the customer wants or expects from the bank and the factors that cause a customer to leave the bank, ending the bank and customer relationship (Symonds et al., 2007).

Very little focus has been placed on understanding and investigating factors that cause a business banking customer to switch banks by leaving their current bank (Satendra et al., 2012, Hamilton et al., 2017;). Most studies have only focused on retaining customers and a business customer’s happiness but not looking at them simultaneously by associating them to one another in the form of a customer retention model (Gan et al., 2006). When a customer retention strategy is not well maintained, no matter how long a customer banks with a specific bank, the customer can still get out of the relationship with the bank at any time, regardless of how hard bank management and the employees work (Williams, 2014).

The study will focus on the effect of business failure on South African business banks and see if a logistic regression model or a linear regression model can be used as a customer retention strategy for predicting business failure, in order to help the business bank to retain customers.

1.2 PROBLEM STATEMENT

For many years, the South African business banking sector has seen a high number of new to the bank, business banking customers closing their cheque accounts within three years of having their account and operating their business, which then lead to them ending their relationship with the bank, either due to lack of funding, business insight or support from various stakeholders.
When a business banking customer ends the bank and customer relationship, the question arises: Is it due to the business banking customer’s business failing or could it be because of a bank failing the customer by not understanding and catering to the customer’s needs?

Currently, there is no financial predictive model to predict the reasons why a customer would end their relationship with the bank before they actually end the relationship. Thus, the bank can determine, which customers are about to leave the bank through an attrition model, which is a model that predicts account closure, but not the reasons why the customer wants to end his/her relationship with the bank.

A business vulnerability index (BVI) model has been done by the Bureau of Market Research (BMR), which is a research unit in UNISA, regarding the vulnerability of South African sectors where they were testing what sector is vulnerable and the vulnerability reasons. According to Fedderke (2014), South African economies are said to be highly sensitive to global economic conditions as South Africa is heavily dependent on commodities and markets.

A business banking vulnerability index (BBVI) model will be built using business banking segment data, to better understand and measure the vulnerability of business banking customers’ businesses.

1.3 RESEARCH OBJECTIVES

1.3.1 Primary objectives

The objective of this study is to predict business failure in a South African business bank.

1.3.2 Secondary objectives

In order to address the primary objective, the secondary objectives below have been formulated:

- To do a literature review on the business banking sector.
• To understand the role and impact of the South African Reserve Bank (SARB) in banking systems.
• To understand customer retention and factors that influence it.
• To identify factors that predict business failure using a modelling methodology.
• Determine if business customers with a probability of business failure can be retained.

1.3.3 Theoretical objectives

The theoretical objectives are formulated to understand what is business failure, what causes business failure and the key variables that can be identified to help predict business failure through a modelling process.

According to Castano et al. (2017: 60), factors such as the business age, business size, the industry the business is in, business financial standing and the business risk could lead to business failure whilst the reasons thereof differ according to the researchers focus on the study. Williams (2014) suggests that the causes of business failure are associated with management failure, failed marketing strategies, failure in customer retention, failure to manage finance, systems and structural failure. Businesses are said to fail due to internal factors and not external factors as a result of lack of good management decisions according to organisation ecology scholars who study business failure.

**Business size** is said to be a contributing variable in terms of business failure, based on business turnover and the number of employees a business has (Bloodgood, Sapienza, & Almeida, 1996; Williams 2013, 2014).

**Business age** is seen as a predictive variable for business failure as it shows the experience and maturity of the business (Autio, Sapienza, & Almeida, 2000; Satendra et al., 2012, Williams, 2014;). Older businesses are said to have a better prospect of continued existence than smaller or younger businesses as they have more experience in their respective industries (Pretorius, 2009).
Industry plays a role in business failure, as the industry in which a business operates in plays a role in its ability to succeed (Campbell et al., 2012:90; Avi-Yonah et al., 2017:7). Access to resources is limited to specific sectors, the same way performance differs per sector (Avi-Yonah et al., 2017:12). The level of competition and influence of other factors in a sector determines whether a business will succeed in the sector or exit the sector. Businesses with greater resources will be more likely to survive and support smaller firms in the same industry, which will lead to other smaller firms surviving compared to sectors where there is a lack of support in terms of resources. However, there are cases where bigger businesses with more resources do not want to support small businesses to eliminate the competition in the same industry.

Financial resource is an important variable that is easily observed especially when it is time to release financial reports. A business’ financial standing plays a role in its ability to get credit, manage its funds, business credit score and turnover (Anani, 2010). When a business owner does not invest a large capital amount, it may indicate that the owner might want to take time to learn about operating the business instead of expanding the business immediately. Therefore, one can assume that investing less capital into a business could lead to business closure (Williams, 2014; Avi-Yonah et al., 2017:14).

Some of the variables above will be used in the study together with other financial variables that are relevant to the business bank, as the aim is to build a propensity model to predict business failure using financial data. A propensity model is a statistical model that is used to help predict the behaviour of a customer.

1.3.4 Empirical objectives

For the purpose of the study, a logistic regression model and a linear regression model will be built instead of compiling a survey study. The two models, namely logistic regression and linear regression, will be compared as follows:

- To test if a logistic regression model or a linear regression model could help predict business failure.
To use the best performing model (either a logistic regression model or a linear regression model) to predict business failure.

1.4 RESEARCH METHODOLOGY AND DESIGN

The study will include a review of the literature of an empirical basis. Modelling and quantitative techniques will be used to help identify factors that cause business banking customers to end their relationship with the bank and create a customer retention strategy to retain existing business banking customers.

1.4.1 Literature review

Reviewing of the literature will give more information into relevant discussions and the practicable factors that cause business failure in South African business banks. Relevant sources that are used in the study are:

- Dissertations
- Thesis
- Academic journals
- Articles
- Google Books
- Google Scholar

1.4.2 Empirical study

The empirical study focuses on the following dimensions: target population, sampling frame, sampling methodology, measuring instrument and data collection methodology.

1.4.2.1 Target population

The study will focus on all BankX business banking customers in South Africa with a turnover between R0 – R10 million.

1.4.2.2 Sampling frame

The study will focus on BankX business banking customers in South Africa with a business cheque account and a turnover between R0 – R10 million.
1.4.2.3 Sampling methodology

A non-probability convenience sampling method will be used, which focuses on business banking customers with a turnover of R0 – R10 million, that have a business cheque account. A non-probability convenience sampling method is a technique where businesses participating in the study will be chosen due to convenient accessibility and proximity to the researcher.

1.4.2.4 Sample size

The sample size is a significant part of the empirical study that will allow conclusions to be made based on evidence and reasoning about the population from a sample. The business banking customer population is around 582 000, of which a sample of the population is used.

1.4.2.5 Measuring instrument and data collection method

According to the primary objective, the experimental objectives below have been formulated:

The analytical framework

A logistic regression is a regression analysis method for analysing a dataset with one or more independent variables that determine an outcome (Rodriguez, 2007:3). The outcome is measured with a binary variable that should only be zero or one. The dependent variable is a binary variable and the independent variables are different variables of different formats such as ordinal, nominal or interval (Piech, 2016:15). The objective of a logistic regression model is to explain the relationship between the dependent binary variable and the independent predictive variables by creating coefficients of a formula to predict a logit transformation to get the probability of the presence of characteristic of interest such that the prepared model becomes:

\[
\text{Logit} \ (p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_6X_6 + b_7X_7 + \ldots + b_iX_i \quad (1)
\]

where:

\( p \) is the probability of presence of the characteristic of interest
\[ X_1 \text{ is variable 1} \]
\[ X_i \text{ is variable } i \]

The logit transformation is defined as the log odds, where

\[
\text{odds} = \frac{p}{1-p} = \frac{\text{probability of presence of characteristic}}{\text{probability of absence of characteristic}} \quad (2)
\]

and

\[
\text{logit}(p) = \ln \left( \frac{p}{1-p} \right) \quad (3)
\]

Variables are entered into the model in one of the three methods:

- **Forward stepwise**: allows significant variables to be entered in a sequence.
- **Backward stepwise**: enters all variables into the model and then removes the non-significant variables sequentially.
- **Bi-directional stepwise**: enters significant variables sequentially; after entering a variable in the model, checks and possibly remove variables that became non-significant.

A variable is entered into the model if its associated significance level is less than the P-value and it is removed from the model if its associated significance level is greater than the P-value.

**Odds ratios with 95 percent CI**

Piech (2016) notes that by taking the exponential of both sides of the regression equation as given above, the equation can be rewritten as:

\[
\text{odd} = \frac{p}{1-p} = e^{b_0} \times e^{b_1X_1} \times e^{b_2X_2} \times e^{b_3X_3} \times \ldots \times e^{b_iX_i} \quad (4)
\]

It is clear that when a variable \( X_i \) increases by one unit, with all other factors remaining unchanged, then the odds will increase by a factor \( e^{b_i} \)

\[
e^{b_1(1+X_1)} \div e^{b_1X_1} = e^{b_1(1+X_1) - b_1X_1} = e^{b_1 + b_1X_1 - b_1X_1} = e^{b_1} \quad . (5)
\]

This factor \( e^{b_1} \) is the odds ratio (O.R.) for the independent variable \( X_i \) and it gives the relative amount by which the odds of the outcome increase (O.R. greater than one).
or decrease (O.R. less than one) when the value of the independent variable is increased by one unit.

The prepared model will represent predictive variables of business failure.

**Table 1.1: Data collection and model building process**

![Diagram of data collection and model building process]

**Source:** SAS (2016)

*Theoretical objectives that will be answered:*

- Identify the total number of customers that ended their relationship with the bank and the period in which it took place.
- Identify the number of products a customer has with the bank.
- Determine if the business banking customer is profitable.
• Identify customers that will end their relationship with the bank because of business failure or banking failure.
• Determine which customer is more feasible to retain between a business failure customer and a banking failure customer.
• Identify factors that cause business failure.
• Create customer retention strategies for business banking customers who have been identified to end their relationship with the bank due to banking failure.
• Through a customer retention strategy, determine the number of customers that can be retained by ensuring that business banking customers do not end their relationship with the bank, by offering them the right products and support for their business.

1.4.3 Statistical analysis

The research methodology and the BBVI model building process will involve data collection and analysis using SAS, which is a statistical system that will allow results to be statistically interpreted and analysed.

The data will be manipulated, summarised, extracted and analysed using SAS procedures.

A DATA step process in a SAS language is used to create a SAS dataset view and read a dataset from internal and external data sources within the server. When the data are ready and accessible, a PROC step is used to produce a set of procedures such as tables, charts, reports and statistics using the data provided by the business bank.

SAS Enterprise Miner will be used to provide descriptive and predictive modelling insights that will drive and improve better decision making. This will help with designing the data mining process to develop models quickly.

1.5 ETHICAL CONSIDERATIONS

The following ethical considerations were adhered to during the conducting of this study:
• Getting permission to conduct the research study at BANKX
• Protecting the data provided for the research by BANKX
• Treating customer data with confidentiality.

1.6 LIMITATIONS OF THE STUDY

Focus will be placed on the business banking segment of BANKX. The study will also be limited to the financial data provided by BANKX that only relates to the businesses included in the study.

1.7 DEFINITION OF TERMS

ATTRITION MODEL: An attrition model is a model that predicts business banking cheque account closures.

CUSTOMER ATTRITION: This is the process of losing a business banking customer through account closure to another business bank.

BUSINESS BANKING: Business banks provide banking services to business customers across different banking platforms such as branches and online banking.

BUSINESS BANKING CUSTOMER: A business banking customer is defined as a customer who either operates or manages a business and has a business cheque account in South Africa.

VALUE ADDS: Value adds are products offered to business banking customers free of charge.

VERTICAL SALES INDEX: This shows the number of profitable products a customer has, which exclude products that are offered to business banking customers free of charge.

PRICEWATERHOUSECOOPERS: PwC, also known as PricewaterhouseCoopers, is one of the big four auditors along with KPMG, EY and Deloitte. They are based in multiple countries and their head office is in London, United Kingdom.
1.8 CHAPTER CLASSIFICATION

Chapter 1: Introduction and problem statement of the study

Chapter 1 will provide the introduction and background on the study including detailed descriptions of the objectives as well as the research design and methodology for the study.

Chapter 2: Literature review

The literature review gives insight into understanding what is business banking, how business banking is segmented, the role of the SARB, banking industry market share, customer retention and business failure.

Chapter 3: Research design and methodology

Chapter 3 will focus on the research design and methodology used in the study and the results obtained will be discussed in detail.

Chapter 4: Model building, results and findings

Chapter 4 will focus on the results found using the research design and methodology used in the study to obtain knowledge.

Chapter 5: Conclusions and recommendations

In Chapter 5, the results obtained will be evaluated to distinguish if the primary and theoretical objectives were met. Thereafter, recommendations will be given and the study concluded.
CHAPTER 2: LITERATURE REVIEW

2.1 INTRODUCTION

A bank offers financial services to any individual or group of individuals who own or manage a registered business such as a sole proprietorship, partnership, private company or public company.

This chapter will look into topics such as business banking in South Africa, banking market share, customer retention, business failure, banking sector overview and business banking segmentation.

2.2 BUSINESS BANKING IN SOUTH AFRICA

2.2.1 Introduction

Previously, banks were only seen as a place where money could be deposited for safe-keeping and a place where customers would go if they needed credit, which led to customers only seeking those two services from their respective banks (Banking Association South Africa, 2017). The first South African bank was the Lombaard Bank in Cape Town, which was established more than 200 years ago in the year 1793 (SARB, 2018). Before the 21st century, customers would get a service from the bank when they went to the bank branch and queued for the service.

However, with the introduction of the 21st century, came financial innovation where customer behaviours started to shift from the traditional way of banking, such as going to bank branches to do their transactions to the digital world were customers now have access to the Internet, cell phone banking, Apps and a designated banker to help them with their banking and queries. Digital banking platforms have enabled customers to do banking in the comfort of their own home; as a result, fewer customers are seen walking into the bank’s physical branches due to a change in customer behaviour (Nhundu, 2016).

According to PwC’s (18th Annual Global CEO Survey) the focus is currently placed on how business leaders can find ways in which they can compete in the ever changing
digital world by finding ways in which they can respond to the digital disruption (PwC, 2015).

The PwC survey focused on the five factors that business leaders should focus on:

- Finding new ways in which they can create value through the digital transformation, which leads to innovation.
- Vision and flexibility in thinking.
- Listening and learning to make clear informed decisions.

The PwC survey was conducted amongst 1400 CEOs in 77 countries, where they shared their views on the impact of growth, talent, trust and society.

In the ever-changing world, it is not just the economy that worries CEOs but the over regulation that spans across different industries such as tourism, manufacturing, mining, agriculture and communications to name a few. With the digital storm taking over, there have been a high number of cyber-attacks and threats to high profiled individuals on social media. The rapid pace of the digital world is seen as a challenge by 58 percent of CEOs, which highlights a shortage of key skills and growth. Eighty percent of the CEOs use data analytics and mobile technology as a strategy.

Major industry disruptions:

- Regulatory changes.
- Increasing competition.
- Behavioural patterns of customers.

Thirty percent of CEO’s say that most organisations will be forced to penetrate new industries every three years, to allow their organisations to remain competitive amongst others. In South Africa, most banks have entered into new industries by creating their own branded cellular phones and selling them to their current customers. They have their own cellular network through sim cards, moving more customers to digital platforms, offering insurance, offering convenience banking by giving their customers a team of bankers to do their banking via telephone and allowing users to get an electronically stamped bank statement on the APP without going to a branch (PwC, 2015).
2.2.2 South African overview of the banking sector

Out of 137 countries, South Africa is currently ranked 61 according to the Global Competitiveness Report (GCR) for 2017/18 and it is still viewed as the most competitive in Africa by being 39th in innovation. The GDP for 2017 was forecast at 1 percent whereas for the year 2018 it is forecast to be 1.2 percent. The GDP is affected by the low demand of commodities from international countries, a high unemployment rate that is currently estimated at 25 percent and lack of confidence of South African leaders caused by political uncertainty in 2017.

There have been quite a number of changes with respect to regulations, types of products banks now offer and the bank’s competitors due to new banks penetrating the sector, which causes greater competition amongst banks regardless of their size.

The South African banking industry consists of the following banks, according to the Banking Association South Africa and SARB (2018):

- 11 banks that are controlled locally.
- three joint banks (a joint bank is a credit union that has followed processes to get approval to use the name bank as part of its company name).
- seven foreign controlled banks.
- 14 foreign banks.
- two banks in liquidation.

Representative offices are created by a company to do marketing and not perform transactions normally in a country where a branch or a subsidiary of that company is not allowed. From the year 2006 to 2015, the number of representative offices in South Africa has decreased from 43 to 40 in 2015 and the number of international bank branches has remained constant at 15. The number of registered banks has remained constant between 17 and 20 from 2006 until 2015 as shown in Figure 2.1:
Figure 2.1: Number of banks in SA

*Includes banks that are active and banks that have been exempted by the Registrar of Banks with effect from 1 July 1990.

Source: SARB (2016)

2.2.3 Banking market share

Based on previous studies, there seems to be a correlation among profit and market share where they associated a greater return on investment to market share. Banks with greater revenue can build more branches across the country, which will lead to smaller banks actually closing down their business (Abir & Chokri, 2010:17; Kerai & Saleh, 2017).

Research conducted shows that a small number of customers would actually continue their relationship with their respective banks regardless of whether the bank now has a greater return on investment. Banks that penetrate the market with customer insight, large revenue and capital, tend to grow more than small, single market entrants (Borio et al., 2017).
The banking market share currently consists of FNB, ABSA, Nedbank, Capitec and Standard bank and they collectively hold a market share of 89.2 percent, which shows that the South African market is growing and experienced, even though the banking market share decreased slightly from 90.6 percent.

Figure 2.2 shows the number of customers each of the banks have; Standard bank has a customer base of 11.8 million as at December 2016, which is bigger than the other four banks and Capitec, which penetrated the market late and was seen as a small bank had a customer base of 8.3 million. When comparing all five banks, ABSA has been losing customers, dropping from 8.9 million in June 2016 to 8.7 million in June 2017, Nedbank has been consistent, while Standard Bank, Capitec and FNB have been increasing their customer base.

**Figure 2.2: Number of customers in each bank**

![Number of customers in each bank](image)

**Source:** BusinessTech (2017)
Market capitalisation and price/earnings ratio shows that FNB has the highest market capitalisation of R265 billion with Standard bank being second at R231 billion. FNB and Standard bank have a high market capitalisation compared to the other three banks, Capitec has a high price earnings ratio of 24.82, which is higher than the other four banks. The price earnings ratio of Capitec is high due to a combination of factors such as increase in sales from newly developed products, creating new trends, innovative solutions and cost management (BusinessTech, 2017).

This helps inform potential investors about their earnings based on historical data, the sector and sustainability. Price earnings ratios seem to remain consistent for banks that beat expectations. Banks that have a low price earnings ratio compared to Capitec could be impacted by increasing interest rates, increase in unemployment, customers spending less money, therefore, creating a low demand for banking products, customers not qualifying for credit and increased operating costs.

Table 2.1 shows the number of ATM’s and branches each of the top five banks has and their market capitalisation which is a company’s total value that is being traded on the stock market and is calculated by multiplying the total number of shares by the current share price.

### Table 2.1: South African banking market 2016/17

<table>
<thead>
<tr>
<th>Bank</th>
<th>Market capitalisation 2016</th>
<th>Price/earnings</th>
<th>Branches 2017</th>
<th>ATMs 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Bank</td>
<td>R231.36 billion</td>
<td>9.84</td>
<td>1211</td>
<td>7410</td>
</tr>
<tr>
<td>Absa Bank</td>
<td>R121.65 billion</td>
<td>8.07</td>
<td>774</td>
<td>8885</td>
</tr>
<tr>
<td>Nedbank</td>
<td>R103.86 billion</td>
<td>8.63</td>
<td>786</td>
<td>4052</td>
</tr>
<tr>
<td>Capitec</td>
<td>R94.16 billion</td>
<td>24.83</td>
<td>796</td>
<td>4024</td>
</tr>
<tr>
<td>FNB</td>
<td>R265.89 billion</td>
<td>11.15</td>
<td>676</td>
<td>4641</td>
</tr>
</tbody>
</table>

**Source:** BusinessTech (2017)
Network and reach

Standard bank has a big branch footprint of 1 211 branches in South Africa and ABSA has a big ATM footprint of 8 885 ATMs.

The market capitalisation comparison for 2016 and 2017

Market capitalisation, which is a company’s total value that is being traded on the stock market, is calculated by multiplying the total number of shares by the current share price. It shows that the value of FNB, ABSA, Nedbank and Standard Bank has decreased except for Capitec, which had a value of R65.17 billion in 2016, which increased to R94.16 billion in 2017 (BusinessTech, 2017).

Figure 2.3: Market capitalisation comparison

![Market capitalisation comparison graph](image-url)

Source: BusinessTech (2017)

2.2.4 Business banking segmentation

Business banking is segmented according to the business customer’s turnover. A product offering is then bundled, based on the customer’s business type, business needs and business size (ABSA, 2018).
Value adds refer to benefits such as online banking, in-contact, business legal advice, concierge service and free email statements (ABSA, 2018).

2.2.5 The role of the South African Reserve Bank (SARB) and its impact on banking systems

The role of SARB is to provide regulation and supervision as it regularly checks the efficiency and stability of important components of the South African financial system. SARBs financial system ensures that the financial system remains stable, which will lead to minimum cases of an intervention. It is key to have enough information about the behaviour of banks in the South African market, irrespective of institutional arrangements (SARB, 2018). The stability of the South African financial systems’ success is not done in isolation but through the participation of other financial
institutions, which require reliable and good information that provides insight about trends and developments regarding the financial system.

According to Banking Association South Africa, SARB supervises registered banks as they are allowed to take deposits from the public, therefore, the role of SARB is to ensure that public funds are not misused and to protect the depositors, should registered banks run into trouble of not being able to pay its depositors back.

Through financial surveillance and exchange controls, SARB is responsible for daily operations by ensuring that limits are put in place by Exchange Control Regulations (ECR) about the amount of money South African residents and registered companies can send abroad (SARB, 2018). National Payment System (NPS) ensures that payments between individuals and companies are recorded using an interbank settlement payment of transactions according to the central bank books. NPS creates efficiency by allowing parties to transact (SARB, 2018).

2.3 CUSTOMER RETENTION

Years ago, there was not enough attention being given to customers, which led to customers being neglected. Due to the lack of suppliers for customer goods and services, customers were unable to replace their current supplier. A shift was witnessed, as there was an increase in the number of suppliers in the same industry and competition started to grow, which led to the importance of retaining existing and new customers.

Jain et al. (2017) define customer retention as a measure that companies, businesses and organisations can take to reduce the number of customers that attrite by retaining as many customers as possible, whether new or existing, to ensure that they do not go to their competitors. However, most efforts are put on retaining existing customers as it is easier to adopt new retention strategies. It is more expensive to acquire new to bank customers, as more money is spent on acquiring new customers than retaining existing customers (De Meyer et al., 2010:28).

According to Satendra et al. (2012) for any business to succeed customer retention strategies need to be put in place and competitors have to be continuously innovative to attract new customers and retain existing customers. When customer loyalty
decreases and sales become less, customer retention remains an essential component, because if retention is not well managed the key customer will go to the competitor, which affects the businesses growth and profitability (Jain et al., 2017).

Customer relationship management is a process that uses customer information to understand customer behaviour by using the information to maintain relationships with the customer (Ascarza et al., 2017:349). CRM process involves data insight, market planning, improving customer interactions and analysis.

**Stage One:** Early stage customer retention is a time in months which the customer has had their business bank cheque account. It is generally from one day to twelve months. During this time, a customer has a small number of product holdings and possibly some of the products have not been activated (Aslam et al., 2014:55). Therefore, the focus of the stage is to ensure that products that the customer has are activated to avoid customer attrition. During this period, customer education takes place, where different products and uses thereof are explained to give them a better understanding of the product’s use and value (Haenlein et al., 2018).

**Stage Two:** Customer relationship management (CRM) is divided into two parts.

Part one is the proactive stage; the focus is to retain the existing customer by ensuring that triggers are put in place so that when a customer has an insufficient amount in their account a trigger message can be sent out to them to make them aware of the decrease in their account balance. Therefore, retention strategy would be to recommend product options that are suited to the customer based on the type of business they have and transactional behaviour by sending customer personalised communications in the form of an email, phone call or messaging.

Part two is the reactive customer retention stage, which applies when a customer is on the verge of ending their relationship with the bank whether due to high service fees, bank service or getting a low interest rate on their investment. Customer retention priority will be placed on profitable customers by retaining them by offering them loyalty reward points (Aslam et al., 2014:56). Reactive customer retention could
potentially be too late or too expensive, therefore, focusing on early customer retention stage and proactive customer retention stage is important.

**Figure 2.5: Customer retention stages**

![Customer retention stages diagram]

**Source:** Aslam *et al.* 2014

### 2.3.1 The importance of customer retention

Customer retention has many benefits and some of them are discussed below.

#### 2.3.1.1 Customer acquisition costs are high:

For a business bank to acquire new customers requires time and a strategy needs to be rolled out and an acquisition plan needs to be implemented to obtain customer transactional behaviour data so that an analysis can be done in order to get insight into customers who have had their business bank cheque accounts for less than three months (Satendra *et al.*, 2012). The customer acquisition process costs more money than retaining existing customers does; only if a customer has been with the organisation for some time can the acquisition process costs be recovered (Anani, 2010).
2.3.1.2 The business bank knows the existing customers: Through the use of analytics, organisations start to understand customer spending trends, price sensitivity, desires and expectations to name a few. This helps with marketing strategies (Williams, 2014).

2.3.1.3 Customer longevity equals profitability: Studies such as Aslam et al. (2014) have shown that the more customers a business can retain the more profitable a business can become. Only with a thorough understanding of customer retention factors can customer attrition decrease (Anvari & Amiem, 2010:17-18).

2.3.1.4 Innovation: Research studies by the GCR (2017), Ngonyama et al. (2018) and PwC shows that it is key for businesses to continuously evolve with their customers, to understand them better and to offer products specific to customer needs, as well as to remain relevant and to be able to differentiate themselves from their competitors (Kallmuenzer et al., 2018).

2.3.1.5 Customer lifetime value (CLV) calculations: A research study done by Panda (2006), Casteran et al. (2017:20) and Haenlein et al. (2018) shows that it is important for organisations to calculate a customer’s value to the organisation based on the customer’s current net worth to the organisation based on transactional behaviours and product holdings.

2.3.2 Factors that influence customer retention

One of the scarcest resources of an organisation, besides products and usage, is a customer (Peppers & Rogers, 2013).

2.3.2.1 Business size is said to be a contributing variable in terms of business failure based on business turnover and number of employees a business has (Bloodgood, Sapienza, & Almeida, 1996; Williams 2014).
2.3.2.2 Firm age is seen as a predictive variable for business failure variable as it shows the experience and maturity of the business (Autio, Sapienza, & Almeida, 2000, Williams, 2014). Older businesses are said to be in business longer due to their experience as opposed to less experienced businesses (Watson, 2006).

2.3.2.3 Industry sector plays a role in business failure as the industry in which a business operates plays a role in its ability to succeed. Access to resources remains limited to specific sectors, the same way performance differs per sector (Watson, 2006). The level of competition and influence of other factors in a sector determines whether a business will succeed in the sector or exit the sector. Businesses with greater resources will be more likely to survive and support smaller firms in the same industry, which will lead to other smaller firms surviving compared to sectors where there is a lack of support in terms of resources (Dias & Teixeira, 2014).

2.3.2.4 Financial resource is an important variable that is easily observed especially when it is time to release financial reports. A business’ financial standing plays a role in its ability to get credit, manage its funds, business credit score and turnover (Avi-Yonah et al., 2017:27).

2.4 BUSINESS FAILURE

Business failure is becoming a common factor in businesses as competition increases in the same industry and across industries. Understanding business failure grants a theoretical and practical challenge that is still to be met. The lack of understanding of business failure is mainly due to not having an adequate definition and identifying business failure predictors (Cybinski, 2001:39; Shepherd, 2005:126 & Castano 2018:71).

Through researching business failure, different authors have different views on the definition, effects and causes of business failure even though some share common factors in terms of causes and effects, even though there is no common definition of
business failure. There is a high business failure rate for new business but with every business that has created and started, business failure is the one thing that any new business owner does not want to think about. Given that business is seen as ‘survival of the fittest’ dependent on the industry and noticed gap in the market, business failure is seen as a natural step in a business cycle especially on a new venture (Castano, 2018:67).

2.4.1 Causes of business failure

2.4.1.1 Internal causes of business failure

Salman et al. (2015), in their research study, classify business failure into two categories, namely management-related issues and finance-related issues. In order to have a successful business there needs to be an ethical management system in place or the business could easily fail.

2.4.1.2 Management practices contributing to business failure

Adopting management strategy and practices could lead to business failure; when they are adopted from CEOs and then filtered down to analysts. There’s a probability that the management strategies and practices are not well communicated and understood, therefore could be implemented incorrectly.

Lukason & Hoffman’s (2015) study has voluntarism theories, which say that business failure is caused by voluntary decisions made by management that showcase the need to understand business and business failure. Making wrong decisions in a business leads to business failure (Dubrovski, 2009)

2.4.1.3 Comparing small and large firm’s business failure

According to Franco et al. (2009) and Castano et al. (2017:71), small businesses are said to be more flexible in terms of making decisions and making business contingency plans. This allows them to be able to call emergency meetings and make decisions quicker for the benefit of the business as there are fewer members involved in the decision making of the business. For example, most small businesses are mainly
managed by family members, an individual or a small group of owners, which makes it easier for small business owners to come together to make a decision regarding their business. Large businesses find difficulty in terms of making decisions regarding a business’ future prospects.

2.4.1.4 External causes of business failure

External causes usually occur due to factors that are not related to the business environment such the consumer confidence index and inflation rate (Dias & Texeira, 2014).
CHAPTER 3: RESEARCH METHODOLOGY AND DESIGN

3.1 INTRODUCTION
This chapter involves the data collection process, information about SAS, Microsoft Excel, including VBA and the statistical analysis. Tables and figures are used to present the data that was used to build models.

3.2 RESEARCH PARADIGM AND METHODOLOGY

3.2.1 Research paradigm

The research paradigm and methodology is explained below:

Ontology:
The study will focus on BankX business banking customers in South Africa with a business cheque account and a turnover between R0 – R10 million.

The sample size is a significant part of the empirical study that will allow conclusions to be made based on evidence and reasoning about the population from a sample. The business banking customer population is around 582 000, of which a sample of the population is used.

Epistemology:
According to Chau (1986: 612) a research study follows a research paradigm that is either positivist, interpretivist or critical. The context of the research study influences the paradigm choice including other factors in relation to the research problem, environment and the researcher (Trauth, 2009:2587). A choice is made to conduct the research using a positivistic research paradigm also known as a quantitative approach, a quantitative approach involves a statistical or numerical approach into the research design. According to Kumar (2014), research is specific in its experiment as it builds upon existing research and theories, it’s methodology is consistent with a researcher whereby the research remains independent of the researcher (Clark, 2003:211). As a result, the data is used to measure the probability of the objective taking place and the
research gains meaning through collecting data. It’s also used to ascertain the validity of the research and to show if the right research methods were used (Myers, 2013).

**Methodology:**

The purpose of every quantitative research requires understanding the problem statement, which involves the formation of the hypothesis, a literature review, data analysis and research methodology. Some quantitative approaches include strategies such as surveys and data collection on pre-established variables that will produce statistical data.

A quantitative approach is more suitable as it correlates to the problem of the study and the study objectives. Through a quantitative approach, predictive variables could be measured, which helps predict business failure in a South African business bank.

**Method:**

The research method requires the understanding and formulation of the problem so that the data preparation and collection process can begin. Permission must first be obtained in order to use the banks customer data for the modelling process, to ensure that customer information is treated with confidentiality and protection. After exploring the data provided by the business bank, data transformation and selection is done using the extraction rules to have a workable base for the model building process. A model will be built in SAS enterprise miner and get validated and deployed so that the model results will be monitored (SAS, 2016:4).

**3.2.2 Statistical Software tools used**

**3.2.2.1 Microsoft Excel**

For the purpose of this study, Microsoft Excel will be used. Microsoft Excel is a software programming tool that is found within the Microsoft Office suite of programmes, *inter alia* Microsoft Word that is used to type documents, Microsoft PowerPoint, which is used for presentations and Microsoft Outlook that is used to create and send out emails. Each programme in the Microsoft Office Suite is used according
to each user’s specific need (Weterings, 2017). Microsoft Excel is used to analyse data by creating different types of graphical views using the graphical tool, pivot tables and calculations and allows users to edit spreadsheets and to save an excel file in an extension that they prefer such as .xls or .csv.

An Excel spreadsheet has basic features that use a number of cells arranged in rows and columns, which helps with data manipulation. It allows users to hide rows or columns of data that they do not want to view. Microsoft excel is not designed to be used as a database as it can only handle limited data.

Excel provides some of the following security features such as password protection to modify document, to open a document, to protect a workbook, to unprotect a workbook and to protect the sharing of a workbook (Weterings, 2017).

3.2.2.2 Microsoft Excel VBA

Visual basic applications (VBA) is a computer programming language that allows the creation of user-defined functions and the automation of specific processes as well as calculations based on the user’s needs (Anon, 2017). VBA controls Microsoft Office products that are VBA compatible such as Microsoft Word, Microsoft Access and Microsoft Outlook, which come standard with VBA. In order for a VBA programme to work effectively, a user must understand the macro that they are creating and its functionality to help achieve a specific goal (Anon, 2017).

3.2.2.3 Microsoft Excel VBA benefits

Below is a list of some of the VBA benefits within Microsoft Excel:

- Allows a user to create a macro, which can help a user to automate tasks within Excel.
- Allows user to declare variables within a spreadsheet.
- Allows user to make if statements to create categorical variables.
- Arrays allow user to refer to an element within an array by using an array name and its index number.
3.2.2.4 Microsoft Excel disadvantages

Below are some of the disadvantages of using Microsoft Excel:

- By analysing spreadsheets with a large volume of raw data, important information could be hidden, making it difficult to analyse, find trends and interpret data, which could lead to incorrect conclusions and business decisions (Shankar, 2012).

- Spreadsheets do not function as a data warehouse meaning that they were not designed to store historical data. A business will not be able to store all their information in a spreadsheet because whenever a spreadsheet is updated, historical data might be lost, making it difficult to spot trends in historical data and compare it on a year-on-year basis (Denizon, 2012).

- The bigger the size of an excel spreadsheet the more difficult it is to share via email due to the size of the file (Weterings, 2017).

- Microsoft Excel cannot process two million rows in a sheet.

- Microsoft Excel cannot perform all statistical tests.

3.2.2.5 SAS

A SAS enterprise guide is a tool developed by the SAS Institute to perform multivariate analysis, business intelligence, predictive analytics and data management, to name a few (SAS Institute Inc., 2017). It is a tool that can be used in a research study to perform functions such as data extraction, data manipulation and data summaries.

SAS can be used to do the following:

- Access data in any format including Microsoft excel and SAS data tables.
- Manage data and manipulate existing data to get the data into the format that the user wants.
- Data analysis using statistical techniques such as correlations, descriptive statistics of the data.
- Allows the creation of reports in an understandable format that can be saved in different formats such as PDF (SAS Institute Inc., 2017).

By using the SAS tool, businesses will be able to increase their business functions by creating a data warehouse where all the data will be stored and can be retrieved if
needed by the business for data analysis such as customer trends and behaviours to better understand their customers.

3.2.2.6 SAS data access

SAS sequel programming language is used to perform tasks that consist of formats, macros, functions, tables and options (SAS Institute Inc., 2017). In this study, many DATA steps and PROC steps were used.

Below are some of the processes that can be done using a DATA step:
- Importing external files.
- Exporting files into Excel and SAS views.
- Creating SAS views and data sets.

When data becomes available, they can be used to do analysis and reporting to help provide insights into the study.

3.2.2.7 SAS software

The purpose of the study is to use SAS to help identify factors that can help predict business failure and to be able to measure the predictive power of every single factor. Through SAS, data can be analysed and controlled to help with reporting. The main function of SAS is to execute data steps and procedures.

3.2.2.8 Enterprise miner package of SAS

Enterprise miner package is used to build a predictive logistic regression model and a linear regression model to provide insight, draw relationships and find key trends that are important to the data mining process (SAS, 2016:4). The package helps map out the modelling process through an interactive flow diagram that helps in producing the best results. It allows results to be compared by displaying them next to one another.

It helps create correct predictions and descriptive modelling using big data (SAS Institute Inc., 2017). Data mining process is used by a large number of people across different industries to help with solving business problems and analysis of big data such
as customer retention, fraud detection, customer attrition, bankruptcy prediction, customer satisfaction and risk analysis to name a few.

The package follows the following steps for data mining:

- Sampling helps create one or more data sets, provided that there is significant data available to process the information. The process involves tools that help prepare the data such as the ability to import data, merge, filter, append and provide statistical sampling techniques.
- Explore is a node that helps to find trends, anomalies and relationships within the data in order to gain insight, knowledge and ideas. Explore provides a graphical view and variable selection methods.
- Model helps to train a statistical model such as a logistic regression, decision tree, linear regression, partial least squares and neural networks to help predict a desired outcome. The model node also allows for other models created by other users to be imported into enterprise miner.
- Assess evaluates the meaningfulness and trustworthiness of the output received from the data mining process. An assess node includes the model comparison node, which helps to compare different model performances and determines the best performing one by providing new fit statistics, cut-off point analysis and score code.

SAS enterprise miner package has a windowpane that helps with the data mining process by dragging and dropping nodes into the flow diagram. The package interface allows anyone with little statistical experience to be able to navigate through the data mining SEMMA methodology and someone who is a statistical expert will be able to explore each node in detail and be able to fine tune the analytical process (SAS, 2016:7).

3.2.2.9 Enterprise miner package benefits

- The package process helps with a large set of tools, which help the user by addressing complex problems, by taking a raw data and making it more insightful through its analysis tools.
• It allows users to be able to build models in less time by using the easy-to-use Graphical User Interface (GUI) through its common design principles. A GUI is an interface through which a user interacts with electronic devices.
• Gives a user the ability to compare results from different models and nodes.
• Helps drive insight in a self-efficient and automated manner. The SAS rapid model builder allows business analysts with limited statistical knowledge to build models automatically (SAS Institute Inc., 2017).

3.2.2.10 SAS enterprise miner more favourable over Microsoft Excel

Microsoft Excel has some statistical abilities which are less than SAS but depending on which version of Excel the user has for modelling purposes, Excel might require some add-ons and at times the statistics could be incorrect depending on the Microsoft Excel version (Shankar, 2012).

For as much as SAS enterprise miner provides users with the ability to drag and drop nodes onto the flow diagram, it is also imperative for users to comprehend the importance of data analysis and to be able to interpret statistical analysis; if it is not understood it could lead to errors and false hypotheses.

SAS is more favourable to use as it can analyse big data, read almost any data source and perform extractions while joining Microsoft Excel files to SAS data sets through an import statement. SAS gives the user the ability to create and join datasets with millions of rows, whereas Microsoft Excel is limited to just over one million rows and is not designed to do extreme analysis on big data. Through SAS Enterprise Miner and SAS Enterprise guide, a user has the ability to save all temporary datasets in a work library that stores all your datasets until the programme is closed (Shankar, 2012).

3.3 STATISTICAL METHODOLOGY

3.3.1 SAS

The use of SAS will allow the results to be statistically interpreted and analysed. The data will be manipulated, summarised, extracted and analysed using SAS procedures.
A DATA step in the SAS language is used to read a dataset from internal and external data sources within the server, create SAS datasets, perform views and data manipulation (Santana, 2009).

Table 3.1: DATA step logical flow

<table>
<thead>
<tr>
<th>DATA</th>
<th>SAS data set table rename</th>
</tr>
</thead>
<tbody>
<tr>
<td>SET</td>
<td>SAS data set library</td>
</tr>
<tr>
<td>WHERE</td>
<td>Expression is met</td>
</tr>
<tr>
<td>KEEP</td>
<td>List of variable(s) to keep in the data set</td>
</tr>
<tr>
<td>RUN</td>
<td>To execute the statement</td>
</tr>
</tbody>
</table>

When the data is ready and accessible, a PROC step is used to produce a set of procedures such as tables, charts, reports and statistics using the data provided by the business bank.

Table 3.2: PROC SQL logical flow

<table>
<thead>
<tr>
<th>PROC SQL CREATE TABLE</th>
<th>Allows you to create a table</th>
</tr>
</thead>
<tbody>
<tr>
<td>SELECT</td>
<td>Select column(s)</td>
</tr>
<tr>
<td>FROM</td>
<td>SAS data set library</td>
</tr>
<tr>
<td>WHERE</td>
<td>Expression is met</td>
</tr>
<tr>
<td>GROUP BY</td>
<td>Group by column(s)</td>
</tr>
<tr>
<td>QUIT</td>
<td>Execute statement</td>
</tr>
</tbody>
</table>

Table 3.3: PROC FREQ logical flow

<table>
<thead>
<tr>
<th>PROC FREQ</th>
<th>options</th>
</tr>
</thead>
<tbody>
<tr>
<td>BY</td>
<td>Calculates column frequency</td>
</tr>
<tr>
<td>EXACT</td>
<td>Tests specific statistics</td>
</tr>
<tr>
<td>OUTPUT</td>
<td>Creates an output table as per specifications</td>
</tr>
<tr>
<td>TABLES</td>
<td>Specifies frequency</td>
</tr>
<tr>
<td>TEST</td>
<td>Calculates measures of association and agreements</td>
</tr>
<tr>
<td>RUN</td>
<td>Execute statement</td>
</tr>
</tbody>
</table>
SAS Enterprise Miner will be used to provide descriptive and predictive modelling insights that will drive and improve better decision making. This will help with designing the data mining process to develop models quickly.

### 3.3.2 SAMPLE

The study will focus on business banking customers with a turnover of R0 – R10 million that have a business cheque account that meets the following extraction rules:

- Segment – Business banking
- Account status code
- Marketing Indicator – Yes
- Know Your Customer (KYC) – Yes
- Risk category = 0 and 1 (Customers who have low risk behaviour as determined by the bank will be included in the model as risk categories are from level 0 to level 5 and level 0 and 1 are seen as the lowest risk levels as opposed to level 2 upwards as they represent insolvency and fraud.)
- Account closure date
- Account age in months
- Last transaction date in months
- Business reward – Yes/ No

The sample of the study will consist of all business banking customers that meet the extraction rules, which will be used in building a predictive model. Population size is approximately 582 000 customers and after applying the extraction rules noted above the sample size becomes 116 786.

Through the use of the extraction rules above, customer behaviour and business activity can be tracked such that:

- Account status – informs the bank about the business cheque account status as to whether the cheque account is still active or inactive, whereby if its inactive it shows that the customer hasn’t been using the account for more than three months either by swiping, cash withdrawal or doing customer initiated transactions on the banks digital platforms.
• Last transaction date – shows the last time and date a transaction was done on the business cheque account.
• Account age – is the age of the business cheque account.
• Marketing consent – is important as customers have a right to subscribe or unsubscribe to marketing content and according to the Protection of Personal Information act, it’s illegal to market to customers without their consent and the bank can be fined if found in violation.

3.3.3 DATA COLLECTION AND DATA ANALYSIS
The data collection and analysis process of the modelling process involves data manipulation and data analysis of previously closed accounts for a period of five years, to provide an understanding of the number of accounts that are closed yearly and reasons behind why business customers would want to close their accounts.

3.3.3.1 Data tables used
The data used for analysis is sourced from different warehouses within BankX to get an understanding of customer behaviour before account closure. The following SAS data set tables were used to get the variables that are used in the models:

3.3.3.1.1 Demand deposit table
The demand deposit SAS tables contain monthly and daily transactions that business customers perform. The tables contain savings and cheque account information. Key variables that are found in the demand deposit data table are the product type, account status, business customer account number, account closing balance and some categorical variables such as the customer business type, business customer risk category and quarterly credit amounts, to name a few.

3.3.3.1.2 Customer information table
The customer information tables are divided into monthly and daily tables whereby daily tables are updated daily with new customer information and monthly data tables are only updated at the end of every month. Key variables that are found in a customer information table are customer name, customer surname, business telephone number, customer marketing consent, marketing type of communication they would like to receive if they consented such as an SMS, MMS, email or a phone call.
3.3.3.3 Reward point table
Reward point table shows the number of reward points and the reward level the business customer currently has together with the potential reward, had they swiped the cheque account or credit card account at certain retailers.

3.3.3.4 Product table
Shows the number of products and sub products a customer currently has from a loan account, investment account, vehicle finance account and home loan account etcetera.

3.4 MODELLING ANALYSIS
Past behaviour is a good predictor of future behaviour, so by analysing closed business banking cheque accounts data to build a model helps identify variables that can cause business failure. The graphical views below show accounts that were closed with the bank to help identify trends that will ensure that business banks put measures in places to retain profitable customers. Figure 3.1 shows the number of accounts that were closed over a five-year period and the account closure reasons show that most of the accounts were closed due to the automatic system closure, which takes place when a customer’s business cheque does not receive credit amounts into their accounts for more than six months of having the account. More than 110 943 customer business cheque accounts were closed due to the automatic system closure.

Figure 3.1: Business customer cheque account closure reasons

<table>
<thead>
<tr>
<th>Closure reasons</th>
<th>Number of account closures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Satisfied with Conduct</td>
<td>6409</td>
</tr>
<tr>
<td>Banks Charges Too High</td>
<td>6643</td>
</tr>
<tr>
<td>Business Closed/Sold</td>
<td>7137</td>
</tr>
<tr>
<td>Account Write Off &gt;R250.00</td>
<td>9017</td>
</tr>
<tr>
<td>Funds Required</td>
<td>9382</td>
</tr>
<tr>
<td>Dormant Account</td>
<td>91118</td>
</tr>
<tr>
<td>Consolidated Accounts</td>
<td></td>
</tr>
<tr>
<td>Business Closed/Sold</td>
<td></td>
</tr>
<tr>
<td>Consolidated Accounts</td>
<td></td>
</tr>
<tr>
<td>System Closure (No activity in 3 months)</td>
<td>110943</td>
</tr>
</tbody>
</table>
Figure 3.2 shows that most of the business cheque accounts that were closed belonged to customers who own private companies and close corporations.

**Figure 3.2: Business customer’s type of business**

<table>
<thead>
<tr>
<th>Business Types</th>
<th>Number of businesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stokvel</td>
<td>27</td>
</tr>
<tr>
<td>Public Company</td>
<td>35</td>
</tr>
<tr>
<td>Friendly Society</td>
<td>80</td>
</tr>
<tr>
<td>Foundation</td>
<td>107</td>
</tr>
<tr>
<td>Incorporated Company</td>
<td>332</td>
</tr>
<tr>
<td>Partnership</td>
<td>1594</td>
</tr>
<tr>
<td>Registered Trust</td>
<td>2581</td>
</tr>
<tr>
<td>Non Profit Company</td>
<td>3869</td>
</tr>
<tr>
<td>Other Association, Club etc.</td>
<td>4915</td>
</tr>
<tr>
<td>Sole Proprietorship</td>
<td></td>
</tr>
<tr>
<td>Close Corporation</td>
<td>20217</td>
</tr>
<tr>
<td>Private Company</td>
<td>30775</td>
</tr>
<tr>
<td>Foreign Company Registered in SA</td>
<td>47091</td>
</tr>
</tbody>
</table>

**Figure 3.3: Yearly credit turnover of businesses that closed**

<table>
<thead>
<tr>
<th>Yearly credit turnover</th>
<th>Number of businesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>R 100 000 &gt;</td>
<td>2260</td>
</tr>
<tr>
<td>R 40 001 - R100 000</td>
<td>1889</td>
</tr>
<tr>
<td>R 30 001 - R40 000</td>
<td>693</td>
</tr>
<tr>
<td>R 20 001 - R30 000</td>
<td>1067</td>
</tr>
<tr>
<td>R 10 001 - R20 000</td>
<td>2054</td>
</tr>
<tr>
<td>R 1001 - R10 000</td>
<td>17807</td>
</tr>
<tr>
<td>R 0 - R1 000</td>
<td>91016</td>
</tr>
</tbody>
</table>

Figure 3.3 above shows that most of the business cheque accounts that were closed had a yearly credit turnover between R0 – R1000, thus showing that the business wasn’t making a good turnover.
Figure 3.4 shows the number of products vs. Count the graphical view shows that customers who ended their relationship with BankX only had one product, which shows that customers with a low VSI have a high attrition rate, whereby customers who had more than four products did not close their accounts with the business bank.

Figure 3.4: Yearly credit turnover of business that closed

Figure 3.5 shows the business age (in months) of business banking customers businesses, most of the customers who ended their relationship with the bank have had their business for two to five years. This shows that there is a correlation between the business age and the account age, if a business has had a long relationship with a business bank. The more products the business has with the bank the more difficult it is for the business to end their relationship with the bank, as business customers with less product holdings within the bank ended their relationship with the bank early.
3.5 WHICH MODEL WILL BE USED IN THE STUDY

For the purpose of the research, two models will be used, namely a logistic regression model and a linear regression model, which will help in comparing the two models’ performance and see if the predictive variables from the two models remain the same or differ according to the variable statistics. This will in turn assist in selecting the best model to predict business failure in business banks. Below is more information regarding the two modelling techniques.

3.5.1 Logistic regression model

For the purpose of the study, a logistic regression model will be built as it has a binary outcome. A logistic regression model is a predictive model similar to an ordinary least squares (OLS) regression. By predicting a binary outcome, the approach poses a problem such that one assumes that residuals are normally distributed whereby they are in essence following a logistic approach. A linear regression equation written as:

\[ Y = \beta_0 + \beta_1 X + \varepsilon \]  

(1)

For logistic regression, variables that are not independent are said to be continuous and a logistic regression model predicts the probability of the \( Y \) variable being equal to one instead of it being equal to zero (Piech, 2016:14). When the \( X \) and \( Y \) variables have a strong relationship, the probability of a customer’s business failing is equal to one.
If $Y = 0$ is $1 - \hat{\rho}$ the equation is formulated as follows:

$$\ln\left(\frac{\hat{\rho}}{1-\hat{\rho}}\right) = \beta_0 + \beta_1 X \quad (2)$$

In the formula above (2), the $\ln$ function is acknowledged as a natural log and $\beta_0 + \beta_1 X$ signifies the regression line.

$$\hat{\rho} = \frac{\exp(\beta_0 + \beta_1 X)}{1+\exp(\beta_0 + \beta_1 X)} = \frac{e^{\beta_0 + \beta_1 X}}{1+e^{\beta_0 + \beta_1 X}} \quad (3)$$

If the regression equation is known, the value of $P$ can be formulated to calculate the expected probability for any $X$ value.

### 3.5.2 Linear regression model

For the study, a linear regression model will be built so that its performance and predictive power can be compared to the logistic regression model. The model that has better performance and accuracy will be used. The model results will be compared and discussed in Chapter 4 (Piech, 2016:20).

A linear regression model is a model with multiple independent variables, which are used to estimate the dependent variable; if we have $k$ variables that are not dependent on the dependent variable denoted by $x_1, x_2, \ldots, x_k$ the multiple regressions model will be denoted as

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon, \quad (1)$$

$y$ represents the dependent variable, which is also known as the response variable and $\beta_0, \beta_1, \ldots, \beta_k$ are unknown variables and $\varepsilon$ is known as an error term.

For example, if $k=1$ the notation will then be

$$y = \beta_0 + \beta_1 x_1 + \varepsilon \quad (2)$$

In a case where $k \geq 2$ dependent variables are plotted on the vertical axis and independent variables are plotted on the horizontal axis as indicated in figure 3.6:
Figure 3.6 Strong positive linear relationship between the $x$ and $y$ variable
Where the $x$ variable increases the $y$ variable also increases.

Figure 3.7 Strong negative relationship between two variables
Where the $x$ variable increases as the $y$ variable decreases.

Figure 3.8 A non-linear relationship among the $x$ and $y$ variable
Figure 3.9 Between the $x$ and $y$ variable there exists a mix of positive and negative relationships due to wide variations among data points

3.6 PURPOSE OF COMPARING TWO MODELS

It is better to compare two modelling methodologies to one another in a study instead of using one modelling method. By using two models in the study, it helps with comparing the two models and deciding on which model will be the best fit in predicting business failure in a business bank.

The model comparison process involves data testing, validation and the analysis of variable statistics such as the Gini, information value, simple statistics, Pearson correlation and interactive grouping to name a few. All these statistics play a role in deciding on the best model to avoid misleading and insufficient information.

The following rules must be met for the modelling methodologies:

- The variables in the data need to be compatible with the modelling method such that a logistic regression, neural networks, decision trees are meant to have binary classifications and a linear regression is meant to have a single classification.
- Statistics must be well interpreted and understood in order to get results.
- Model must be able to fit the sample data effectively.
- Get good prediction accuracy levels using data.
Based on the above, a quantitative approach is used in the study as it is more suitable because it correlates to the problem of the study and the study objectives.

The comparison of statistical software tools show that SAS enterprise miner and SAS enterprise guide are more beneficial tools to use for the study whereas Microsoft Excel spreadsheets do not function as a data warehouse meaning that they are not designed to store historical data. By using SAS enterprise miner and SAS enterprise guide, users can manage data and manipulate existing data to get the data into the format that the user wants. Analysing closed accounts data helps in finding trends and comparing two modelling methodologies to one another in a study instead of using one modelling method, helps with deciding which model will be the best fit in predicting business failure in a business bank.
CHAPTER 4: MODEL BUILDING, RESULTS AND FINDINGS

This chapter looks into the results found using the research methodology to obtain the required knowledge for the research. A linear regression model and a logistic regression model will also be built and compared in terms of their performance, accuracy and predictive power.

4.1 DATA SELECTION

The population of businesses with a turnover between R0 – R10 million is 582 000. After applying the extraction rules below, the modelling data set now contains a sample size of 116 786 observations with 60 variables, that is used in the modelling process for the time frame 01 January 2015 – 30 November 2015.:

- Segment – Business banking
- Account status code
- Marketing Indicator – Yes
- Know Your Customer (KYC) – Yes
- Risk category = 0 and 1 (Customers who have low risk behaviour as determined by the bank will be included in the model as risk categories are from level 0 to level 5 and level 0 and 1 are seen as the lowest risk levels as opposed to level 2 upwards as they represent insolvency and fraud.)
- Account closure date
- Account age in months
- Last transaction date in months
- Business reward – Yes/ No

Data fields in the data are categorised as follows:

Table 4.1: Metadata summary

<table>
<thead>
<tr>
<th>Role</th>
<th>Input</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>Interval</td>
<td>1</td>
</tr>
<tr>
<td>Input</td>
<td>Interval</td>
<td>34</td>
</tr>
</tbody>
</table>
Table 4.1 defines the variable types that are part of the sampling data set imported. Sampling data contains significant information and it is small enough to process for modelling. Through the sample data, analysis data preparation and simple statistics are done.

### 4.2 DATA PARTITION

The data partitioning process is used to assess the quality of the model from the data source. Data partitioning consists of a data set that is used for training where its primary function is the first step of model fitting, whereby the rest of the observations are kept aside for empirical validations.

The observations are split into two data sets where one is for validating the data and the other is for testing the data; whereby, data validation helps prevent the model from overfitting the training data, which will help with comparing the logistic regression and linear regression model against one another. The testing data are used to test the model performance.

**Table 4.2: Data partition summary**

<table>
<thead>
<tr>
<th>Data</th>
<th>116 786</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training data set</td>
<td>81 748</td>
</tr>
<tr>
<td>Validate data set</td>
<td>35 038</td>
</tr>
</tbody>
</table>

The partition is split 70 percent training data set and 30 percent validation data set.
Table 4.3: Summary statistics for class targets

<table>
<thead>
<tr>
<th>Variable</th>
<th>Numeric value</th>
<th>Formatted value</th>
<th>Frequency count</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target</td>
<td>0</td>
<td>0</td>
<td>99 204</td>
<td>84.9451</td>
</tr>
<tr>
<td>Target</td>
<td>1</td>
<td>1</td>
<td>17 582</td>
<td>15.0549</td>
</tr>
</tbody>
</table>

Data = TRAIN

<table>
<thead>
<tr>
<th>Variable</th>
<th>Numeric value</th>
<th>Formatted value</th>
<th>Frequency count</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target</td>
<td>0</td>
<td>0</td>
<td>69 442</td>
<td>84.9464</td>
</tr>
<tr>
<td>Target</td>
<td>1</td>
<td>1</td>
<td>12 306</td>
<td>15.0536</td>
</tr>
</tbody>
</table>

Data = VALIDATE

<table>
<thead>
<tr>
<th>Variable</th>
<th>Numeric value</th>
<th>Formatted value</th>
<th>Frequency count</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target</td>
<td>0</td>
<td>0</td>
<td>29 762</td>
<td>84.9421</td>
</tr>
<tr>
<td>Target</td>
<td>1</td>
<td>1</td>
<td>5 276</td>
<td>15.0579</td>
</tr>
</tbody>
</table>

4.3 LINEAR & LOGISTIC REGRESSION MODELLING

The sections below follow the process of building the linear regression model and the logistic regression model where their statistics, performance and accuracies will be compared.

4.3.1 Interactive grouping for linear and logistic regression models

Interactive grouping is a different method used in the study to help process data before starting the modelling process. By grouping initial bins into quantiles and thus allowing the interactively splitting and combining of the initial bins, which helps to improve the variables predictive power by means of generated weight of evidence (WOE) for each of the groupings of the variable. WOE is defined as a logarithm; a fraction of the non-event counts of the observations in the interactive group over the proportion of the event count of the observations. The relative risk of any interactive grouping is measured by the WOE. Interactive grouping allows users to choose variables with better quality and prediction based on the Gini coefficient and information value.
When variables are grouped, a decision tree is fitted for each of the data variables. For the model, there are four options available for the grouping methods, namely an optimal criterion, quantile, monotonic event rate and a constrained optimal. A monotonic event rate groups variables in a monotonic distribution of event rates across all the attributes, the event rate is equal to \( P(\text{equal} \mid \text{attribute}) \). This is a conditional probability of an event given that a customer shows certain behaviour.

Figure 4.1 shows the list of predictive variables once interactive grouping has been done.

**Figure 4.1: Interactive grouping – event rate**

![Interactive grouping – event rate](image)

4.3.2 **Gini coefficient for a linear regression and logistic regression model**

The Gini coefficient is a statistic used to measure the inequality of a distribution whereby the numerator is the area below the Lorenz curve and the 45-degree line. The denominator is the area below the 45-degree line. The Gini coefficient ranges from zero
to 100 percent; whereby, the closer the Gini coefficient is 100 percent better, this shows the predictability of each variable. The Lorenz curve is defined as follows:

\[
\text{Gini coefficient} = \frac{A}{A + B}
\]

For the interpretation of the logistic regression, the Gini coefficient as well as the information value will be looked at. The information value is used to measure the predictive power of the predictor and is used to measure the predictive power without any regard to an ordering of a predictor. An information value can be computed for any predictor given that \( g_k \) or \( b_k \) is not zero.

As a formula, information value is given by:

\[
IV = \sum_{k=1}^{L} (g_k - b_k) \ast \log(g_k/b_k)
\]

Where \( L \geq 2 \) and where \( g_k \) and \( b_k > 0 \) for all \( k = 1, ..., L \)
For the interpretation of the information value for a binary logistic regression model, guidelines are taken from Siddiqi (2006:81) such that:

**Figure 4.2: Information value interpretation**

- **Information Value > 0.03**
  - Shows that the predictor is strong

- **Information Value > 0.01 and Information Value ≤ 0.3**
  - Shows that the predictor is medium

- **Information Value > 0.02 and Information Value ≤ 0.1**
  - Shows that the predictor is weak

- **Information Value ≤ 0.02**
  - Shows that the predictor is not predictive
Table 4.4: Logistic regression predictive variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Gini coefficient</th>
<th>Information value</th>
<th>Statistics interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Closing balance amount</td>
<td>41.686</td>
<td>1.279</td>
<td>Strong predictor</td>
</tr>
<tr>
<td>Last month’s account risk category</td>
<td>25.612</td>
<td>1.237</td>
<td>Strong predictor</td>
</tr>
<tr>
<td>Profile risk category</td>
<td>25.577</td>
<td>1.232</td>
<td>Strong predictor</td>
</tr>
<tr>
<td>Yearly credit turnover</td>
<td>30.837</td>
<td>0.479</td>
<td>Strong predictor</td>
</tr>
<tr>
<td>Account age</td>
<td>31.418</td>
<td>0.349</td>
<td>Strong predictor</td>
</tr>
<tr>
<td>Customer type</td>
<td>25.35</td>
<td>0.228</td>
<td>Medium predictor</td>
</tr>
<tr>
<td>Number of sub product</td>
<td>20.589</td>
<td>0.179</td>
<td>Medium predictor</td>
</tr>
<tr>
<td>Number of cheque accounts</td>
<td>15.957</td>
<td>0.112</td>
<td>Medium predictor</td>
</tr>
<tr>
<td>Last deposited amount</td>
<td>10.494</td>
<td>0.042</td>
<td>Weak predictor</td>
</tr>
<tr>
<td>Customer total number of products</td>
<td>7.599</td>
<td>0.027</td>
<td>Weak predictor</td>
</tr>
</tbody>
</table>

Table 4.4 shows a list of variables that were found to be predictive based on the logistic regression model and there are two weak predictors, three medium predictors and five strong predictors.

4.3.3 Further analysis of the predictive variables using interactive grouping

The following variables were said to be predictive based on the Gini coefficient, information value (IV) and WOE that measures how much of the evidence supports or undermines the hypothesis.

- Closing balance amount

The closing balance amount is a variable that refers to the closing balance amount the customer had in their business banking cheque account prior to closing it. This shows that customers who had a negative closing balance amount less than R58.17 have a higher probability of closing their account due to business failure; with a group event rate of 0.684 and a WOE of -2.5013. The high negative of value of WOE correlates to
the high risk of business failure whereas high positive values refer to low risk of business failure.

Table 4.5: Interactive grouping on closing balance amount variable

<table>
<thead>
<tr>
<th>Variable</th>
<th>Group</th>
<th>WOE</th>
<th>Event count</th>
<th>Non-event count</th>
<th>Group rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Closing balance amount &lt; -R58.17</td>
<td>1</td>
<td>-2.5013</td>
<td>5588</td>
<td>2585</td>
<td>0.684</td>
</tr>
<tr>
<td>Closing balance &gt;= -R58.17</td>
<td>2</td>
<td>0.56736</td>
<td>6718</td>
<td>66857</td>
<td>0.091</td>
</tr>
</tbody>
</table>

Original Gini is 41.686
Information value (IV) is 1.279

- Last month’s account risk category

Last month’s account risk category variable refers to the customer’s risk category code on their business cheque account at the end of every month.

Customers who have a risk category of 3 and 4 have a group event rate of 0.956 and a WOE of -4.80199, which shows a likelihood of their businesses failing. Risk category 3 and 4 represent business customers who are at a risk of insolvency and fraud, ideal customers to model with based on their customer behaviour are customers with risk category code of 1 and 0 as they resemble normal customer behaviour, normal profile and have not been detected or monitored for fraud and insolvency.

Table 4.6: Interactive grouping on last month’s account risk category variable

<table>
<thead>
<tr>
<th>Variable</th>
<th>Group</th>
<th>WOE</th>
<th>Event count</th>
<th>Non-event count</th>
<th>Group rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0,1,missing</td>
<td>1</td>
<td>0.29669</td>
<td>9113</td>
<td>69186</td>
<td>0.116</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>-3.06178</td>
<td>496</td>
<td>131</td>
<td>0.791</td>
</tr>
<tr>
<td>3,4</td>
<td>3</td>
<td>-4.80199</td>
<td>2697</td>
<td>125</td>
<td>0.956</td>
</tr>
</tbody>
</table>

Original Gini is 25.612
Information value (IV) is 1.237
• Profile risk category

Profile risk category variable refers to the risk category code of the customer based on the customer’s overall profile behaviour based on all the products a customer holds within the bank. Customers who with a risk category of 3 and 4 have a group event rate of 0.955 and a WOE of -4.79365, which resembles a high probability of business failure. This shows there is a correlation between ‘last month’s account risk category code’ and ‘profile risk category’ when comparing the WOE, Gini coefficient and IV.

Table 4.7: Interactive grouping on account risk category variable

<table>
<thead>
<tr>
<th>Variable</th>
<th>Group</th>
<th>WOE</th>
<th>Event count</th>
<th>Non-event count</th>
<th>Group rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0,1, missing</td>
<td>1</td>
<td>0.29623</td>
<td>9116</td>
<td>69177</td>
<td>0.116</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>-2.99847</td>
<td>494</td>
<td>139</td>
<td>0.78</td>
</tr>
<tr>
<td>3,4</td>
<td>3</td>
<td>-4.79365</td>
<td>2696</td>
<td>126</td>
<td>0.955</td>
</tr>
</tbody>
</table>

Original Gini is 25.577
Information value (IV) is 1.237

• Yearly credit turnover

The yearly credit turnover variable refers to the turnover a business makes in one year. This shows that businesses with a yearly credit turnover less than R178.77 yearly have a high event rate of 0.348 with an event count of 5685. The calculated WOE is -1.101; the smaller your turnover the higher the probability of you closing your account.

Table 4.8: Interactive grouping on yearly credit turnover variable

<table>
<thead>
<tr>
<th>Variable</th>
<th>Group</th>
<th>WOE</th>
<th>Event count</th>
<th>Non-event count</th>
<th>Group rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yearly turnover &lt;=178.77</td>
<td>1</td>
<td>0.45308</td>
<td>6621</td>
<td>58776</td>
<td>0.101</td>
</tr>
<tr>
<td>Yearly turnover &gt;=178.77</td>
<td>2</td>
<td>-1.10117</td>
<td>5685</td>
<td>10666</td>
<td>0.479</td>
</tr>
</tbody>
</table>

Original Gini is 30.837
Information value (IV) is 0.479
- Number of products a customer has

The number of products variable refers to the number of products a business banking customer has within the bank. The vertical sales index (VSI) that is used within the bank tracks the number of products a customer holds within the bank. The more products a customer has, the less likely it is for them to close their cheque accounts and leave the bank. However, if a customer has less product holdings the more likely it is for them to end their banking relationship.

The table below shows us that those customers who have less than two products within the bank have a high probability of leaving the bank, looking at their WOE of -0.34975.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Group</th>
<th>WOE</th>
<th>Event count</th>
<th>Non-event count</th>
<th>Group rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer total number of products &lt;2</td>
<td>1</td>
<td>-0.34975</td>
<td>8585</td>
<td>34147</td>
<td>0.201</td>
</tr>
<tr>
<td>Customer total number of products &gt;2</td>
<td>2</td>
<td>0.51934</td>
<td>3721</td>
<td>35295</td>
<td>0.179</td>
</tr>
</tbody>
</table>

Original Gini is 20.589
Information value (IV) is 0.179

4.3.4 Linear regression model statistics

A proc correlation procedure was run on SAS Enterprise Guide to calculate the correlation and partial correlation among variables, to test if the relationship between the variables is significant or not (Santana, 2009). The relationship between the variables signifies the degree of linearity among them.

The proc correlation procedure took into account the following variables:

- Monthly credit turnover
- Profile risk category
- Quarterly credit amount
- Last deposit amount
• Current balance amount
• Credit limit amount
• Average credit balance amount
• Net balance amount
• High account balance amount
• Number of loans
• Account age (in months)
• Account last transaction date (in months)

Table 4.10: Proc correlation simple statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std dev</th>
<th>Sum</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quarterly credit amount</td>
<td>470219</td>
<td>159045</td>
<td>2928550</td>
<td>747860000000</td>
<td>0</td>
<td>1582787547</td>
</tr>
<tr>
<td>Last deposit amount</td>
<td>470219</td>
<td>22939</td>
<td>1215212</td>
<td>107865000000</td>
<td>0</td>
<td>791528471</td>
</tr>
<tr>
<td>Current balance amount</td>
<td>470219</td>
<td>23635</td>
<td>237967</td>
<td>111134000000</td>
<td>-17416793</td>
<td>39680505</td>
</tr>
<tr>
<td>Number of cheque accounts</td>
<td>470219</td>
<td>2.26776</td>
<td>4.06471</td>
<td>1066342</td>
<td>1</td>
<td>153</td>
</tr>
<tr>
<td>Number of investments</td>
<td>470219</td>
<td>0.32825</td>
<td>0.90248</td>
<td>154350</td>
<td>0</td>
<td>180</td>
</tr>
<tr>
<td>Number of loans</td>
<td>470219</td>
<td>0.00661</td>
<td>0.086</td>
<td>3110</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Account age in Months</td>
<td>470219</td>
<td>16.92907</td>
<td>16.58442</td>
<td>7960370</td>
<td>1</td>
<td>323</td>
</tr>
<tr>
<td>Account last transaction date in Months</td>
<td>470219</td>
<td>38.76181</td>
<td>100.7534</td>
<td>1010404</td>
<td>-9583</td>
<td>1379</td>
</tr>
</tbody>
</table>

Table 4.10 shows the proc correlation simple statistics for each variable, this list contains the number of observations, mean, standard deviation, sum, minimum and the maximum. This shows that the population of the observations is 470219 whereby for the ‘Number of loans’ a business customer has with the bank is a minimum of one (1) loan and a maximum of four (4) loans and the ‘Account age in months’ for the business cheque account is one month and a maximum account age is 323 months.
Table 4.11 shows the Pearson correlation coefficients and it measures the linear relationship between two variables whereby correlation could be clouded by relationships that exist among other variables. For example, the increase in monthly credit turnover will increase the quarterly credit turnover or the decrease in monthly credit turnover could reduce the quarterly credit turnover. Therefore, a change in factors relating to the monthly credit turnover is causing the change known as a spurious relationship.

Table 4.11: Pearson correlation coefficients

<table>
<thead>
<tr>
<th></th>
<th>Monthly credit turnover</th>
<th>Last month’s end balance</th>
<th>Quarterly credit turnover</th>
<th>Yearly credit turnover</th>
<th>Number of products</th>
<th>Customer total number of sub-products</th>
<th>Customer number of cheque accounts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monthly credit turnover</td>
<td>1</td>
<td>0.36959</td>
<td>0.93963</td>
<td>0.93963</td>
<td>0.93963</td>
<td>0.00058</td>
<td>0.00219</td>
</tr>
<tr>
<td>Monthly credit turnover</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>0.6471</td>
<td>0.0861</td>
<td></td>
</tr>
<tr>
<td>Quarterly credit turnover</td>
<td>0.93963</td>
<td>0.4356</td>
<td>1</td>
<td>0.84021</td>
<td>0.0205</td>
<td>-0.00236</td>
<td>0.00424</td>
</tr>
<tr>
<td>Quarterly credit turnover</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>0.0638</td>
<td>0.0009</td>
<td></td>
</tr>
</tbody>
</table>

By assuming that the data used is normal, the p-values can be used to test the hypothesis that the true correlation between two variables is zero. Hypothesis to be tested using the proc correlation is as follows:

\[ H_0 = 0 \text{ vs } H_A \neq 0 \]
By using a significance level of $\alpha = 0.05$, we cannot reject a hypothesis of $H_0$ because $\rho > \alpha$ meaning that we cannot reject a statement that states that the Pearson correlation between two variables is zero.

Looking at table 4.11 there appears to a strong linear relationship between the ‘monthly credit turnover’ and ‘last month’s end balance’ as the sample Pearson correlation between the variables is 0.36959, which is greater than the $\alpha = 0.05$ where $\rho > \alpha$ . Other strong relationships are seen by the following variable pairs ‘monthly credit turnover’ and ‘customer total number of products’ with a correlation of 0.96963; the number of ‘cheque accounts’ a customer has is correlated to the ‘number of customer of sub-products’. There appears to be little or no linear relationship between ‘customer number of products’ and ‘last month’s end balance’ due to a Pearson correlation coefficient of 0.00246.

The rest of the proc correlation statistics will be found in appendix C and the Pearson correlation coefficients will be found in appendix D.

**Table 4.12: Linear regression Gini statistics of predictive variables**

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Label</th>
<th>Group</th>
<th>Gini coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Profile risk category code &gt; 0</td>
<td>&gt;0</td>
<td>1</td>
<td>46.37231</td>
</tr>
<tr>
<td>Profile risk category code = 0</td>
<td>0</td>
<td>2</td>
<td>61.71816</td>
</tr>
<tr>
<td>Account age &lt; 13</td>
<td>&lt;13</td>
<td>1</td>
<td>49.55349</td>
</tr>
<tr>
<td>Account age &lt; 25</td>
<td>&lt;25</td>
<td>2</td>
<td>55.00953</td>
</tr>
<tr>
<td>Account age &lt; 36</td>
<td>&lt;36</td>
<td>3</td>
<td>58.1531</td>
</tr>
<tr>
<td>Account age &lt; 121</td>
<td>&lt;121</td>
<td>4</td>
<td>63.07645</td>
</tr>
<tr>
<td>Account age &lt; 181</td>
<td>&lt;181</td>
<td>5</td>
<td>65.41986</td>
</tr>
<tr>
<td>Account age &lt; 241</td>
<td>&lt;241</td>
<td>6</td>
<td>70.13873</td>
</tr>
<tr>
<td>Account age ≥ 241</td>
<td>&gt;=241</td>
<td>7</td>
<td>72.98899</td>
</tr>
<tr>
<td>Yearly credit turnover &lt;30000</td>
<td>&lt; 30000</td>
<td>1</td>
<td>37.3506</td>
</tr>
<tr>
<td>Yearly credit turnover &lt;75000</td>
<td>&lt;75000</td>
<td>2</td>
<td>39.56083</td>
</tr>
<tr>
<td>Yearly credit turnover &lt;150000</td>
<td>&lt;150000</td>
<td>3</td>
<td>41.48047</td>
</tr>
<tr>
<td>Yearly credit turnover ≥150000</td>
<td>&gt;=150000</td>
<td>4</td>
<td>43.84993</td>
</tr>
<tr>
<td>Last transaction date in months&lt;3</td>
<td>&lt;3</td>
<td>1</td>
<td>44.96313</td>
</tr>
</tbody>
</table>
4.3.5 Receiver operating characteristics chart for a linear regression model

The ROC (receiver operating characteristics) curve is used to help identify customers whose businesses are more likely to fail due to business failure based on a cut off of 0.65.

Below is a list of things to help demonstrate different cut off points:

- The decrease in specificity and the increase in sensitivity showing the trade-off between the specificity and sensitivity
- The further away the curve is from the 45-degree line, the more accurate it is
- If the curve is closer to the 45-degree curve, the test is less accurate
- Below the ROC curve lays an area that shows the measure of accuracy
- Training Gini coefficient of 0.333 and misclassification rate of 0.301238
- Valid Gini coefficient of 0.339 and misclassification rate of 0.30017
The misclassification rate of the linear regression model is 0.301238=30.12 percent, which is a false positive rate that shows the proportion of misclassified observations. The linear regression has a prediction power of 69 percent.

<table>
<thead>
<tr>
<th>Model</th>
<th>Node</th>
<th>Model description</th>
<th>Train - misclassification rate</th>
<th>Train - squared error</th>
<th>Valid - misclassification rate</th>
<th>Valid - squared error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y Reg</td>
<td>Linear regression</td>
<td>0.301238</td>
<td>0.194225</td>
<td>0.30017</td>
<td>0.193745</td>
<td></td>
</tr>
</tbody>
</table>

### 4.3.6 Receiver operating characteristics chart for a logistic regression model

The ROC (receiver operating characteristics) is a graph that shows the indicative capability of a binary classification system that shows in which of the two groupings an element is predicted to belong to, as its bias threshold differs. Binary classification is normally used in neural networks, Bayesian networks, decision trees and logistic
regression. The vertical axis of the graph shows the true positive rate, known as the sensitivity and the horizontal axis of the graphs shows the false positive rate, known as the error prediction.

The curve is used to help identify customers whose businesses are more likely to fail due to business failure and customers who would not fail based on a cut off of 0.65. Below is a list of things to help demonstrate different cut-off points:

- The decrease in specificity and the increase in sensitivity showing the trade-off between the specificity and sensitivity
- The further away the curve is from the 45-degree line, the more accurate it is
- If the curve is closer to the 45-degree curve, the test is less accurate.
- Below the ROC curve lays an area that shows the measure of accuracy
- Training Gini coefficient of 0.69 and misclassification rate of 0.10380
- Valid Gini coefficient of 0.702 and misclassification rate of 0.105114.

**Figure 4.4: Logistic regression ROC chart**
The misclassification rate of the logistic regression model is 0.10380=10.38 percent, which is a false positive rate that shows the proportion of misclassified observations. The logistic regression has a prediction power of 89 percent, which is far greater than that of a linear regression model.

<table>
<thead>
<tr>
<th>Model</th>
<th>Node</th>
<th>Model description</th>
<th>Train - misclassification rate</th>
<th>Train - squared error</th>
<th>Valid - misclassification rate</th>
<th>Valid - squared error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>Reg3</td>
<td>Logistic regression</td>
<td>0.10380</td>
<td>0.084064</td>
<td>0.10519</td>
<td>0.082795</td>
</tr>
</tbody>
</table>

Table 4.13 below looks into the confusion matrix of a logistic regression model whereby a confusion matrix is a technique that calculates the performance of the classification algorithm. By calculating classification accuracy alone could be misleading if the observations in a data set are not equal so by calculating the confusion matrix it gives a better idea of the types of errors the model is making and the classifications it’s getting right. The confusion matrix below describes the performance of the logistic regression model and confirms its predicting power:

**Table 4.13: Confusion matrix**

<table>
<thead>
<tr>
<th><strong>True positive:</strong> customers correctly predicted for business failure.</th>
<th><strong>False positive:</strong> other incorrectly predicted as Business failure.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>False negative:</strong> business failure incorrectly predicted as other.</td>
<td><strong>True negative:</strong> other correctly predicted as other.</td>
</tr>
</tbody>
</table>
Sensitivity and specificity are calculated as follows:

\[
\text{Sensitivity} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} = 0.709944
\]

\[
\text{Specificity} = \frac{\text{False Positive}}{\text{False Positive} + \text{True Negative}} = 0.906329
\]

Sensitivity is the percentage of customers who are correctly identified to have closed their accounts due to business failure.

Specificity is the percentage of customers who are correctly identified to have not closed their accounts due to business failure.
4.3.7 The cumulative lift chart

By determining customers who are more likely to close their business cheque accounts due to business failure, a cumulative lift chart is used in Enterprise Miner to explain the effectiveness of the model. The axis of the cumulative graph reflects different outcomes whereby the vertical axis shows the lift of the chart and the horizontal axis shows the percentile. No model also known as the default is when a horizontal line intersects the y-axis at one, in a case where the y-axis is one means that 10 percent of the customer base can randomly contacted without using the model then 10 percent of the customers who are at risk of business failure can be identified.

Using the chart, one should be able to capture about 32 to 34 percent of the customers who are at risk of business failure by contacting customers for an intervention by selecting customers with high probabilities of business failure in the top 10 percentile. For example, can contact the top 20 percent of the customers and could end up with a cumulative lift curve of about 3.1, which means that 62 percent of customers who run the risk of business failure can be captured if we could just communicate to 20 percent of them.

A cumulative lift curve shows the advantage of using a predictive model to determine, which customers to communicate with regarding business failure as a customer retention strategy to remedy the situation.
Figure 4.5: Cumulative lift chart for training data and validation data
4.4 PILOT CAMPAIGN RESULTS

Seeing that a logistic regression model is the best model compared to the linear regression, therefore the results of the logistic regression model will be used to identify customers who have a probability of business failure. By using the model scoring code created in the enterprise miner guide for the logistic regression model. After executing the scoring code on the current business banking customers using the same extraction rules, an output with business failure probabilities per business customer will be used in order to know which customers have a probability of business failure and will therefore be campaigned to as a customer retention strategy, so that the bank can see which customers they can retain through the use of a campaign or by offering them business banking value added products or services such as a business incubation program. An SMS and phone call campaign is used to engage with customers thus focusing on business cheque accounts that are inactive and dormant.

An inactive account is an account that has not been used for three to four months and when an account is inactive, customers can reactivate their accounts by performing a transaction via the APP, online banking, ATM deposit to the account or by swiping their cheque card. A dormant account is an account that has not been used for more than five months and in order for the account to be activated, the account holder has to go to the bank branch to get the account reactivated. The pilot campaign compares the inactive and dormant accounts to one another as they have different reactivation methods.

These leads had to be subjected to the same filters mentioned in Section 4.4.1 of this document. Additionally, these had to be customer records, which were not used in the model building process (for remaining unbiased). Before a campaign can be launched, the data needs to be filtered according to the extraction rules by checking the following:

- Marketing consent – marketing consent checks if a customer consented to be marketed to by BankX.
- Know Your Customer (KYC) – KYC checks if the customers have FICA’d their documents
- Business age – customer business age must be greater than six months
- Account age – checks the account age
- Region – checks which region the customer resides in.
- Redslit – redslist is checked to ensure that BankX complies with marketing rules such that a customer can only be contacted via SMS and email every 15 days and can be campaigned via phone call every 30 days. Therefore, redslit is customer leads base (CLB) that checks if there are no duplicates on the same channel for the same customer, to avoid a business customer receiving different marketing communication via SMS within a 15-day period.

4.4.1 Campaign structure

Using the same set of leads for three months, the campaign will be structured as follows:

Figure 4.6: Campaign structure

From the first set of leads that will be used for three months, a customer will receive a reactivation SMS in month 1 and if the business customer does not reactivate their account they will be sent a reminder SMS in month 2. The last reactivation method will take place in month 3 whereby the business customer will be called to reactivate their
account and told of the different methods in which they can reactivate their accounts while the call centre agent cross sells other products to the customer.

4.4.2 Campaign messaging

The marketing communication is structured as follows:

Month 1: SMS Campaign

Dormant with reward: Do you want to increase your rewards balance of R125? Simply visit your nearest BankX branch to reactivate your business account ending with xxxxxxx1111 that is currently dormant and get one step closer to earning rewards. Ts, Cs and rules apply.

Month 2: Reminder SMS

Dormant with rewards: Customer reminder: your business account ending with xxxxxx1111 is currently dormant. Please visit your nearest BankX branch to reactivate and get one step closer to start earning Rewards for Business. Ts, Cs and rules apply.

The messages above were created by the marketing team of BankX ensuring that the campaign messaging is compliant with the bank and meets the objective of the campaign, which is to reactivate the accounts of business banking customers who have behaviours of business failure.

Month 3: Phone call campaign

A follow up phone call campaign will be done for customers who do not reactivate their business cheque accounts after receiving an initial SMS campaign and a reminder SMS campaign. During the phone call campaign, the call centre agent will explain the purpose of the phone call, explain how to reactivate the cheque account and to cross sell products that the business customer qualifies for.
4.4.3 Campaign results

The campaign started in October 2017. Below are the results of the three-month campaign whereby a customer base of 13,426 business banking customers was selected to be part of the campaign using the campaign rules and extraction rules:

Month 1: SMS

An initial SMS was sent out to 13,426 business banking customers to reactivate their business cheque accounts. The SMS campaign included customers with inactive business cheque accounts and customers with dormant business cheque accounts. Customers who do not reactivate their accounts using the initial SMS campaign, will be cycled through another campaign method to measure the effectiveness of the campaign reactivation rate.

Figure 4.7: The list of customers to be campaigned to using the initial SMS were distributed as follows:

<table>
<thead>
<tr>
<th>Initial SMS Campaign</th>
<th>Number of customers who were sent an SMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dormant customers with reward programme</td>
<td>767</td>
</tr>
<tr>
<td>Dormant customers without reward programme</td>
<td>1826</td>
</tr>
<tr>
<td>Inactive customers with reward programme</td>
<td>5003</td>
</tr>
<tr>
<td>Inactive customers without reward programme</td>
<td>5830</td>
</tr>
</tbody>
</table>

Figure 4.8 shows the holistic view for the reactivation rate for customers who have ‘inactive cheque accounts without a reward programme is 22 percent, ‘inactive cheque accounts with a reward programme is 20 percent, ‘dormant accounts without reward programme is 5 percent and ‘dormant accounts with reward programme is 9 percent.
For dormant accounts, the reactivation rate is less than 10 percent, which could be caused by the reactivation process for dormant accounts that requires business customers to go to branch to activate their business cheque accounts instead of swiping their cards or performing a transaction. Most business owners do not have a lot of time to go to a branch, which results in a low reactivation rate.

**Figure 4.8: Holistic view of the initial SMS campaign**

<table>
<thead>
<tr>
<th>Holistic view of the initial SMS Campaign</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inactive customers without reward programme</td>
</tr>
<tr>
<td>Inactive customers with reward programme</td>
</tr>
<tr>
<td>Dormant customers without reward programme</td>
</tr>
<tr>
<td>Dormant customers with reward programme</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Activation rate = (activated accounts/number of...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of customers who activated their accounts...</td>
</tr>
<tr>
<td>1089</td>
</tr>
<tr>
<td>Number of SMS’s not delivered</td>
</tr>
<tr>
<td>Number of SMS’s that got delivered</td>
</tr>
<tr>
<td>Number of customers who were sent an SMS</td>
</tr>
</tbody>
</table>

**Month 2: Reminder SMS**

Figure 4.9 a reminder SMS is only sent out if a business customer does not respond to the first SMS that was sent out regarding the status of their account and did not reactivate their business cheque account. A reminder SMS was sent out to 8 835 business banking customers whereby the campaign leads were split among customers with inactive cheque accounts and dormant cheque accounts with or without a reward programme.

There were 6 548 customers with inactive cheque accounts and 2 287 customers with dormant accounts. The graphical view below shows the distribution of leads for the reminder SMS:
Figure 4.9: Reminder SMS

Reminder SMS Campaign

Activation rate = (activated accounts/number of SMS’s delivered)*100

<table>
<thead>
<tr>
<th>Number of customers who activated their accounts after receiving SMS</th>
<th>Inactive customers without reward programme</th>
<th>Inactive customers with reward programme</th>
<th>Dormant customers without reward programme</th>
<th>Dormant customers with reward programme</th>
</tr>
</thead>
<tbody>
<tr>
<td>334</td>
<td>12%</td>
<td>9%</td>
<td>2%</td>
<td>2%</td>
</tr>
<tr>
<td>604</td>
<td>3%</td>
<td>15%</td>
<td>25%</td>
<td>25%</td>
</tr>
<tr>
<td>2899</td>
<td>12%</td>
<td>9%</td>
<td>2%</td>
<td>2%</td>
</tr>
<tr>
<td>3503</td>
<td>3%</td>
<td>15%</td>
<td>25%</td>
<td>25%</td>
</tr>
</tbody>
</table>

Number of SMS’s not delivered

<table>
<thead>
<tr>
<th>Number of SMS’s not delivered</th>
<th>Inactive customers without reward programme</th>
<th>Inactive customers with reward programme</th>
<th>Dormant customers without reward programme</th>
<th>Dormant customers with reward programme</th>
</tr>
</thead>
<tbody>
<tr>
<td>604</td>
<td>3%</td>
<td>15%</td>
<td>25%</td>
<td>25%</td>
</tr>
<tr>
<td>2899</td>
<td>12%</td>
<td>9%</td>
<td>2%</td>
<td>2%</td>
</tr>
<tr>
<td>3503</td>
<td>3%</td>
<td>15%</td>
<td>25%</td>
<td>25%</td>
</tr>
</tbody>
</table>

Number of SMS’s that got delivered

<table>
<thead>
<tr>
<th>Number of SMS’s that got delivered</th>
<th>Inactive customers without reward programme</th>
<th>Inactive customers with reward programme</th>
<th>Dormant customers without reward programme</th>
<th>Dormant customers with reward programme</th>
</tr>
</thead>
<tbody>
<tr>
<td>256</td>
<td>9%</td>
<td>2%</td>
<td>2%</td>
<td>2%</td>
</tr>
<tr>
<td>155</td>
<td>3%</td>
<td>15%</td>
<td>25%</td>
<td>25%</td>
</tr>
<tr>
<td>1344</td>
<td>12%</td>
<td>9%</td>
<td>2%</td>
<td>2%</td>
</tr>
<tr>
<td>670</td>
<td>3%</td>
<td>15%</td>
<td>25%</td>
<td>25%</td>
</tr>
</tbody>
</table>

Number of customers who were sent an SMS

<table>
<thead>
<tr>
<th>Number of customers who were sent an SMS</th>
<th>Inactive customers without reward programme</th>
<th>Inactive customers with reward programme</th>
<th>Dormant customers without reward programme</th>
<th>Dormant customers with reward programme</th>
</tr>
</thead>
<tbody>
<tr>
<td>3503</td>
<td>3%</td>
<td>15%</td>
<td>25%</td>
<td>25%</td>
</tr>
<tr>
<td>3045</td>
<td>12%</td>
<td>9%</td>
<td>2%</td>
<td>2%</td>
</tr>
<tr>
<td>1617</td>
<td>3%</td>
<td>15%</td>
<td>25%</td>
<td>25%</td>
</tr>
<tr>
<td>670</td>
<td>3%</td>
<td>15%</td>
<td>25%</td>
<td>25%</td>
</tr>
</tbody>
</table>

Figure 4.10 the holistic view below shows that the reactivation rate for customers who have ‘inactive cheque accounts without a reward programme’ is 12 percent, ‘inactive cheque accounts with a reward programme’ is 9 percent, ‘dormant accounts without reward programme’ is 2 percent and ‘dormant accounts with reward programme’ is 2 percent. For dormant accounts, the reactivation rate is less than 10 percent, which could be caused by the reactivation process for dormant accounts that requires business customers to go to branch to activate their business cheque accounts instead of swiping their cards or performing a transaction. Most business owner’s do not have a lot of time to go to a branch, which results in a low reactivation rate.
Figure 4.10: Holistic view of the reminder SMS campaign

<table>
<thead>
<tr>
<th></th>
<th>Inactive customers without reward programme</th>
<th>Inactive customers with reward programme</th>
<th>Dormant customers without reward programme</th>
<th>Dormant customers with reward programme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activation rate = (activated accounts/number of...</td>
<td>12%</td>
<td>9%</td>
<td>2%</td>
<td>2%</td>
</tr>
<tr>
<td>Number of customers who activated their...</td>
<td>334</td>
<td>326</td>
<td>257</td>
<td>257</td>
</tr>
<tr>
<td>Number of SMS’s not delivered</td>
<td>604</td>
<td>598</td>
<td>273</td>
<td>273</td>
</tr>
<tr>
<td>Number of SMS’s that got delivered</td>
<td>2899</td>
<td>2890</td>
<td>1344</td>
<td>1677</td>
</tr>
<tr>
<td>Number of customers who were sent an SMS</td>
<td>3503</td>
<td>3045</td>
<td>1617</td>
<td>670</td>
</tr>
</tbody>
</table>

Month 3: Phone call campaign

Figure 4.11 below shows the structure of the follow-up phone call campaign. 1,963 customers were selected to be part of the phone call campaign as they had not reactivated their business cheque account from the first SMS they received and the reminder SMS. From the 1,963 customers who were part of the phone call campaign, 885 customers were contacted but only 303 phone calls were completed where the agent managed to discuss the purpose of the phone call, reactivate the cheque account and to cross sell products that the business customer qualified for. The activation rate based on the follow-up phone call campaign was 70 percent.

This shows that 70 percent of the customers who were contacted via the phone call campaign reactivated their accounts by either doing a transaction on online banking, the APP, ATM or by swiping their transactional card at any of the retail stores or fuel stations.
According to BankX marketing requirements, a successful SMS campaign should have a minimum activation rate of 5 percent and for a successful phone call campaign, it should be a minimum of 15 percent. This shows that the three-month campaign has been very successful and has met BankX marketing requirements.

SMSs that could not be delivered were due to network failure or the customer’s cell phone being off.

In chapter 4, the linear regression model and the logistic regression model get compared to test the model performance, accuracy and predictive power. For the two models the same extraction rules were applied and a workable sample size base is imported into the enterprise miner model builder to create the two models.

The data partition is split 70 percent training data set and 30 percent validation data set for both models. Interactive grouping is a different method that is used in the study on the enterprise miner tool to help process the data before starting the modelling process, interactive grouping involves the grouping of initial bins into quantiles and thus allowing the interactive splitting and combining of the initial bins, which helps to
improve the variables predictive power by means of generated weight of evidence (WOE) for each of the groupings of the variables.

The model has two weak predictors, three medium predictors and five strong predictors.

The misclassification rate of the linear regression model is 0.301238 = 30.12 percent, which is a false positive rate that shows the proportion of misclassified observations. The linear regression has a prediction power of 69 percent.

The misclassification rate of the logistic regression model is 0.10380 = 10.38 percent, which is a false positive rate that shows the proportion of misclassified observations. The logistic regression has a prediction power of 89 percent, which is far greater than that of a linear regression model. Therefore, the logistic regression model is the better performing model compared to the linear regression model.
CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS

5.1 INTRODUCTION

This chapter includes conclusions and recommendations relating to the study. Conclusions made are based on the data analysis and are formulated based on the literature review and empirical research. It also includes limitations and areas for further research. The aim of the study is to confirm if a logistic regression model or a linear regression model can be used to predict business failure and be used as a customer retention strategy.

5.2 CONCLUSIONS

Through the research study, the following conclusions were drawn:

- A linear regression model and a logistic regression model were built, to compare the model’s performance and to choose a better performing model.
- The misclassification rate of the linear regression model is 0.301238 = 30.12 percent, which is a false positive rate that shows the proportion of misclassified observations. The linear regression has a prediction power of 69 percent.
- The misclassification rate of the logistic regression model is 0.10380 = 10.38 percent, which is a false positive rate that shows the proportion of misclassified observations. The logistic regression has a prediction power of 89 percent, which is far greater than that of a linear regression model.
- A logistic regression model was a better performing model due to its accuracy and predicting power when compared to the linear regression model.
- The following variables were found to be significant strong predictors that can be used to predict and identify business failure; closing balance amount, last month’s risk category, profile risk category, yearly credit turnover, account age, customer type and product holdings.
- Business banking financial significant variables were identified, to predict if a business customer’s business is failing.
• Using the scoring code from the logistic regression model, probabilities of business failure can be obtained by running the scoring code on a current business banking base.

• The primary objective which is to test if modelling can be used to predict business failure in the South African business bank, has been achieved.

• Significance of the model variables are based on the WOE which is defined as a logarithm; a fraction of the non-event counts of the observations in the interactive group over the proportion of the event count of the observations. The relative risk of any interactive grouping is measured by the WOE. Interactive grouping allows users to choose variables with better quality and prediction based on the Gini coefficient and information value.

• Business banking customers can be successfully retained using a customer retention strategy. Through a customer retention strategy, business banking customers will be provided with products that meet their business needs, be advised on their business’s financial positioning and be given support through the bank’s entrepreneurship programme, which is generally given to business banking customers at no cost.

• With the availability of big data and analytical tools, there is also the challenge of data quality, data integrity and data access. As business banks data generally are situated across different servers and warehouses, it will require the data to be merged from different warehouses and be put into a sensible format, which is a complex process.

• Analytical tools and statistical methodologies can be used to predict and detect business failure across the business banking sector.

• Business banks differentiate themselves through service fees, quality of service, incentives, customer value, business image, innovation and making the customer comfortable by offering them different servicing platforms where customers can transact, submit bank-required documents and update their details such as branch, online banking, ATM, APP, cell phone banking as well as WhatsApp banking, which is the first social media banking platform to be introduced in South Africa by ABSA Group Limited. WhatsApp banking was previously launched in India in May 2018. The WhatsApp banking platform allows ABSA Group Limited
customers to check their bank balance, buy electricity, data, airtime as well being able to pay beneficiaries (BusinessTech, 2018).

- The SAS tool is used to perform data analysis, data management, data manipulation and importing of data, to name a few. This helped to process thousands of rows in a short period of time thus being very effective in providing results. The SAS tool helps banks and businesses to increase their business functions.

- Banking sectors have a lot of responsibility in terms of managing the banks data and ensuring that their systems are always updated. Through the study, it can be seen that the value of data analysis and manipulation, as it provides insight and shows correlations among variables. Analytics provide more information in terms of developing strategies.

- The more products a customer has, the less chances of customer attrition, therefore, the more products a customer has, the less likely it is for them to end their relationship with the bank.

- If a customer has a long relationship with the bank the less likely it is for a customer to attrite.

- When customer loyalty decreases and sales become less, customer retention remains an essential component, because if retention is not well managed the key customer will go to the competitor, which affects the businesses growth and profitability

5.3 STUDY RECOMMENDATIONS

Study recommendations are done using the study conclusion. Below is a list of recommendations:

- SAS tool should be used in any research regarding data analysis, manipulation and model building; many researchers do not normally use SAS.

- Business banks should continuously analyse their customers’ data to better understand customer behaviour, find trends and retain profitable customers so that they will be able to distinguish themselves from their competitors.

- Business banks need to ensure that, with the acquisition process, customers are communicated with effectively in terms of the type of product they have, product
benefits, product education and a call to action if a customer has a query that they call in.

- By mapping out an acquisitions process, a business bank will know what communications are sent out to the customer and at which period in the process. This will ensure that a customer receives a welcome communication 24 hours after opening their account. The communication will have more information about how the customer can activate their account and the available product benefits. Once the customer activates his/her account after receiving his/her welcome communication, the next step should be to send the business customer communications such as a debit order switching communication.

- Business banks need to start targeting customers based on trends observed through data mining by ensuring that each campaign becomes meaningful. That can only be done when a business bank has an objective that needs to be met. Therefore, it is important to provide a customer with what they need.
REFERENCES


Myers, M.D. 2013. Qualitative Research in Business and Management. SAGE Publications.


APPENDIX A

DECLARATION BY LANGUAGE EDITOR
Ms Linda Scott
English language editing
SATI membership number: 1002595
Tel: 083 654 4156
E-mail: lindascott1984@gmail.com

26 September 2018

To whom it may concern

This is to confirm that I, the undersigned, have language edited the dissertation of

Nonvula Lydia Motsale

for the degree
Master of Science in Computer Science

entitled:

Predicting business failure in a South African business bank

The responsibility of implementing the recommended language changes rests with the author of the dissertation.

Yours truly,

Linda Scott
APPENDIX B

LIST OF SOUTH AFRICAN BANK INCLUDING MUTUAL BANKS, LOCALLY CONTROLLED BANKS, FOREIGN CONTROLLED BANKS, BANKS IN LIQUIDATION, BRANCHES OF FOREIGN BANKS AND FOREIGN BANK REPRESENTATIVES
<table>
<thead>
<tr>
<th>Categories</th>
<th>Names</th>
</tr>
</thead>
</table>
| Mutual banks                | • Finbond Mutual Bank  
                               | • GBS Mutual Bank  
                               | • VBS Mutual Bank |
| Locally controlled banks    | • African Bank Limited  
                               | • Bidvest Bank Limited  
                               | • Capitec Bank Limited  
                               | • Firstrand Bank Limited  
                               | • Grindrod Bank Limited  
                               | • Investec Bank Limited  
                               | • Nedbank Limited  
                               | • Sasfin Bank Limited  
                               | • The Standard Bank Of South Africa Limited  
                               | • Ubank limited  
                               | • Discovery Bank Limited  |
| Foreign controlled banks    | • ABSA Bank Limited  
                               | • Albraka Bank Limited  
                               | • Habib Overseas Bank Limited  
                               | • HBZ bank Limited  
                               | • Mercantile Bank Limited  
                               | • The South African Bank Of Athens Limited  
                               | • COMMONWEALTH BANK OF SOUTH AFRICA also known as TymeDigital by Commonwealth Bank SA  |
| Banks in liquidation        | • Islamic Bank Limited (In Final Liquidation)  
<pre><code>                           | • Regal Treasury Private Bank Limited (In Liquidation)  |
</code></pre>
<p>| Branches of foreign banks   | • Bank of Baroda  |</p>
<table>
<thead>
<tr>
<th>Foreign bank representatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>• AfrAsia Bank Limited</td>
</tr>
<tr>
<td>• Banco BIC</td>
</tr>
<tr>
<td>• Banco Santander Totta S.A.</td>
</tr>
<tr>
<td>• Bank of America</td>
</tr>
<tr>
<td>• National Association</td>
</tr>
<tr>
<td>• Bank One Limited</td>
</tr>
<tr>
<td>• Banque SYZ SA</td>
</tr>
<tr>
<td>• CaixaBank</td>
</tr>
<tr>
<td>• Commerzbank AG Johannesburg</td>
</tr>
<tr>
<td>• Ecobank Ghana Limited</td>
</tr>
<tr>
<td>• Export-Import Bank of India</td>
</tr>
<tr>
<td>• Hellenic Bank Public Company Limited</td>
</tr>
<tr>
<td>• Industrial and Commercial Bank of China African Representative Office</td>
</tr>
<tr>
<td>Bank Name</td>
</tr>
<tr>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>KfW Ipex-Bank GmbH</td>
</tr>
<tr>
<td>Millennium BCP</td>
</tr>
<tr>
<td>Mizuho Bank Limited</td>
</tr>
<tr>
<td>National Bank of Egypt</td>
</tr>
<tr>
<td>Notenstein La Roche Private Bank Limited</td>
</tr>
<tr>
<td>Novo Banco</td>
</tr>
<tr>
<td>Société Générale Representative Office for</td>
</tr>
<tr>
<td>Southern Africa</td>
</tr>
<tr>
<td>Sumitomo Mitsui Banking Corporation,</td>
</tr>
<tr>
<td>Swedbank AB</td>
</tr>
<tr>
<td>The Bank of New York Mellon</td>
</tr>
<tr>
<td>The Bank of Tokyo-Mitsubishi UFJ, Ltd</td>
</tr>
<tr>
<td>The Mauritius Commercial Bank Limited</td>
</tr>
<tr>
<td>The Rep. Off. for Southern and Eastern Africa</td>
</tr>
<tr>
<td>The Export-Import Bank of China</td>
</tr>
<tr>
<td>Unicredit Bank AG</td>
</tr>
<tr>
<td>Wells Fargo Bank</td>
</tr>
<tr>
<td>National Association</td>
</tr>
<tr>
<td>Zenith Bank Plc</td>
</tr>
</tbody>
</table>

**Source:** South African Reserve Bank (2018)
APPENDIX C

PROC CORRELATION SIMPLE STATISTICS
<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Sum</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quarterly credit</td>
<td>470219</td>
<td>159045</td>
<td>2928550</td>
<td>747860000000</td>
<td>0</td>
<td>1582787547</td>
</tr>
<tr>
<td>amount</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Last deposit</td>
<td>470219</td>
<td>22939</td>
<td>1215212</td>
<td>107865000000</td>
<td>0</td>
<td>791528471</td>
</tr>
<tr>
<td>amount</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Current balance</td>
<td>470219</td>
<td>23635</td>
<td>237967</td>
<td>111134000000</td>
<td>-17416793</td>
<td>39680505</td>
</tr>
<tr>
<td>amount</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Credit limit</td>
<td>470219</td>
<td>-4414</td>
<td>58231</td>
<td>-20756600000</td>
<td>-19230000</td>
<td>0</td>
</tr>
<tr>
<td>amount</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average credit</td>
<td>470219</td>
<td>25655</td>
<td>213828</td>
<td>120634000000</td>
<td>-7677</td>
<td>37177683</td>
</tr>
<tr>
<td>balance amount</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Net balance</td>
<td>470219</td>
<td>24139</td>
<td>239310</td>
<td>113506000000</td>
<td>-17416793</td>
<td>45150057</td>
</tr>
<tr>
<td>amount</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High account</td>
<td>470219</td>
<td>128316</td>
<td>3429353</td>
<td>603367000000</td>
<td>-15437235</td>
<td>791528971</td>
</tr>
<tr>
<td>balance amount</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overdraft excess</td>
<td>470219</td>
<td>-138.554</td>
<td>5023</td>
<td>-65150873</td>
<td>-1223400</td>
<td>0</td>
</tr>
<tr>
<td>amount</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low balance</td>
<td>470219</td>
<td>13077</td>
<td>189433</td>
<td>6148925309</td>
<td>-18566793</td>
<td>35364884</td>
</tr>
<tr>
<td>amount</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of cheque</td>
<td>470219</td>
<td>2.26776</td>
<td>4.06471</td>
<td>1066342</td>
<td>1</td>
<td>153</td>
</tr>
<tr>
<td>accounts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of investments</td>
<td>470219</td>
<td>0.32825</td>
<td>0.90248</td>
<td>154350</td>
<td>0</td>
<td>180</td>
</tr>
<tr>
<td>Number of loans</td>
<td>470219</td>
<td>0.00661</td>
<td>0.086</td>
<td>3110</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Account age in Months</td>
<td>470219</td>
<td>16.92907</td>
<td>16.58442</td>
<td>7960370</td>
<td>1</td>
<td>323</td>
</tr>
</tbody>
</table>
APPENDIX D

PEARSON CORRELATION COEFFICIENTS

PROB > |R| UNDER H0: RHO=0

NUMBER OF OBSERVATIONS
<table>
<thead>
<tr>
<th></th>
<th>Monthly credit turnover</th>
<th>Last month’s end balance</th>
<th>Quarterly credit turnover</th>
<th>Yearly credit turnover</th>
<th>Number of products</th>
<th>Customer total number of sub products</th>
<th>Customer number of cheque accounts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monthly credit turnover</td>
<td>1</td>
<td>0.36959</td>
<td>0.93963</td>
<td>0.93963</td>
<td>0.00058</td>
<td>0.00219</td>
<td></td>
</tr>
<tr>
<td>Monthly credit turnover</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>0.0861</td>
</tr>
<tr>
<td>Last month’s end balance</td>
<td>0.36959</td>
<td>1</td>
<td>0.4356</td>
<td>0.36634</td>
<td>0.00269</td>
<td>0.00246</td>
<td>0.0056</td>
</tr>
<tr>
<td>Last month’s end balance</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>0.0345</td>
<td>0.0539</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>Quarterly credit turnover</td>
<td>0.93963</td>
<td>0.4356</td>
<td>1</td>
<td>0.84021</td>
<td>0.0205</td>
<td>-0.00236</td>
<td>0.00424</td>
</tr>
<tr>
<td>Quarterly credit turnover</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>0.0638</td>
<td>0.0009</td>
<td></td>
</tr>
<tr>
<td>Yearly credit turnover</td>
<td>0.67364</td>
<td>0.36634</td>
<td>0.84021</td>
<td>1</td>
<td>0.04508</td>
<td>-0.01592</td>
<td>0.00802</td>
</tr>
<tr>
<td>Yearly credit turnover</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>Quarterly minimum credit bal amount</td>
<td>0.2138</td>
<td>0.73039</td>
<td>0.2664</td>
<td>0.27156</td>
<td>0.017</td>
<td>-0.01077</td>
<td>0.007</td>
</tr>
<tr>
<td>Quarterly minimum credit bal amount</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>Credit limit amount</td>
<td>-0.01896</td>
<td>0.07327</td>
<td>-0.04674</td>
<td>-0.10774</td>
<td>-0.10171</td>
<td>0.05079</td>
<td>0.0082</td>
</tr>
<tr>
<td>Credit limit amount</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>Customer total number of products</td>
<td>0.00903</td>
<td>0.00269</td>
<td>0.0205</td>
<td>0.04508</td>
<td>1</td>
<td>0.45781</td>
<td>0.03412</td>
</tr>
<tr>
<td>Customer total number of products</td>
<td>&lt;.0001</td>
<td>0.0345</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>Customer total number of sub products</td>
<td>0.00058</td>
<td>0.00246</td>
<td>-0.00236</td>
<td>-0.01592</td>
<td>0.45781</td>
<td>1</td>
<td>0.20874</td>
</tr>
<tr>
<td>Customer total number of sub products</td>
<td>0.6471</td>
<td>0.0539</td>
<td>0.0638</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td></td>
</tr>
</tbody>
</table>