FACULTY OF COMPUTING AND INFORMATION MANAGEMENT

LOGISTICAL REGRESSION MODEL FOR PREDICTING SMALL AND MEDIUM ENTERPRISES' (SMEs) CREDIT RISK FOR COMMERCIAL BANKS IN KENYA.

By

BENARD MURIITHI MURIUKI

MASTER OF SCIENCE IN DATA ANALYTICS

KCA UNIVERSITY

OCTOBER 2021

DECLARATION

I declare that this dissertation is my original work and has not been previously published or submitted elsewhere for award of a degree. I also declare that this contains no material written or published by other people except where due reference is made, and author duly acknowledged.

Student Name: Benard Muriithi Muriuki Reg No. KCA/19/00438

Sign: Date: 08th Oct. 2021

This research project has been presented for examination with my approval as the appointed supervisor.

Digitally signed by Simon N. Mwendia
DN: cn=Simon N. Mwendia, o=KCA University, ou=College of Technology, email=smwendia@kca.ac.ke, c=KE
Date: 2021.10.12 17:32:32 +03'00'

Signed: _____ **Date:** 12/10/2021

DR. SIMON MWENDIA

ABSTRACT
Small and medium enterprises play a very vital role in the growth of any country economy. They provide employment to both the owner(s) and the employee(s). However, their growth has been hampered by lack of capital for expansion and operational expenses. Commercial banks in Kenya provide the main source of funding to these SMEs through the various loan products they offer. As a result of these borrowings, banks have been exposed to default risk which affects their profitability. The purpose of this research is help to generate a predictive model for accessing SMEs probability of default. Convenient sampling was used for selecting the entire population consisting of commercial banks in Kenya thereby utilising the census as opposed to sampling criterion. The study utilises the KDD model to direct the development of the research study. The collected data will be analysed using R software. Additionally, the study used logistic regression to testing the statistical significance of the relationship between variables with a p-value of $P > 0.05$ considered significant, and $P \leq 0.05$ considered not statically significant.

Keywords: *Commercial banks, Credit score, Default risk, Logistic regression, SME's.*

AKNOWLEDGEMENT

I do wish to thank my supervisor Dr. Simon Mwendia for his unwavering support during the entire time of my study period.

To my wife Edith, daughters Mitchell and Jasmine, I salute you for the support you have offered to me during my study period.

Thanks to the Almighty for the grace and courage He has granted me throughout this project.

DEDICATION

To my wife Edith, and my daughters Mitchell and Jasmine, you give me strength to face each day.

ACRONYMS AND ABREVIATIONS

CBK: Central Bank of Kenya

COVID-19: Coronavirus Disease of 2019

CRB: Credit Reference Bureau

EAD: Exposure at Default

LGD: Loss Given Default

OECD: Organisation for Economic Co-operation and Development

PD: Probability of Default

SMEs: Small and Medium Enterprises

Table of Contents

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE
INTRODUCTION

1.1 Background of the Study

Small and Medium Enterprises (SMEs) form the backbone of many countries' economies including Kenya (Wu, 2008). In the OECD countries, 97% of firms are SMEs, and they provide up to 80% of the nation's jobs. The likelihood of defaulting (the probability that a customer will fail to pay the amount borrowed from the banks) is a significant business variable for commercial banks that requires effective assessment of credit risk. The assessment of credit risk is conducted using a credit rating model as banks do not base their lending decision on financial criterion alone. The credit scoring system is a multivariate assessment system that effectively predicts the risk of loan default by weighing the set of defined variables and delivering a single value that shows the risk of default (Mugenda, 2008). As such a low score could be interpreted as a low risk of default that is used to imply that the client is able to service the loan, with a high score indicating a high risk of defaults thus the rejection of the loan request. Apart from individual factors such as cash flow, the commercial banks in Kenya also use the Credit Reference Bureau (CRB) listing to assess the credit risk factor of individuals as well as organisations. Currently in Kenya, the Central Bank of Kenya (CBK) is the Banking sector regulatory body overseeing the conduct and performance of commercial banks, Forex bureaus, and non-bank organisations. The country has 42 commercial banks and one mortgage finance organisation that has two banks namely; Imperial bank and Chase bank (CBK, 2014).

The relevance of using predictive models in assessing eligibility to receive loans was supported by the reduction in the cost of computers, rapid advancement in underwriting technology, and the inefficiency of the house-rate criterion that was implemented through interview procedures in the 1990s. As such, the application of predictive models for the likelihood of default using an automated credit scoring system sets a standardised criterion that saw the banks improve their performing loan rates in the past decade (Samreen and Zaidi, 2012). However, political violence and instability experienced in the year 2007 and 2013 increased the inflation rate as the economic growth of the nation slowed down at a higher rate than its recovery, with the worst-hit on both the nation's economy and loan performance experienced in the year 2020. As a result, banks have been placed in critical situations in determining loan eligibility with increased use of data mining methods in improving the accuracy of the credit scoring system to minimise the risk of non-performing loans. Adem and Waititu (2012) indicate a concern in the steadily increasing level of non-performing loans in Kenya. The conventional

wisdom that SMEs in Kenya are better suited for engaging in relationship lending due to personalised direct contact that offers the chance for soft information gathering has been disputed. Instead, banks are realising that they can gain a comparative advantage on their loans to SMEs using lending technologies as opposed to relationship lending. The aim of this research is to develop a credit scoring model that will evaluate the credit risk factor among SME borrowers in Kenya.

1.2 Statement of the Problem

The country economy is dependent on the performance of SMEs through wealth creation as well as the provision of employment opportunities for local residents. As a result, SMEs play a significant role in the reduction of poverty resulting in a steady growth of the economy and an increase in demand for banking services. SMEs face a significant lack of adequate capital and depend on the credit facility provided by the banks to meet their financial capital demands in terms of production expenses, use of technologies and business growth. Wanjohi and Mugure, (2008) indicates that loans are the main sources of capital for SMEs. Additionally, Waweru and Kalani (2009). indicates that loans are the major source of revenue for commercial banks. As a result, the need to provide performing loans that is beneficial for both the SMEs and banks is paramount in promoting profitability, and competitiveness of both organisations. The lending contract requires the SMEs to pay their loan with interest in a steady way as provided in the amount and duration specifications however, Munene and Guyo (2013) indicate that the liquidity and profitability aspect of the SMEs significantly determines whether the SME will default on its Loan. Loan defaulting is detrimental to both the SME and the bank, as the bank stands to make hefty losses for the default, while SMEs that default is likely to collapse because of poor business performance in terms of profitability, competitiveness and production, with an additional financial debt that might result to a loss of security assets contributing to the 75% rate of collapsing SMEs in Kenya. Hence the need to ensure that banks only approve loans that have a low risk of default (Kipyego & Moses, 2013). As a result, this study seeks to develop a predictive model that will help the commercial banks in Kenya approve loans for SMEs that have financial potential to service their loans to term, consequently maximising financial gains for the SME in question, the bank and the nation at large.

In the year 2016, 8% of total loans borrowed by SMEs were defaulted, supporting the set limit of 4% for non-performing gross loans. Additionally, a survey conducted by the Credit Survey Report by the Central Bank of Kenya indicated that the ratio of non-performing loans to gross loans stand at 9:1 in the year 2017 (CBK, (2017). This study seeks to fill the gap in

literature by providing a model that is suitable to the local context of the Kenyan commercial banks. While researcher such as Goriunov and Venzhyk (2013) effectively applied the logit models and neural networks to successfully develop a predictive model in Ukraine, however, this model is not suitable for the Kenyan environment as it is highly dependent on the credit-base for car loans, mortgages, and retail borrowers that is utilises the existent credit score aspect of the Ukrainian financial systems. Additionally, this system differs from the existing manual-based credit scoring system used in commercial banks that depend on the bank account activity, financial statement in terms of the stability of the sources of income, and Credit Reference Bureau (CBR) rating if available. Despite the development of a number of predictive models for lending to SMEs in Kenya, Kipyego and Moses (2013) indicate that overreliance on scoring methods such as credit histories retrieved from credit bureau has led to unreliable credit scoring leading to an increase in the number of non-performing loans. Lagat, Mugo and Otuya (2013) also point out that current predictive models need improvement that can factor in the additional data available about SMEs in Kenya. In addition, existing predictive models for SMEs take a similar form that measure the investment gap against the financing requitements. However, current research findings do not justify the effectiveness of this approach for predicting credit scoring (Lemay, 2016). This study fille the gap by providing an automated system that considers all variables that influence the probability for default risk

1.3 Main objective

This study intends to develop a predictive model that can be used by Kenyan commercial banks to determine the Credit risk for SMEs.

1.4 Specific Objectives

1. To examine and identify attributes that are considered to determine the credit risk for SMEs Loans in Kenya Commercial banks
2. To develop an appropriate predictive model for the assessment of credit risk.
3. Evaluate the developed model.

1.5 Research Questions/hypothesis

1. What attributes are considered to determine the credit risk of SMEs Loans in Kenya Commercial banks?
2. Which predictive model is most effective in predicting the credit risk for SMEs?
3. How can the developed model be evaluated?

1.6 Significance of the Study

The study will have great importance to commercial banks in Kenya in implementing it as it will reduce the prevalence of non-performing loans. This study will benefit the potential loan applicants in assessing and understanding their capacity to service the loan, hence rejecting loans that would lead to economic problems such as loss of assists. This study will also be beneficial to researchers as they can use the developed concept in coming up with a predictive model to access risks on other clients such as the risk of default on individual loans, retail loans and corporate loans.

1.7. Motivation of the Study

The high prevalence of non-performing loans lent to SMEs has significantly affected the banking business in terms of its profitability. The loan given to SMEs is of central concern because they come in larger amounts as compared to loans lent to individuals. The motivation for choosing SMEs rather than other enterprises is because most SMEs require capital injection to remain competitive and sustainable, additionally, SMEs provide employment to 70% of the Kenyan population of employed people and it is the biggest contributor to the Gross domestic product (GDP). The motivation for choosing banks stems from the need to ensure that banks provide profitable loans to SMEs with capacity to service the loans and save the ones without the financial capacity to service the loan from financial crises that could narrow their survival and sustainability chances.

1.8 Scope of the Study

This study focuses on the development of an effective predictive model for determining default risk among SME's because ineffective assessment of the eligibility of SMEs to access loans leads to lending organisations that are unable to meet their repayment obligations, thereby crippling their business foundation as opposed to supporting them for positive contribution to the country's economy. As such, the management of default risk would improve the economic performance of commercial banks and the country's economy. It also assists banks to improve on turnaround time in the loan application process, as the model can process huge amounts of data more accurately as compared to manual verification which has been a problem in the banking industry for long periods of time.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter provides an empirical and theoretical review of extant research to establish the knowledge gap. The theories informing this study are discussed under the theoretical review with the empirical review discussing past studies based on the research objectives. Through the review of the literature, the relationships between various variables that determine credit risk management are identified.

2.2 Theoretical Review

2.2.1 SMEs and its relationship with commercial banks

The problem of loan defaulting has caused significant loan losses to banks in Kenya (Adem and Waititu, 2012). Even though banks utilise the bulk of information provided by SMEs to assess their ability to adequately service the loans, the number of non-performing loans have kept increasing jeopardising the profitability of the organisation (CBK), 2017).

The increase in non-performing loan ratio is attributed to economic challenges experienced by SMEs that led to the increase in short-term liabilities and nonperforming loans since the year 2020 following the negative economic impact of the Covid-19 pandemic on SMEs with a 6.2% regress in economic performance and a large-scale layoff that saw approximately 1.72 million people losing their jobs due to the pandemic, SMEs in the tourism, real estate, trade, and hospitality industry has faced significant economic challenges that led to their inability to service their loans within the year (Business Daily Africa, 2021). Nonetheless, the CBK (2016) acknowledges the persistence of loan default trends in the Kenyan commercial banks and insists on the need to comply with the set guidelines on handling non-performing loan clients. As a result, the question raised on whether the existing credit system is effective in accurately analysing and appraising the ability of borrowers to service their loan effectively.

Loan default is defined by Samreen and Zaid (2012) as the inability of the borrower to adequately fulfil his/her loan obligation leading to subsequent months of missed payments. Bofondi and Gobbi (2003) also assert that default is the threshold risk marking the point when the borrower misses to pay at least three-monthly instalments within a 24-month loan period. As a result, Commercial banks in Kenya have experienced considerable growth in data as a result of advances in technology and growth in banking products as they utilise loan performance as a

positive signal increasing the availability of credit to various organisations and economic sectors. Thun (2011) identifies interest rates, individual characteristics, business performance, and socioeconomic factors as key determinants of loan performance. Nonetheless, Commercial bank's credit default is still very frequent in financial institutions in Kenya (CBK, 2017). Most commercial banks in Kenya have a considerable portion of their loan book issued to SMEs and the bank's profitability as a result of default risk by these SME's affecting the bank's income due to high provision as a result of an increase in non-performing loans. This has created the need to use more advanced analytics and machine learning techniques to analyse the bank's data related to SMEs to predict whether the customers will default on the loan or not (Schreiner, 2010). As a result, the commercial banks in Kenya need the application of an effective predictive model for accurately assessing the risk of defaults to lower the persisting prevalence of non-performing loans.

2.2.2 Loan defaulting among SMEs

a) Overview of loan defaulting

Most SMEs in the world face challenges in growth more so in access to capital and funding. Deng, Liu and Deng (2016) indicate that the international connectivity of countries' debt to globalisation has given rise to global dynamics that influence macroeconomic fluctuations as well as the returns of financial assets. As a result, the current decade has seen an increase in loan default prevalence across the globe especially in the developed nations such as the United States of America, the United Kingdom and Germany (Calabrese, 2012). The great depression and various economic crises caused by international fluctuations in the global markets such as the rise and drop of fuel significantly affect the pricing and economic performance of individual countries. Commercial banks in Kenya have shied away from lending to SMEs since they lack the ability to meet certain criteria and terms set by the banks. It is well-established that one of the main determinants of economic growth is access to finance (CBK, 2014). Therefore, access to funds is a major determinant of the economic success of SMEs in Kenya. Commercial banks in Kenya have implemented several lending conditions that affect the availability of loan products to SMEs in Kenya, but despite this, the rate of default by SMEs is still very high (CBK, 2016). As of January 2021, Metropol data showed that there was an increase in loan accounts listed by Credit Reference Bureau Africa Limited for being in arrears of more than three months. The numbers had risen from 9.673,258 as of August 2020 to 14,035,718 in January 2021 (Business Daily Africa, 2021). The increase in CRB listings shows the challenges that commercial banks

are experiencing in dealing with SME default risk as a result of economic difficulties being experienced by SMEs in Kenya more so during this time of Coronavirus Pandemic (Business Daily Africa, 2021). As the target group of this study, it is therefore prudent for commercial banks to establish a way of evaluating the creditworthiness of SME.

b) Metrics for measuring loan defaulting in SMEs

One of the most vital processes used by banks in decisions regarding credit risk management is credit scoring. The process included the collection, analysis and classification of a range of credit variables and elements for the assessment of credit decisions. CBK (2017) affirms that credit scoring is considered the core tool applied in the last few decades to appraise various financial institutions. Moreover, Wu (2008) points out that the applicability of credit scoring in areas ranging from accounting to finance makes this method generalizable across different sectors.

Models for credit risk assessments provide some of the most successful methods for modelling research in banking and finance, which reflects in the increasing scoring analysis within the industry (Adrea, 2010). However, the phenomenal consumer credit growth within the last few decades is largely attributed to credit scoring. Thun (2011) acknowledges the increased popularity in using creditworthiness of borrowers by lenders based on the credit histories derived from credit bureaus, as well as assessing the salary of borrowers before approving loans. The excellent facilities in developed countries have led to well-established credit scoring, which is lacking in developing countries due to the availability of less information, as well as facilities (Bofondi and Gobbi, 2003).

According to Dastoori and Mansouri (2013), credit risk management is a strategic process used to assess the risk of default in loan repayment to determine the creditworthiness of borrowers. According to Schreiner (2010), this assessment is based on various factors such as age, assets, history of repayment, income, employment and outstanding debt. However, Pompe and Bilderbe (2005) indicate that the manual process applied by banks is cumbersome and presents a myriad of problems. The problem with the manual process includes the need for skilled operators capable of calculating scores manually, which results in increased administrative costs. In addition, Deng et al. (2016) indicate that the associated high costs force some lenders to abort the manual systems and primarily rely on the business judgment employed by their lending officers to assess the creditworthiness of applicants. Therefore, lending financial institutions opted to use credit scoring to reduce costs and improve the process of credit collection and analysis, as well as to reduce costs of credit analysis. Yap, Ong and Husain (2011)

indicates that both quantitative and qualitative credit assessment models are used. However, Adem and Waititu (2012) indicate that the subjective and judgemental nature of qualitative methods is disadvantageous due to the lack of an objective basis over which decisions are made regarding the default risk of borrowers. Therefore, more inclination towards quantitative methods is observed due to the systematic method used for categorization of non-performing and performing loans, which improves accuracy and reliability of assessing creditworthiness.

2.2.3 Attributes that influences SMEs loan defaulting

Business Daily Africa (2021) indicate that credit risk should be assessed based on the perceived loss related to lending the loan, as well as the ability of the loanee to meet his/her contractual obligations. Lanzarini, Villa Monte, Bariviera, and Jimbo (2017) indicate that credit risk will be assumed to be a function of the 5Cs, which includes Collateral, Capital, Conditions, Capacity and Character are considered vital for measuring creditworthiness.

a) Character: Adem and Waititu (2012) indicate that character is a weighting technique that represents the loan applicant's average of several attributes. Lemay (2016) indicates that the creditworthiness of a loan applicant can be examined in terms of the quantified total weighted score of customer-related factors in categories that include personal, cultural, social and economic categories. The typical measurement of the social aspect of the customer involves the customer's lifestyle, which include how the customer lives, entertainment and consumption trends and reference circle (Yang, Zhang and Zhang, 2009). Particularly, credit institutions look at the reference group of the customer, which is considered influential on the creditworthiness. Property ownership and relative ownership in the reference groups can be used to evaluate economic factors while personal factors include occupation, economic standing, personality, and family standing (Adrea, 2010).

b) Capacity: Credit institutions consider the capacity to pay including the cash flow derived from the customer or business, previous loan repayment processes credit repaying frequency (Adem and Waititu, 2012). Therefore, in assessing the capacity for loan repayment, credit institutions can assess the financial ratios of the borrower (Beaver, 1966).

c) Collateral: Yang et al. (2009) define collateral and the assets pledged by the borrower, which can be used an alternative resource for loan repayment. It mainly includes tangible

capital like farming and manufacturing equipment, office equipment and furniture and forms of real estate such as land. The lifetime value of collateral should correspond to the length of the period of loan repayment for credit institutions to consider it (Munene and Guyo, 2013).

d) Capital: This factor represents the invested amount of money by the business or individual. It represents the incurred risk by the borrower should the business fail (Adrea, 2010).

e) Condition: The borrower' sensitivity to external forces represents their condition. Borrowers have different sensitivity levels to forces such as inflation rates, interest rates, competition pressure and business cycle. Conditions are measured by the vulnerability of the customer to these external factors (Lemay, 2016).

2.2.4 Overview of Probability theory

Probability is part of mathematics that branches out to analyse the concept of random phenomena (Calabrese, 2012). The probability theory indicates that while the outcome of a random phenomenon cannot be accurately determined, it can be categorized into a set of alternative outcomes (Yang et al., 2009). According to Deng et al. (2016), the actual results of the events are considered to be based on chance only. As a result, probability can be defined as a relative frequency as provided in a game of coins where a person faces equal chance of getting heads or tails, a game of dice, where there are six possible outcomes, or a game of cards, among other examples. The main aspect of the probability based on the relative frequencies is that the actual outcome cannot be predicted but all the possible outcomes are known with certainty (Hilbe, 2009). However, Dastoori and Mansouri (2013) indicate that probability and statistics focuses on the laws of random events that entail the collection, interpretation, analysis, and display of numerical data that present real-life values such as possibility of occurrence and relationship between variables

2.2.5. Probabilistic prediction data Analytics Techniques

Predictive analytics utilise statistical techniques to analyse data for relationships and trends between variables to predict the outcome of specific events when certain conditions are held constant or in combination (Adem and Waititu, 2012). Predictive models' outcome depends on the associated algorithm that provides outcome based on the relationship between variables (Dastoori and Mansouri, 2013). Some of the effective analytic techniques include:

- **Classification techniques**

The aim of these techniques is to describe a range of predefined classes and classifying a data item into one of these classes (Vitek, 2014). Classification tasks are performed using these techniques by establishing a set of patterns for predicting the class of objects with unknown class labels. Classification techniques include decision tree techniques, IF-THEN rules, Generic Algorithm and Random Forest, among others (Lemay, 2016).

- **Regression Techniques**

Regression refers to the process used to identify patterns and calculate continuous outcomes predictions. The aim of these techniques is to establish the relationship between the dependent and independent variables. The techniques are mostly used to forecast, establish the cause-effect relationships and the time series models (Aladag and Eğrioğlu, 2012). Regression is best applicable in studying relationships like the link between speed driving and the frequency of road accidents. Examples of regression techniques include logistic, linear, ridge, lasso and stepwise regression techniques.

Logistic regression is an efficient multivariate regression approach used for credit assessment. The dependent variable is used as a binary (1 or 0) variable, which marks 1 where the borrower defaults within the period of observation and 0 where the client does not default (Aladag and Eğrioğlu, 2012). Credit risk potential parameters include all the independent variable. As opposed to discriminant analysis, logistic regression present benefits such as no requirements for normally distributed input variables, which allows the inclusion of qualitative creditworthiness variables, and direct interpretation of the probability of default (Aladag and Eğrioğlu, 2012).

- **Unsupervised Learning Algorithms**

These techniques are often used in exploring customer information to inform adjustment of services accordingly. The techniques are also applicable in developing more efficient targeting strategies for advertising content. Through these techniques, the dataset can be split automatically into groups based on their similarities (Hastie et al. 2009).

2.3 Empirical Review

Reviewing extant literature indicates a significant impact of credit scoring methods on consumer lending. Chaudhary (2003) used a logistic regression model in Pakistan to predict the risk of default. This study indicated that the most significant variables are level of education (higher

level of education), nature of business such as non-farm business, gender (female clients) increases the likelihood of payment of loan, while subsidies on interest rate does not influence the risk of default. This model is inconsistent with the Kenyan environment where men and agriculture-based business have a steady income stream that could be linked with high chance of loan repayment.

On the other hand, Berhanu and Fufa, (2008) applied the legit regression model to predict the risk of loan default among Ethiopian Small-scale farmers. This study indicated that farmers with large heard of livestock, bigger farms, alternative sources of income, use agricultural-related technologies and those located in areas that record high amounts of rainfall have higher likelihood to repay their loan. This model cannot be applied to assess the risk of default in SMEs as they are not all based on agriculture-related businesses.

A comparative analysis of developed credit scoring techniques with manual statistical credit scoring techniques including discriminative analysis and logistics regression indicated that developed scoring models present increased accuracy and reduced errors compared to the manual methods (Pompe and Bilderbe, 2005). Moreover, an analysis of the financial performance of banks based on credit risks management by Mugenda (2008) established a significant impact of credit risks management on banks' profitability. Similarly, an investigation of the performance of bank credit policy in Rwanda by Adem and Waititu (2012), indicates that increasing the spread of average interest rate, as well as the margins of interest rate results in high ineffective and poor competition. The results were replicated by Yap et al. (2011) who established that poor credit policy and terms, credit analysis, lending, and credit appraisal correlated with loan performance and credit risk control hence the need to use standard sources of information to assess borrowers' default (see Fig. 1).

FIGURE 2.3

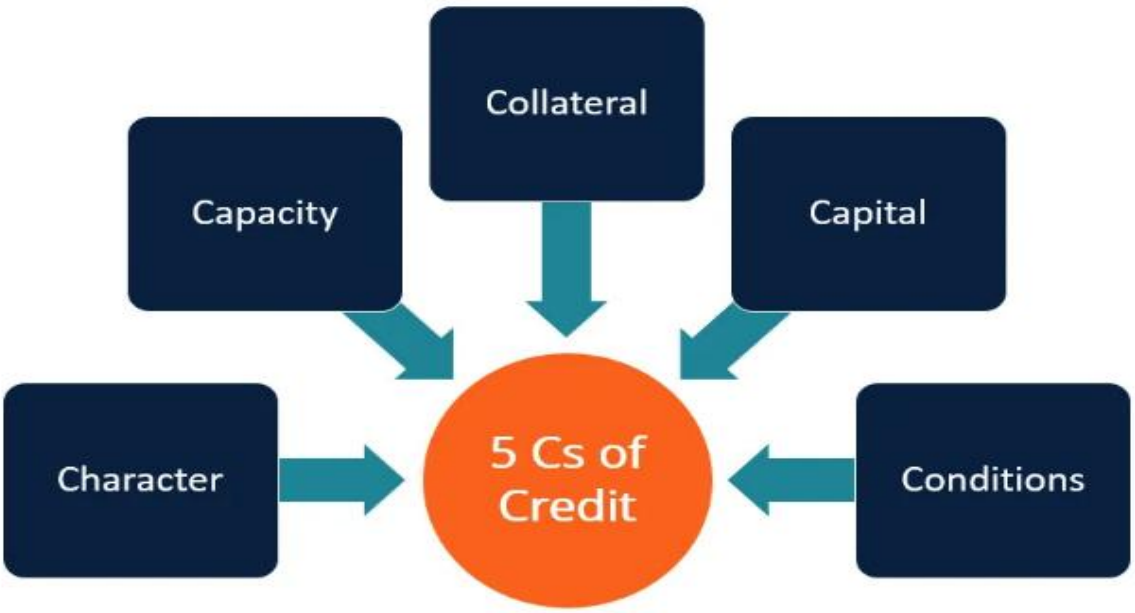
Standard Sources of Information for Assessing the Risk of Borrowers Default

Source: Yap et al. (2011)

A case study approach by Deng et al. (2016) to investigate the influence of credit policy on the profitability of a bank also established minimisation of bad debts incidences as a result of implementing good credit policy. Einav et al. (2013) established that the main technique used by microfinance institutions that are known for their focus on SMEs is the 5Cs technique, which includes Collateral, Capital, Conditions, Capacity and Character for credit risk management (Lagat et al., 2013). The prevalent use of credit matrix was established in measuring default risk and credit management in the majority of microfinance institutions. A further study on the loan performance of these institutions by Calabrese (2012) indicates that the use of the 5Cs technique for client appraisal was effective which resulted in more emphasis on these elements.

FIGURE 1.3.1

5Cs Technique



Source: Lagat et al. (2013)

The review of extant literature dealing with automation of credit scoring systems and their relationship to loan performance indicates that this subject has attracted the attention of many researchers with several researchers indicating the ability of credit assessment tools to categorize loan applications into low-risk and high-risk applicants. In addition, some of these studies have ventured into assessing the rate of accuracy in predicting the risk of loan repayment between credit assessment tools and manual methods. However, most of the evidence retrieved is based on studies performed in developed countries, which makes the generalisation of these findings in Kenya challenging.

Research Gap

While there is numerous predictive and automated credit scoring system using various statistical techniques ranging from multiple regression, machine learning, to logistic regression models, Waweru and Kalani (2009) show a low uptake and application of these models in the Kenyan commercial banks is attributed to the lack of structured scoring model that for best practice in credit scoring. This study seeks to fill the gap by providing a structured model that reflects the local economic environment to encourage banks to adopt the use of automated probabilistic

predictive model in determining the risk of default for optimal decision on loan eligibility. Many scholars have explored the probability prediction models of credit default in banks and proposed some methods for forecasting, which present various defects and restrictions that this study seeks to address. The findings by Adrea (2010) indicated the implementation of an automated credit scoring technology in financial institutions leads to increased profitability; however, the study failed to provide a predictive model for assessing credit risk factors. In addition, Lagat et al. (2013) indicated that proper risk assessment leads to the deployment of stringent requirements on a down payment for high-risk loan applicants, as well as expansion of lending to low-risk borrowers, however; the impact of loan default on the banks was not mentioned. A study conducted by Dastoori and Mansouri (2013) also established benefits associated with automation of credit risk scoring technologies including targeting generous local to low-risk borrowers increased ability to screen high-risk loan applicants, however, this study did not focus on SMEs, as it was more directed towards loans given to individuals. The application of a single-factor method of financial ratios for analysis of credit financial performance of enterprises was proposed by Beaver (1966) which Pompe and Bilderbe (2005), used multivariate discriminant analysis to construct a default predictive model. On the other hand, Yang et al. (2009) employed Regression to establish a probability prediction model of credit default for listed companies which resulted in the identification of the most significant corporate financial indicators. Notably, these credit default prediction models were mostly applicable to large companies hence cannot be generalised on SMEs due to the different parameters under which these SMEs operate. However, the results point to the potential benefits of developing predictive models through a data mining classification algorithm to establish a classification model for new loan applicants. This approach enables the identification of the risk customers hence reducing the risk of default on loan repayment. Therefore, despite the development of probability prediction models for large companies by many scholars, the review of literature indicates that such models have not been developed for SMEs, which can be evaluated in the Kenyan context.

This project employs the logistic regression method as a suitable classification algorithm for analysis of the risk of credit default of SMEs in Kenya to provide an effective approach for loan approval by banks.

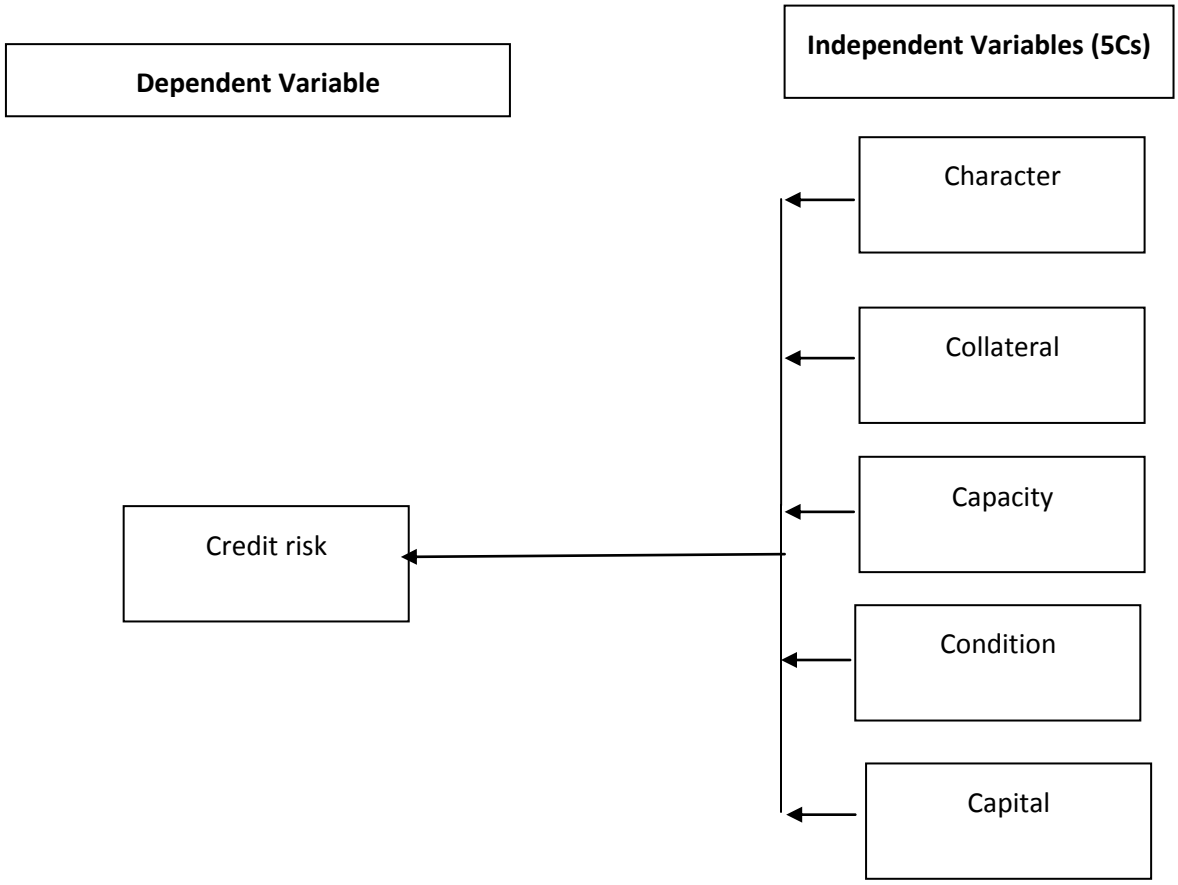
2.4 Conceptual Framework

Based on the reviewed literature, a conceptual framework was identified as stated below. The framework illustrates the link between dependent and independent variables employed in this

study. The independent variables identified include the 5Cs, which include Capacity, Character, Collateral, Condition and Capital, which are useful to determine the dependent variable, which is the credit risk of the customer.

FIGURE 2.4

Independent and Dependent Variables



Source: Author (2021)

2.5 Operationalisation of Variables

The identified independent variables and their explanations are provided in the table 2.5.

TABLE 2.5

Operationalisation of Variables

Variables	Indicator measuring variable	Data to be collected for the indicator
Conditions	Level of sensitivity to external forces	Rates of inflation and interest rates

Collateral	Pledged assets	Farming equipment, manufacturing equipment, office furniture, office equipment, real estate.
Capacity	Capacity to repay loan	Cash flow, credit repaying frequency, process of previous loan repayment
Character	Loan applicant's average of social, cultural, economic, and personal attributes	Consumption per month, financial standing of reference circle, property ownership, relative ownership, occupation, personality, economic standing
Capital	The amount the borrower stands to lose should the business fail	Amount invested

Source: Author (2021)

2.6 Summary

The review of literature recognised the development of various credit scoring tools and predictive models for the performance of bank loans. Potential variables for assessing credit risk for SMEs are identified and used to develop the conceptual framework on which the predictive model in this project is designed.

TABLE 2.6

Literature Matrix

Literature Matrix		
Author and Date	Model used	Variables
Chaudhary (2003)	Logistic Regression Model	Education, gender, nature of business, purpose of loan significantly influences risk of default.
Pompe and Bilderbe (2005)		
Berhanu and Fufa, (2008)	Multiple Regression Model	Size of the farm, number of livestock, climate, use of technology, and alternative sources of income contributes to loan repayment
Yap et al. (2011)	logistic regression model and decision tree model were	Credit Repayment history data is a significant determinant of credit risk with equal classification error rate obtained from both methods.
Goriunov & Venzhyk (2013)	logit models and neural networks	Loan repayment history, and frequency of loan is a significant determinant of credit risk

Lagat et al. (2013)		5c's for credit risk management
Adem and Waititu (2012)		indicates that increasing the spread of average interest rate, as well as the margins of interest rate results in high ineffective and poor competition

Source: Author (2021)

CHAPTER THREE
METHODOLOGY

3.1 Introduction

This chapter addressed the research method employed to achieve the study objectives. Covered parts in this chapter include the research design used, target population, sampling techniques and sample size, research instruments and their reliability and validity, data collection process, processing and analysis and research schedule and budget.

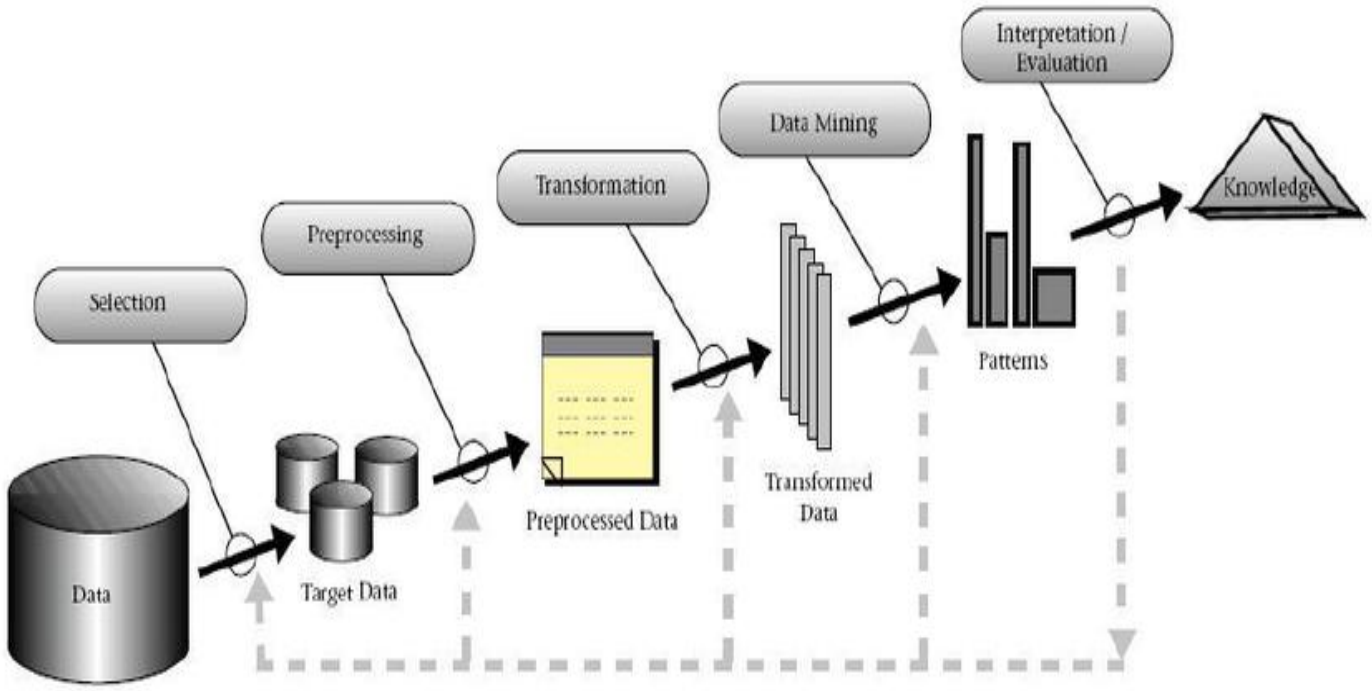
3.2 Data collection method

This study used secondary data collected from bank records and annual reports of banks that provide statistics of SMEs loan default in different sectors. The data was collected through the application of data mining techniques used to extract usable knowledge from the bulk of data contained in the 42 banks' databases by applying the most suitable Knowledge Discovery Processes (KDP) techniques. KDP is a process in which databases are used as tool for retrieval of information and develop applicable knowledge by combining the KDP with selection, subsampling, pre-processing, and transformational techniques. The combination of KDP and the named processes enables the researcher to adequately identify pattern and this study applied the Knowledge Discovery in Databases (KDD) model proposed by Fayyad et al. (1996) as an effective tool for to extract usable knowledge form the bulk of data contained in the banks' databases. The model was modifying to sit the requirement of this study. The KDD has the five steps starting with the data selection, data prepossessing, data transformation, data mining and the interpretation /evaluation is the last stage of the KDD model. The data selection stage has to sub steps, the first entails the development and understanding of domain of the application, and the second sub step entail the development of a target data set from the large data store. The pre-processing process ensured that the vales are consistent through the elimination of missing data issues. This second step entails the data cleaning process that yield pre-processed and cleaned data. At the third stage of data transformation, the data is further processed by identifying useful attributes and the researcher applies data transformation and dimension reduction methods. At the data mining stage there are three sub steps starting with the choosing data mining task, the second entails selecting the data mining method matching it to data mining task. In this case, the researcher used the regression technique to identify the relationship between the dependent and

independent variables. The last step entails applying the selected algorithm to develop patterns and models by interpreting and visualizing the mined patterns.

FIGURE 3.2

Data Collection Process



Source: Li et al. (2016)

3.3. Research Design

This study used descriptive and analytical research designs, which is considered an effective method for studies that are based on human behaviour (Parylo, 2012). The descriptive design is effective in the collection of data that answered the question on how to successful develop and use a credit scoring model for predicting the risk of default for SME. While the analytical research design focuses on drawing of relationships between variables to show patterns and develop a basis for a predictive model (Balcaen and Ooghe, 2004). Hence, the study utilized a quantitative research design. Data mining technique was applied in the collection of data from historical data in the banks' database. According to Bernard (2013), the application of the KDD methodology was adopted in this study as demonstrated in the table below.

TABLE 3.3

KDD model, Adopted from Fayyad et al. (1996) Study

Research tasks	Objectives	Applied research techniques	Indicators
i. Data selection	To examine and identify attributes that are considered to determine the credit risk for SMEs Loans in Kenya Commercial banks	-Developing and understanding domain, creating a dataset -Convenience sampling method to choose data sources like banks - Random sampling method by banks to select 5 SMEs with performing and non-performing loans	-Retrieval of raw data -Both the banks and the SMEs were assigned numerical means of identification such as Bank 1 to Bank 42, and Organisation 1 to 5 for each bank.
ii. Data pre-processing	To examine and identify attributes that are considered to determine the credit risk for SMEs Loans in Kenya Commercial banks	Application of pe-processed algorithm	Cleaned, pre-processed data
iii. Data transformation	To examine and identify attributes that are considered to determine the credit risk for SMEs Loans in Kenya Commercial banks	Transformation methods and dimension-reduction methods	Transformed data
iv. Data mining	To develop an appropriate predictive model that uses the identified attributes to predict SMEs credit risk.	Regression methods, and Data mining algorithm	Models and pattens
v. Evaluation	Evaluate the developed model	Documenting and reporting	Knowledge

Source: Fayyad et al. (1996)

3.4 Target Population

The target population is the Commercial Banks in Kenya. Adopting a census study entailed considering all the 42 banks for the study with the implementation of effective consent seeking procedure to lower the risk of non-response. The census approach is considered a feasible method for a small target population such as the 42 commercial banks.

3.5 Research Instruments

The KDD was used as research interment to extract historical data from the databases of banks. The researcher selected data on financial standing, loan repayment and probability of default of 5 SMEs with performing and non-performing loans in the respective banks. These attributes include Collateral, Capital, Conditions, Capacity and Character (Balcaen and Ooghe, 2004).

The researcher also assumed that the credit risk is a function of Expected Loss and Actual loss. Important variables that were put into consideration. The credit risk was assumed to be a function of the expected and actual loss generalized in the equation below:

$$\text{Credit risk} = \max\{\text{Actual Loss} - \text{Expected Loss}, 0\} \quad (1)$$

In this case, the actual loss represents the loss experienced by the banks, while the expected loss is an estimate that can be represented in the following equation;

$$\text{Expected Loss} = \text{Exposure at Default} \times \text{Probability of Default} \times \text{Loss Given Default} \quad (2)$$

As a result, the credit risk presents the risk of the actual loss exceeding the expected loss (Hastie et al., 2009).

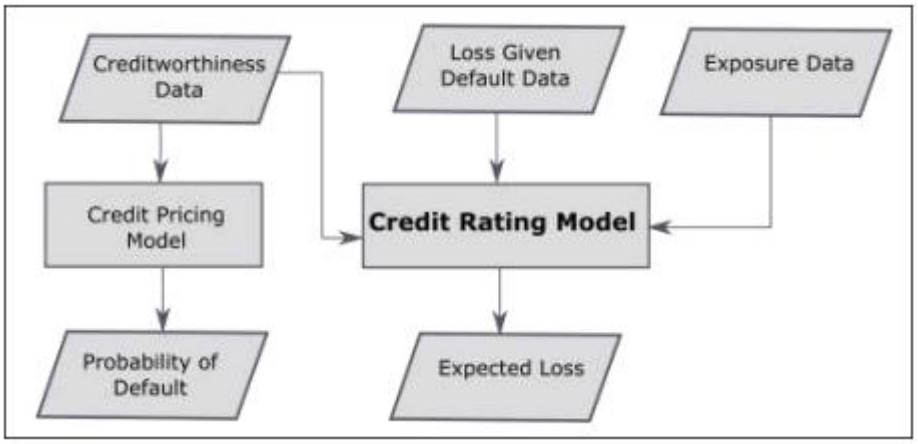
3.5 Reliability and Validity of the Research Instrument

A pilot study and consultation with the experts was used to uphold the validity and validity of the study. A pilot study on 10% (21 SMEs) of the sample (210 SMEs). The use of Cronbach Reliability coefficient was applied in testing the reliability of the data. The researcher also consulted with 3 experts to verify both the data obtained from SMEs are consistent with the defined research objectives process and that the study is not subject to ambiguity. Where the EAD is the amount that the loanee owes the bank, which does not have to be the amount equal to what was borrowed as terms and conditions of borrowing are legally binding. Additionally, the loss given default (LGD) is a percentage of the Actual loss of the EAD that affects the bank. Due to the resultant liabilities, commercial banks tend to hold collaterals to protect themselves from

financial losses. Hence credit rating model that is used as an effective predictive model combines the PDs, EADs and LGDs values and demonstrated in Figure 3.5.

FIGURE 3.5

Credit Rating Model



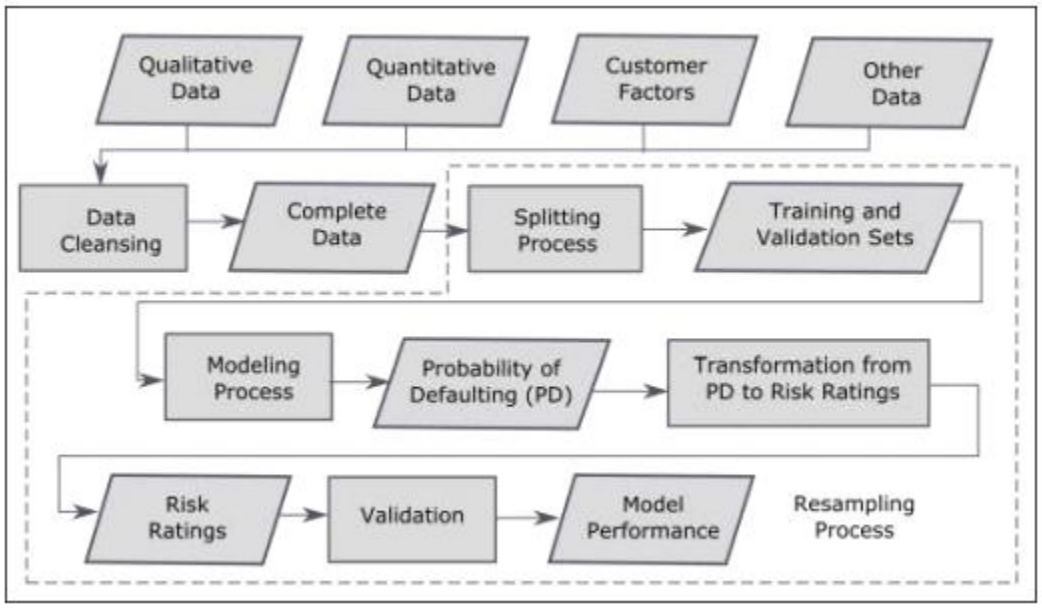
Source: Dastoori and Mansouri (2013)

3.8 Data Processing and Modelling

For equal assessment of all credit risks, the data was categorised into five business sectors; trade, service, real estate, production, and transport industries. According to Chen et al. (2011) real estate category focused on client borrowing money for buying properties and buildings, while trade consisting of business borrowing money to start or expand their venture. Services entail clients dealing with non-farming and non-manufacturing businesses such as IT, while production consists of farmers and manufacturers. On the other hand, transportation entails the borrowing money to invest on start a transport business. The collected data were processed and modelled using a variety of steps that changed the data from quantitative data of customer factors to probabilities of default and credit risk rating as shown in figure 3.8.

FIGURE 3.8

Data Processing Model



Source: Deng et al. (2016)

The credit risk is the outcome of the credit risk scoring model that is a function of the 5Cs, which includes Collateral, Capital, Conditions, Capacity and Character. The given the sample size of 210 (42 banks X 5 SMEs= 210) for fitting a 5 variables model, the obtained data was split into the validation data and training data sets, with 80% of the total data used as a training dataset for fitting models, the 20% of the data used as a validation dataset to estimate the error of prediction for the selected model, and the other 25% used to tests the data set for assessing the error of the final model (Hosmer, 2013).

The data analysis generated both inferential and descriptive statistics, with the descriptive statistics providing measures of relative frequencies, central tendencies, and variability measures. The central tendency provided information on the mean, the variability measure focused on

standard deviation and the frequency distribution tables provided information of relative frequencies. The inferential statistics was generated from the logistic regression model using R statistical tools (Hilbe, 2009). The logistic regression model used the following equation;

$$P(X) = 1/(1+\exp\text{-}z)) \quad (3)$$

The p(X) is the probability of loan default, and z is the depicted as:

$$z = b_o + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 \quad (4)$$

And;

x_1 = capacity

x_2 = character

x_3 = capital

x_4 = collateral

x_5 = condition

b_o = is the b_o the Value of Y, where X=0 (the intercept)

$b_{1\text{-}n}$ = denotes the regression coefficient which represent the change in Y when X changes

Hence, this logistic regression model used the Collateral, Capital, Conditions, Capacity and Character variables to generate the credit risk. The logistic regression is applicable for predicting the credit risk of SMEs within a given period with a high level of accuracy over other credit scoring methods as established by Chaudhary (2003). The model is effective for predicting the risk of credit default as a function of various variables, which are measured before the occurrence of the event (Chaudhary, 2003)

The conceptual framework chapter presents the relation between the independent and dependent variables. Hence the 5Cs was used to assess the credit worthiness of SMEs as follows;

Character: The character is the customers' social economic, cultural, and personal factors. These factors include the existence of a reference group, ownership of property, and ownership of property within a group. The age, personality, gender, occupation, family and economic standing are also significant personal factors.

Capacity: The cashflow from the SME, prevision loan repayment history, and frequency of loan repayment can be used to assess for the capacity to pay back the borrowed money. Adem and Waititu (2012) indicate that the use of financial ratio is useful in assessing the clients' capacity to meet all his financial obligation and the additional expenses incurred by the loan.

Collateral: Collateral are the tangible properties and capitals that the customer used as a loan security for se as an alternative payment of the loan if they fail to meet the loan repayment conditions the collateral lifetime value should be consistent to the length of the loan period hence properties such as land, farming equipment, and other forms of real estate asserts are commonly used as collateral.

Capital: capital is the money invested by the client in the SME capital is a significant indicator of a credit risk. It denotes the risk that the client faces in case the business fails.

Condition: condition is the measure of the level of sensitivity that the borrower is exposed in terms of the effect of external forces such as completion, interest rates, business cycle and inflation rates, on the business. Condition is the measure of the vulnerability of the customer to external forces.

These study variables are applied to assess the credit worthiness of SMEs loan applicant. In this case, the variables were assessed in terms of the ability to predict whether the loan applicant presents a default risk. This is conducted by analysing whether each of the defined 5Cs variables provide a statistically significant effect on the ability of SMEs to repay their loan.

CHAPTER FOUR

DATA ANALYSIS, FINDINGS AND DISCUSSION

4.1 Introduction

The main findings of this study are presented in this chapter. The chapter begins by declaring the results obtained from the summary descriptive statistics. The statistical and mathematical procedure results was then be employed in the analysis of the probability of default, as well as the creditworthiness of customers. The chapter will present results based on various validation methods. The performance of the most significant variables will also be discussed.

4.2 Descriptive Statistics

The sample size of 210 was drawn covering a one-year period of between December 2019 and December 2020 in which 80% of the sample was applied in the development of the predictive model, while the 20% was used for validation. The SME were categories in the form of industry of business. The obtained observations were described in terms of their frequency that is the number of observations in each industry, and the defaulting character in each of the respective industries. The distribution is shown in the table below;

TABLE 4.2.1

Distributions of SMEs across industries and the Defaulting behaviour

Industry	Frequency (number of observations)	Default frequency	Rate of default %
Transport	52 =24.76%	13= 23.64%	25%
Real estate	60 =28.57%	25= 45.45	41.67%
Service	43 =20.48%	9= 16.36	20.93%
Production	30 =14.29%	6=10.91%	20%
Trade	25=11.9%	2=3.64%	8%
Total	210=100%	55=100%	26.19%

Source: Author (2021)

The findings of this table show that the trade industry has the least rate of default followed by the production industry. However, the real estate industry has the highest default rate followed by the transport industry.

TABLE 4.2.2
The Correlation Matrix

##		Default Risk	Collateral	Condition	Capital	Capacity	character
##	Default Risk	1.0	0.85	0.73	0.62	0.61	0.66
##	Collateral	0.85	1.00	0.79	-0.71	0.89	-0.43
##	Condition	0.73	0.79	1.0	-0.45	0.66	-0.71
##	Capital	0.62	-0.71	-0.45	1.0	-0.71	0.09
##	Capacity	0.61	0.89	0.66	-0.71	1.0	-0.17
##	Character	0.66	-0.43	-0.71	0.09	-0.17	1.0

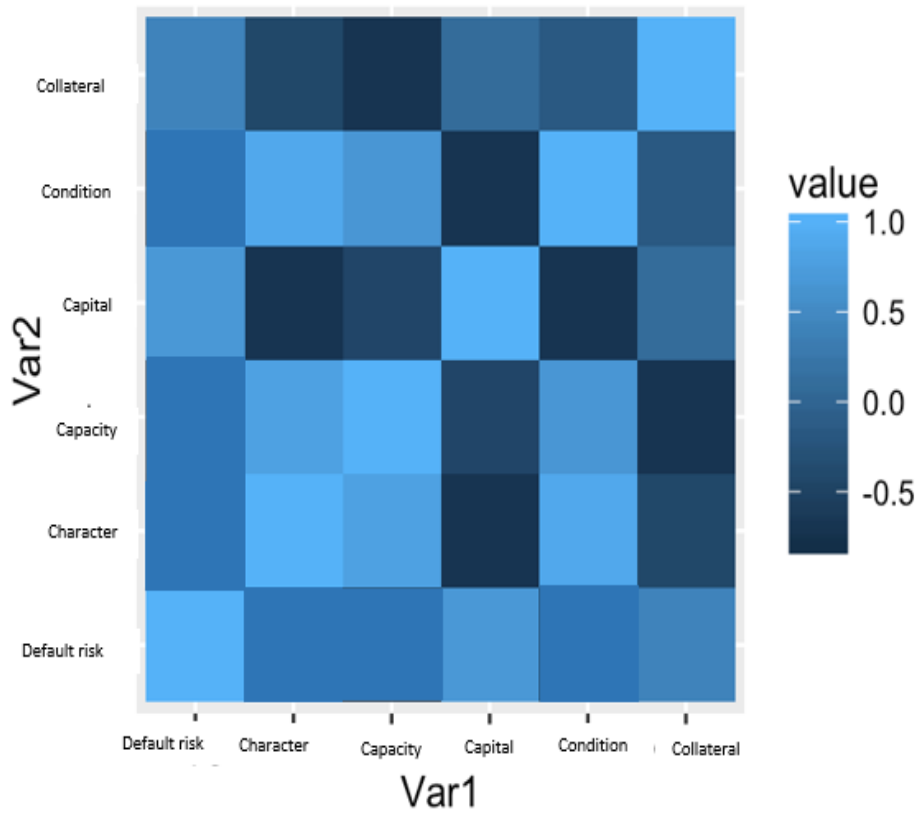
Source: Author (2016)

Table 4.2.3
VISUALISATION OF THE CORRELATION MATRIX

##	Var1	Var2	Value
##	Default risk	Deafault risk	1.0
##	Collateral	Default risk	0.85
##	Condition	Default risk	0.73
##	Capital	Default risk	0.64
##	Capacity	Default risk	0.61
	Character		0.66

Source: Author (2016)

FIGURE 4.2
Correlation Heatmap



Source: Author (2016)

On the correlation heatmap, the light blue colours shows a strong correlation, whith the stongest correlation attaining a avale of 1. Houwerver, the darkening of the colour scemes indicates a reduction in the correllation bewteen variables. The correlation matrix below shows that collateral and capital have a high positive correlation with the probability of default. Additionally, character capacity, and condition are also positively correlated with default risk even though not as high as colletaral and capital. As a result, the correlation analysis shows that

all variables are significantly correlated with the probability of default and they are all included in the predictive model.

4.3 Research Findings

4.3.1 Attributes that define Credit Risk of SMEs

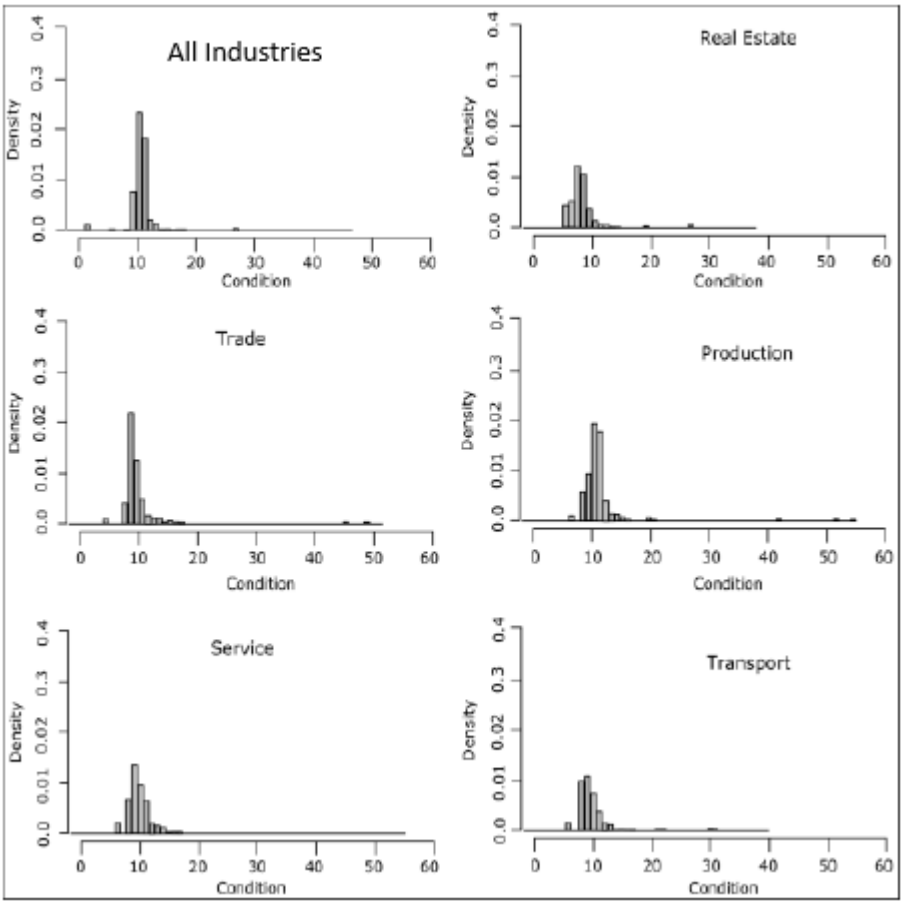
The defined 5Cs are analysed to assess the credit risk of organisations in order to determine whether the organisations re credit worthy. As a result, the firms are assessed based on the following attributes; condition, character, collateral, capital, and capacity.

4.3.1.1 Condition

The condition factors used to assess the credit risk of firms are based of external factors that affect organisations and these factors are used as indicators of vulnerability of firms. The market condition in terms of inflation, and competition, competitiveness of the organisation and inflation can significantly affect the ability of organisation to repay their loans, hence firms with higher consumer base are considered to have lower risk of default than firms with small client base. Additionally, inflation affects all firms, but others perform better than others during this period. From the figure below, the histogram shows distribution on the lower side for all industries indicating that all industries are affected equally by the market conditions.

FIGURE 4.3.1.1

Diagrammatic representation of SMEs condition across industries using Histograms



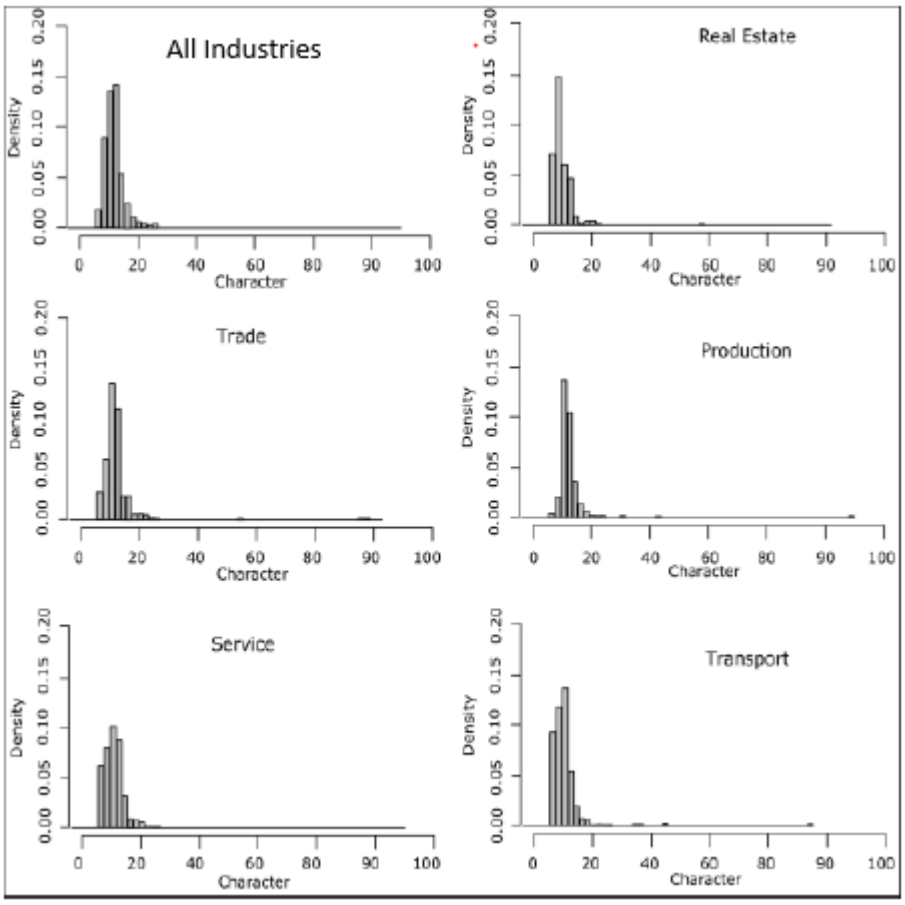
Source: Author (2016)

4.3.1.2 Character

The assessment of the organisational character was based on personal, social, cultural and economic factors to assess its credit worthiness based on the identified reference group. The economic factors included business ownership; the personal factors entail the size and age of the organisation, while socio cultural factors are exhibited in terms of the organisational culture, employee satisfaction influencing the turnover rate, and the quality of customer management strategies of the firms. Based on the findings represented on the figure below, all industries exhibit a normal distribution, however the real estate sector shows a left-sided skewed distribution showing a low score on character in this sector. Nonetheless, both trade and the production industries show a balanced distribution indicating an average score on character for most organisations. However, the transport industry exhibits a skewed distribution on the left tail indicating a low score on character.

FIGURE 4.3.1.2

Diagrammatic representation of SMEs Character values across Industries using Histograms



Source: Author (2016)

4.3.1.3 Collateral

In these findings, obtaining collateral estimates, the organisational assets that are pledged as alternative payment resources for the loans are considered. The collateral estimates were obtained for all industries as well as for each of the individual industry. In this case, when an organisation has a liability value of 0, it is awarded 1000 as the collateral value, is the liability is other values that are not 0, then it is given a collateral value as a fraction of 1000. The findings in the table below show that the real estate industry has the lowest average collateral values, and the service industry has the highest average collateral value. Additionally, the production industry has the highest collateral ratio.

TABLE 4.3.1.3

A Summary of Collateral values for each industry and the score for combing all Industries

Industry	All industries	Real Estate	Trade	Production	Service	Transport
Minimum	-0.42	-o.36	-0.02	-0.02	0.00	-0.006
Average	1.33	0.97	1.48	2.28	1.51	1.17
Maximum	20.13	19.83	50.28	60.70	26.13	11.41
Collateral Value (1000)	0.75%	0.56%	0.34%	2.00%	0.43%	0.17%

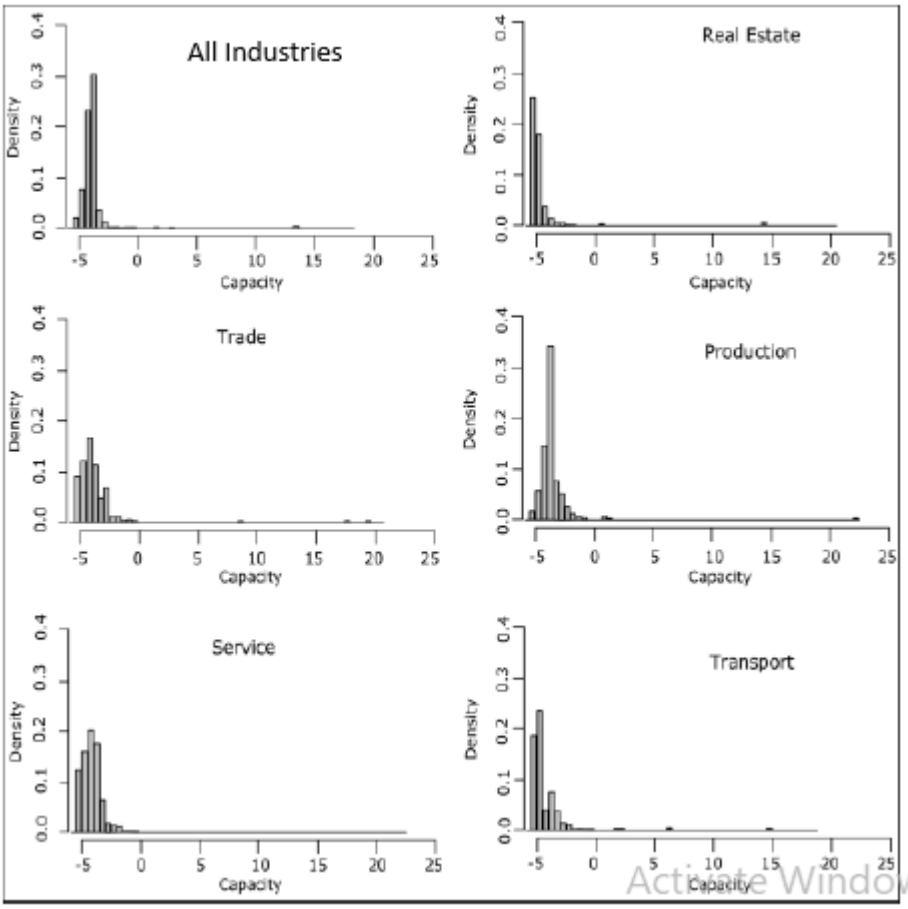
Source: Author (2016)

4.3.1.4 Capacity

In assessing capacity values, the cashflow and the success in paying previous loans is assessed. The capacity measures for all the five industries is skewed to the left with a fraction of the SMEs scoring a negative capacity value indicating low or negative cashflow and nonperforming loans. While observing the SMEs by industries, the transport, service, and real estate industries have skewed distribution at the left tail with some firms scoring negative values, indicating lack of sufficient resources to repay their loans. Hence, these findings show that most of the firms across all industries scored low on capacity estimates as demonstrated in the figure below.

FIGURE 4.3.1.4

Diagrammatic representation of SMEs Capacity values across Industries using Histograms



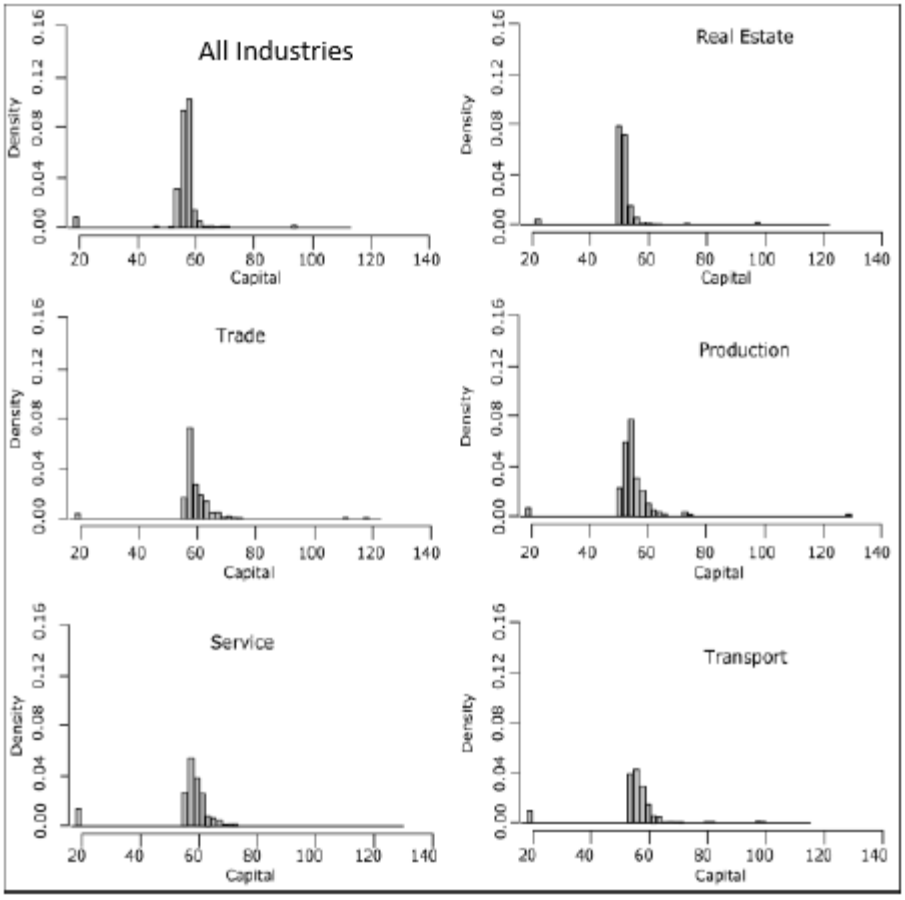
Source: Author (2016)

4.3.1.5 Capital

The capital estimate is assessed in terms of the capital size and the associated risk. The industries that require high capital are associated with greater risk of default, while those with low capital demand are associated with low risk of default. The findings of this study showed that apart from the real estate and the transport industry that require high capital investments, other industries are distributed at the middle showing moderate level of the likelihood of default. Hence the real estate and the transport industries show a high capital demand indicating a low level of credit worthiness due to the perceived risk of default.

FIGURE 4.3.1.5

Diagrammatic representation of SMEs Capital values across Industries using Histograms



Source: Author (2016)

4.3.2 The Predictive Model For Credit Risk: Logistic Regression Model

The findings of the predictive regression model indicated that the most significant variable is high condition with a p-value of 0.010 and the odd ration of 1.266 indicating a low probability of default. High capital was also associated with a low probability of default at an odd ratio of 0.653 and a p-value 0.013. As indicated in the table below, the other significant variable is the 'with collateral' with SMEs with collateral showing a low likelihood of default as indicated a high odd ratio of 1.930. while capital, condition, and character were also significant attributes, the capital, condition and collateral-based information were most effective in predicting the probability of default.

TABLE 4.3.2.1

Prediction of Probability of Default

Variables	**Odd ratio**	**Confidence**	**p-value**

		interval-95%	
Capacity			
High Capacity	0.006	0.28-0.51	0.300
Low capacity	2.080	1.44-2.38	0.047
Collateral			
Without Collateral	1.230	1.41-2.03	0.452
With Collateral	1.930	0.94-1.57	0.020
Capital			
High capital	1.653	1.05-2.08	0.013
Low Capital	0.572	0.28-1.17	0.480
Condition	1.266	1.09-1.73	0.010
Character	0.771	0.48-1.31	0.022
-2log likelihood	10.68		
Chi-square	4.32 ($p<0.07$)		
Degrees of freedom	209		

Source: Author (2016)

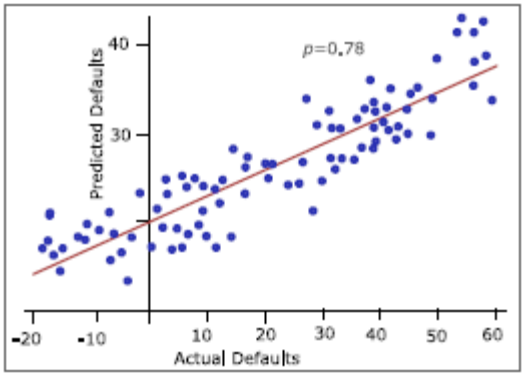
The predictive model of the probability of risk is developed using the logistic regression model provided in chapter 3. As the findings have shown that all the 5 variables have a significant impact on the probability of risk, the predictive model is developed from this study is;

Probability of default = $b_o + b_1$character + $b_2$capacity + $b_3$capital + $b_4$collateral + $b_5$condition (5)

In these findings the predictions obtained from the model were compared to defaults seen and then plotted on a scatter plot with the line of best fit being the regression line. The findings obtained from the observed regression line (the red line in the figure) showed high level of accuracy in predicting the probability of default. Hence, the model is sufficient and valid to provide a prediction of default risk.

FIGURE 4.3.2.2

A plot of predicted defaults agents observed defaults of the Logistic Regression Model



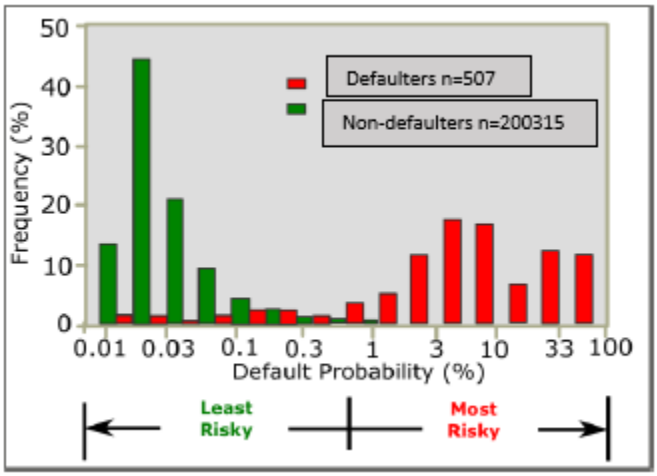
Source: Author (2016)

4.3.3 Validating the Predictive Model For Credit Risk

The validation of the predictive model was conducted by applying the variables from the previous year to the model and assessing its accuracy in predicting the risk of default. Hence, the statistical analysis for validating the model commenced with the backward testing of the model to assess its capacity to predict the defaulting organisation in the year 2017 to 2020. The test was performed by use of the annal walk-forward approach that constituted of 200,315 firms and 507 observations of default. Default prediction for one year was made by utilising information of the year before the targeted year to calibrate coefficients and select the variables for the model. As indicated in the figure below, the model depicted a high accuracy level in predicting the default risk of firms as it was effective in isolating organisations with loan defaults from those who had successfully repaid their loan.

FIGURE 4.3.3.1

Distribution for probability of Defaulting SMEs within a 1-year Period

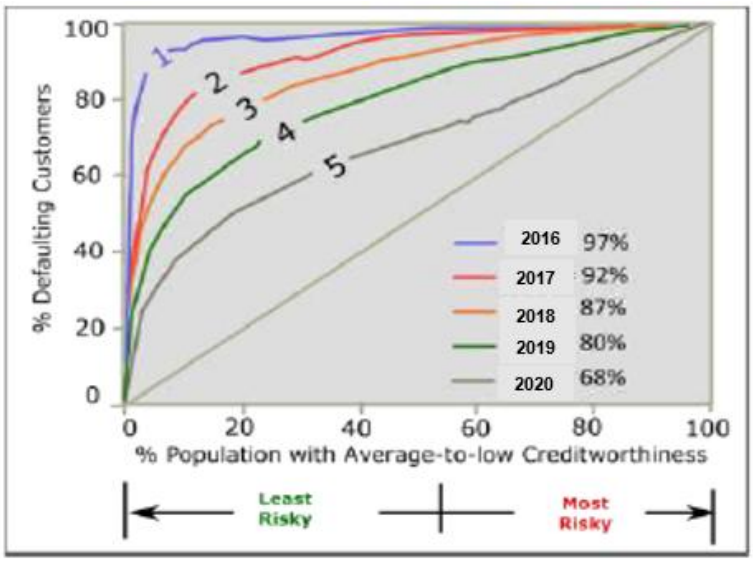


Source: Author (2016)

The accuracy of the model was validated as the organisations in the highest percentile of 97% of defaulted on the loans. However, the prediction power of the regression model decreased significantly with the extension of the prediction period over time. The model demonstrated high level of accuracy in predicting default risk covering a period of 5-year window. The results of the figure above indicates that the defaulting organisations decreased from 95% to 81% in the 105 of the population, followed by a further decrease of 68% to 54% and lastly to 40% in the second, third, fourth and fifth year respectively.

FIGURE 4.3.3.2

Cumulative Accuracy Curves for developing the Default probability from 1-year to 5-year Horizon



Source: Author (2016)

4.4 Discussion of Results

The findings of this study have indicated that an examination of the credit risk, as well as contributing factors enhance the reduction of credit risk for commercial banks in Kenya. An agreement has been established in this study with respect to credit scoring modelling found in the empirical literature. The results indicate that SMEs in industries such as Real Estate have the biggest risk of default in Kenya. These findings are consistent with Adrea (2010) who observed a high default rate in the Real Estate industry.

The use of the 5Cs, which include capital, condition, collateral, capacity and character to assess the creditworthiness of customers has provided a more efficient way of assessing creditworthiness compared to the indicators used within current predictive models. The 5Cs as indicators for creditworthiness of SMEs have highlighted that SMEs in the Transport and Real Estate industries have very low scores. An assessment of Capital indicates that SMEs in heavy investment industries such as Production and Real Estate also have the biggest risk of credit default hence the need for more rigorous risk measurement (Chen et al., 2011).

The application of the logistic regression model indicates a high level of accuracy in terms of default risk prediction. The application of the logistic regression model indicates a high level of

effectiveness in identifying potential credit defaulters, as well as non-defaulters among SMEs (Hilbe, 2009). The level of accuracy was determined to be high over a year period with predictive power decreasing past this period. However, overall, the use of this model for predicting the probability of credit default indicated a good performance level (Deng et al., 2016).

Five main independent variables were determined within the logistic regression model as main contributors in the identification of the probability of SMEs defaulting on loans obtained from commercial banks in Kenya. These variables included the 5Cs, which includes Collateral, Capital, Conditions, Capacity and Character. These variables were selected using the p-value score for measuring significance. The variables proved a high level of accuracy in predicting the risk of credit defaulting by SMEs compared to current predictive models that mainly focus on the gap between the actual and predicted investment by SMEs.

The fitness of the logistic regression model was further tested using two methods. The first method was the change likelihood ratio within the -2 range of likelihood compared to the baseline model. The results of the test indicated strong support for the relationship between the dependent and independent variables. The Akaike Information Criteria (AIC) was used as the second measure for model fitness. The model proved to be a fit for the data due to the lack of significance of the chi-square at the 0.05 level (Hosmer, (2013). As a result, the four independent variables for determining the probability of default of SMEs were determined as significant for use in determining the credit risk of these SMEs by commercial banks in Kenya. Therefore, the model demonstrated a good overall fit for the data as a high sensitivity of the logistic model to the selected variables.

High classification accuracy was also demonstrated by the logistic model for analysing the sample data, The use of measures to evaluate the accuracy of classification indicated a high classification accuracy level. The use of the change likelihood measures, likelihood ratio and AIC to test coefficients indicated that all variables measured were significant. The conclusions drawn from these findings is that the objectives were successfully met and that the developed model was fit for predicting the probability of credit default by SMEs.

4.5 Summary

Considering the findings of this study, the conclusion drawings from assessing the credit risk of SMEs was successful due to the analysis of sufficient variables. The developed model is shown to be effective in the identification of the probability of SMEs defaulting hence guiding commercial banks in the analysis of the creditworthiness of loan applications. In addition, the

developed model is applicable to various other financial institutions in risk assessment and determining the probability of default. Hence financial institutions and commercial banks can apply the model for rating the creditworthiness of customers.

CHAPTER FIVE
CONCLUSIONS AND RECOMMENDATIONS

5.1 Introduction

Three items are addressed in this chapter. The first item is the conclusions drawn from the results of this study. The chapter proceeds to the second item which is to outline and discuss the contribution of this study towards extending previous studies, frameworks and models on the risk of credit default on loans obtained from commercial banks. The final action of the chapter provides recommendations derived from the conclusions for future research. Recommendations for policy action are also provided in the final section of this chapter based on the findings of this current study.

5.2 Conclusions

This study has developed a predictive model for assessing the probability of credit defaulting for loans obtained by SMEs from commercial banks in Kenya. The study, which included a sample of 210 SMEs selected from the 42 commercial banks in Kenya, indicated that credit scoring modelling highlighted a higher risk of defaulting among SMEs in some industries such as Real Estate and Production, which often seek larger loans due to the capital-intensive nature of their ventures. In addition, modelling using the 5Cs to assess the creditworthiness of SMEs indicates that industries such as Transport and Real Estate presented low scores.

The developed logistic regression model proved to be successful in terms of a high level of accuracy in predicting the probability of defaulting on credit by SMEs. Therefore, the model could be effectively used to determine customers who have a high potential of defaulting from those who are non-defaulting.

The use of a logistic regression model enabled the determination of independent variables for assessing the credit risk of SMEs including the Collateral, Capital, Conditions, Capacity and Character. In addition, high accuracy of classification of sample data was possible using the logistic model, which made meeting the objectives of the study possible. This study has filled a missing gap in predictive models for loan default among SMEs in Kenya by providing an adequate and reliable instatement for accurate assessment of the creditworthiness of these SMEs. The identified variables demonstrated high significance for credit scoring in Kenya (Lagat et al., 2013).

Research of this nature presents some limitations. First, the focus on commercial banks and SMEs in Kenya limits the results to SMEs and banks in Kenya since the results may differ if conducted in another country. Secondly, the use of cross-sectional data for credit risk modelling limits the ability of the researcher to address development and changing issues. This limitation could be solved by using time series data collected between 10 to 20 years to capture the impact of time in developing a predictive model (Lemay, 2016).

5.3 Contributions of the study

This project has contributed significantly to credit risk analysis for commercial banks and added more information to the available literature on credit. While most studies focus on either individual asset loans and corporate loans, loans to SMEs are a largely ignored subject considering that these organisations form the backbone of the economy in most developing countries. Due to the significance of SMEs in the economy, the high default rates on loans are damaging to the economy. In addition, this study is significant since it does a proper assessment through the application of both modelling of credit risk assessment and descriptive statistics, which has been missing in extant studies, thereby adding to the current knowledge on the subject. While studies like Lanzarini et al. (2017) have implemented the 5Cs, their application of the concept is considered insufficient for determining the quality of loans. The combination of the 5Cs together with regression modelling in this study, which includes socioeconomic variables for risk assessment is considered more effective in determining the quality of loans. Moreover, Kenya lacks extensive research in the subject of credit risk modelling, such that this study has made a significant attempt into filling this gap. The lack of such studies has made it challenging for commercial banks in Kenya, as well as other financial institutions to determine the credit risk of loan applicants. The findings of this study will help many financial institutions in Kenya to apply the newly proposed credit risk model to define their customers' credit risk. As a result, this study makes a significant contribution to literature pertaining to credit scoring models for Kenya.

The predictive model for credit default developed in this study has resulted in the identification of variables, which will be of significant importance to lending institutions in Kenya. The assumptions made about the model were tested based on data obtained from commercial banks in Kenya. The tests have ensured that the selected variables are fit to support the aim of this study. The compassion of the performance of other credit scoring methods in

extant literature to develop the model used in this study means that the study has contributed new information to credit assessment models.

The continuous growth of credit scoring in investment loans will likely create competitiveness in the banking sector due to an increase in investment loan availability. The use of credit scoring is changing the traditional way of credit risk assessments used by banks, which involved assessing the current information from the bank branch under which the customer holds an account. The centralised and automated systems availed through credit scoring allow banks to avail large investment loans. Therefore, by providing a more accurate credit risk basement, the credit scoring model will support banks and other financial institutions to provide more lending depending on the projected risk.

The project has contributed to the academic knowledge on the analysis of current predictive models for credit risk and added new information more suitable for analysis of credit risk for SMEs. This information is useful considering that most predictive models developed in Kenya are on corporate loans. In addition, the study has contributed significantly to current literature since there is limited research combining credit risk assessment modelling and descriptive analysis for business loans.

5.4 Recommendations for Future Research

One of the main limitations of this research is that by focusing on commercial banks in Kenya, there is a limitation of the focus of the findings to banks in Kenya. As a result, different outcomes may emerge were the study to be conducted in other countries which lower the generalisation of these findings. Therefore, a future area of study would include a comparison of the results of the application of credit scoring models on banks located in different countries to establish better models of scoring for investment loans. In addition, improved models could be realised through a comparison of research conducted on the credit scoring models in Kenya and other developing and developed countries.

The use of cross-sectional data in this research for the credit risk model was advantageous in terms of cost and time-saving. While the cross-sectional design may be attractive, it limits the ability to address the development of changing issues. The selection of loans granted after a specified period of time may limit comparison with the performance of

loans granted after other periods of time especially considering the transitioning economy of Kenya. Future studies should consider using time series data for better results (Lemay, 2016). The use of this type of data may include the impact of time on credit scoring models, as well as the inclusion of other variables connected from literature.

The consideration of select credit assessment techniques such as logistic analysis and the 5Cs in this study reveals omissions of other models for credit assessment due to different requirements by other models for credit assessment. The Kenyan context could benefit from a consideration of other models such as non-parametric models.

REFERENCES

Adem, O., and Waititu, A. (2012). Parametric modelling of the probability of bank loan default in Kenya. *Journal of Applied Statistics, 14*(1), pp.61-74.

Adrea, R. (2010). Measuring the likelihood of small Business default. *Journal of Applied Sciences, 33*(7), pp.1289-1386.

Aladag, C., and Eğrioğlu, E. (2012). *Advances in time series forecasting*. Bentham eBooks.

Balcaen, S., and Ooghe, H. (2004). 35 Years of business failure: An overview of the classic statistical methodologies and their related problems. *British Accounting Review 38*(1), pp.63-93.

Beaver, W. (1966). Financial Ratios as Predictors of Failure. Empirical Research in Accounting: Selected Studied. *Journal of Accounting Research, (4)*.

Berhanu A. & Fufa, B. (2008). Repayment rate of loans from semi-formal financial institutions among small-scale farmers in Ethiopia: Two-limit Tobit analysis. Journal of Socio Economic 37, 2221-2230.

Bernard, R. (2013). *Research methods in anthropology: Qualitative and quantitative approaches.* (4th edn.). Altamira Press, Toronto Canada.

Bofondi, M. and Gobbi, G. (2003). Bad Loans and Entry in Local Credit Markets. Bank of Italy Research Department, Rome.

Business Daily Africa. (2021). SMEs change strategies after pandemic turmoil. Available at: https://www.businessdailyafrica.com/bd/corporate/enterprise/smes-strategies-after-pandemic-turmoil-3300428 (Accessed 7 April 2021)

Calabrese, R. (2012). Modelling SME loan defaults as rare events: The generalized extreme value regression. *Journal Of Applied Statistics, 00*(00), pp.1-17.

Central Bank of Kenya (2014). Bank Supervision Annual Report.

Central Bank of Kenya (CBK), (2016). *The Kenya Financial Sector Stability Report, 2015*. August 2016, Issue No. 7

Central Bank of Kenya (CBK), (2017). *Credit Survey January-March 2017. Annual Report & financial statements 2017.* Available at: https://www.centralbank.go.ke/uploads/banking_sector_reports/623284779_Credit%20S

Chaudhary, M. A. (2003). Credit worthiness of rural borrowers of Pakistan. Journal of

Chen, D., Chou, H., Wang, D., & Zaabar, R. (2011). The predictive performance of a path-dependent exotic-option credit risk model in the emerging market. Physica A, 390(11), 1973–1981.

Cooper, D. R., and Schindler, P. S. (2011). Business Research Methods. (11th ed). New York: McGraw Hill International Edition.

Cooper, D. R., and Schindler, P.S. (2003). *Business Research Methods*. McGraw-Hill.

Creswell, J. W. (2014). *Research Design: Qualitative, Quantitative and Mixed Methods Approaches.* (4th edn.). London: Sage Publications Ltd.

Dastoori, M., and Mansouri, S. (2013). Credit Scoring Model for Iranian Banking Customers and Forecasting Creditworthiness of Borrowers. *International Business Research, 6*(10), pp.25-39.

Deng, X., Liu, Q., and Deng, Y. (2016). An improved method to construct basic probability assignment based on the confusion matrix for classification problem. *Information Sciences, 340*: pp.250-261.

Einav, L., Jenkins, M. and Levin J. (2013). The impact of credit scoring on consumer lending. *RAND J. Econom*, *44*(2), pp.249–274.

Fayyad, U., Piatetsky-Shapiro and Smyth, G. P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 37–54

Goriunov, D. and Venzhyk, K. (2013). Loan default prediction in Ukrainian retail banking. Moscow: Economics Education and Research Consortium

Hastie, T., Tibshirani, R. and Friedman, J. (2009). T*he Elements of Statistical Learning: Data mining, Inference and Prediction*. Berlin: Springer.

Hilbe, J. M. (2009). *Multiple regression models*. CRC Press.

Hosmer J. D. W. (2013). *Lemeshow S, Sturdivant R X. Applied multiple regression*. John Wiley & Sons.

Kamal, S. A., and Hussain, S., and Shafiq, M., and Jahanzaib, M, (2018). Investigating the Adoption of Telemedicine Services An Empirical Study of Factors Influencing Physicians' Perspective in Pakistan The Nucleus,55(3), 153-163.

Kipyego, D. K. & Moses, W. (2013). Effects of credit information sharing on nonperforming loans: The case of Kenya Commercial Bank, Kenya. European Scientific Journal May 2013 edition Vol 19. No 13. ISSN 1857-7881 (PRINT) EISSN 1857-7431.

Lagat, F. K., Mugo, R., and Otuya, R. (2013). *Effect of credit risk management practices on lending portfolio among savings and credit cooperatives in Kenya*.

Lanzarini, L., Villa Monte, A., Bariviera, A., and Jimbo Santana, P. (2017). Simplifying credit scoring rules using LVQ + PSO. Kybernetes, 46(1), 8–16. https://doi.org/10.1108/K-06-2016-0158

Lemay, E. (2016). The Forecast Model of Relationship Commitment. *Journal of Personality and Social Psychology*, *111*(1), pp.34–52.

Li, Y., Thomas, M., and Osei-Bryson, K. (2016). A snail shell process model for knowledge discovery via data analytics. *Decision Support Systems*, *91*, pp.1–12.

Louzada, F., Ferreira-Silva, P., & Diniz, C. (2012). On the impact of disproportional samples in credit scoring models: An application to a Brazilian bank data. Expert Systems with Applications, 39(9), 8071–8078.

Moule, P., and Hek, G. (2015). *Making Sense of Research*. (5th edn.). London: Sage.

Mugenda, A.G. (2008). *Social Science Research: Theory and Principles*. Acts Press, Nairobi.

Mugenda, O. and Mugenda, A. (2003). Research methods: quantitative and qualitative approaches (1st edn.). Nairobi: African Centre for Technology Studies (ACTS)

Munene, H. N. & Guyo, S. H. (March 2013). Factors influencing Loan repayment Default in micro-finance institutions: The experience of Imenti North District, Kenya. International Journal of Applied science & Technology vol 3.

Parylo, O. (2012). Qualitative, quantitative, or mixed methods: An analysis of research design in articles on principal professional development (1998-2008). *International Journal of Multiple Research Approaches*, 6(3), pp.297–313.

Pompe, P.P.M. and Bilderbe, J. (2005). The Prediction of Bankruptcy of Small and Medium-sized Industrial Firms. *Journal of Business Venturing, 20*.

Samreen, A. and Zaidi, F. B. (2012). Design and Development of Credit Scoring Model for the Commercial banks of Pakistan: Forecasting Creditworthiness of Individual Borrowers. *International Journal of Business and Social Science, 3*(17), pp.2219– 1933.

Schreiner, M. (2010). Credit Scoring for Microfinance: Can It Work? J*ournal of Microfinance Risk Management, 2*(2), pp.105-118.

Thun, C. (2011). Credit Risk Models Buy vs. Build. *Moody's Analytics*.

Vitek, F. (2014). *Policy and spillover analysis in the world economy: a panel dynamic stochastic general equilibrium approach*. International Monetary Fund. Socio-Economics, 32, 675-684.

Wanjohi, A.M. and Mugure, A. (2008). Factors affecting the growth of MSEs in rural areas of Kenya: A case of ICT firms in Kiserian Township, Kajiado District of Kenya.

Waweru, N. M. & Kalani V. M. (2009). Commercial Banking crises in Kenya causes and Remedies. African Journal of accounting Economic Finance & Banking Research, 4 (4), 12-33.

Wu, X. (2008). *Credit Scoring Model Validation*. Master's thesis, University of Amsterdam, Korteweg-de Vries Institute for Mathematics.

Yang, P., Zhang, C., and Zhang, X. (2009). Prediction model of credit default probability of listed companies based on Multiple regression analysis. *Economic latitude and longitude, (2),* pp.144-148.

Yap, B. W., Ong, S. H. and Husain, N. H. M. (2011). Using data mining to improve assessment of creditworthiness via credit scoring models. *Expert Systems with Applications, 38*(10), pp.13274-13283.

APPENDIX

Appendix 1: Code for Regression Analysis

```
library(ggplot2)
library(RColorBrewer)
library(gridExtra)
library(png)
library(reshape2)
setwd("C:/Ben")
df <- read.csv("LoansData.csv", header=TRUE)
dir.create(file.path("output"), showWarnings = FALSE)
png(filename="output/LoansData.png", height=750, width=1000,
      bg="white", res=300)
dat <- melt(df, id="DefaultingYear")
ggplot(dat, aes(x=time, y=value, fill=value),stat = "identity", position =
"stack",alpha=.9)+

scale_fill_brewer(palette="Paired",breaks = sort(levels(dat$variable)))
+  geom_density(stat="identity")
dev.off()

dir.create(file.path("output"), showWarnings = FALSE)
png(filename="output/DefaultProbability.png", height=1000, width=1000,
res=300)
dev.off()

ggplot(dat, aes(x=Time, y=value)) +
   geom_area(aes(fill=variable, colour=variable),position='stack')
dir.create(file.path("output"), showWarnings = FALSE)
png(filename="output/DefaultProbability1.png", height=1000, width=1000,
res=300)
dev.off()

dat <- melt(df, id="DefaultingYear")
ggplot(dat, aes(x=Time))+ geom_density(aes(y=value, fill=variable,
colour=variable))
dir.create(file.path("output"), showWarnings = FALSE)
png(filename="output/DefaultProbability2.png", height=1000, width=1000,
res=300)
dev.off()

#  Creating the Cumulative accuracy profiles (CAP) curves
defaultFile <- data.frame("Score"=runif(100,1,100),
                          "hasDefaulted"=round(runif(100,0,1),0))

# Ordering the dataset
defaultFile <- defaultFile[order(defaultFile$Score),]

#  Creating the cumulative density defaultFile$cumden <-
cumsum(defaultFile$hasDefaulted)/sum(defaultFile$hasDefaulted)
```

```
#  Creating the % of defaulting customers
      defaultFile$perpop <- (seq(nrow(defaultFile))/nrow(defaultFile))*100
```

```
# Ploting
plot(defaultFile$perpop,defaultFile$cumden,type="l",xlab="% Population with
Average-to-low Creditworthiness",ylab="% of Defaulting SMEs")

# Creating the regression model

dfdata <- read.csv("Loans.csv", header=TRUE)
summary(dfdata)
par(mfrow=c(1,8))
for(i in 1:8) {
      hist(dfdata[,i], main=names(dfdata)[i])
}
glm.fit <- glm(Default ~ highcapacity + lowcapacity + nocollateral + collateral +
highcapital + lowcapital +
condition+ Character+ RiskMeasures, data = dfdata, family =
binomial)
summary(glm.fit)

xweight <- seq(-20, 60, 10)
yweight <- predict(model_weight, list(wt = xweight),type="response")
plot(dfdata$actual, dfdata$predict, pch = 16, color = "blue", xlab = "Actual Defaults
(g)", ylab = "Predicted Defaults") lines(xweight, yweight)
```

Appendix 2: Project Schedule

Task	Objectives	Start Date	End Date
1. Introduction	Initial discuss and topic approval	10th Jan 2021	12th Jan 2021
	Develop the Introduction chapter	13th Jan 2021	23rd Jan 2021
	Finalize the introduction chapter and submit for review. Revised according to the feedback	24th Jan2021	8th Feb 2021
Literature Review	Conducting extensive research on existing literature regarding the subject matter. And submit for feedback	9th Feb 2021	20th Feb 2021
3. Research Methodology	Define and develop the research methods and submit the supervisor for feedback	22nd March 2021	4th April 2021
	Revise as advised and submit for second opinion	6th April 2021	16th April 2021
	Submit the 3 chapters for feedback	20th April 2021	24th April 2021
	Revise according to the obtained feedback	26th April 2021	29th April 2021
	Defend the proposal to the Panel	3rd May 2021	3rd May 2020
4a. Data collection and preparation	Initiate the data collection process. Submit a consent paper and a data request to the banks	4th May 2021	9th June 2021
4b. Data analysis	Start the data cleaning process and carry out univariate analysis for the identification of variable	7th June 2021	9th July 2021
	Develop the results part of the dissertation by providing the outcomes of the study	20th June 2021	1st July 2021
5. Conclusion	Develop the discussion chapter by discussing the implication of the findings, and analysing its consistency with the findings in the literature review	10th July 2021	20th July 2021
	Develop the collusion and recommendation chapter to highlight the main findings of the research and recommendations for future study	1st Aug 2021	7th Aug 2021
	Work on dissertation draft and supervisor review meeting and Redraft	10th August	14th August 2021
	Defend the dissertation	20th Aug 2021	20th Aug2021

Appendix 3: Resources and Budget

No	Resource	Unit	Price (Ksh)	Total
1	Airtime for communication		4 per minute	12000
2	Internet bundles for online research, Survey administration	4000	2 per MB	8000
3	R software costs	1	00	00
4	Laptop	1	45000	45000
5	Travelling costs both to the University and field work	12	1000 per visit	12000
6	Miscellaneous expense			15000
	Total			**92000**

Appendix 4: Correlation Matrix R- Algorithm

```
mydata <- mtcars[, c(1,3,4,5,6,7)]
head(mydata)

cormat <- round(cor(mydata),2)
head(cormat)

library(reshape2)
melted_cormat <- melt(cormat)
head(melted_cormat)

library(ggplot2)
ggplot(data = melted_cormat, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile()
```