CREDIT RISK MANAGEMENT IN BANKING INDUSTRY: CASE STUDY

ATWIMAN KWANWOMA RURAL BANK

BY

PADDY JONATHAN, BSc. MATHS

A Thesis Submitted to the Department of Mathematics,
Kwame Nkrumah University of Science and Technology
In partial fulfillment of the requirement for the degree of

MASTERS OF SCIENCE
(INDUSTRIAL MATHEMATICS)

Department of Mathematics
Institute of Distance Learning

NOVERMBER, 2012

DECLARATION

I hereby declare that this submission is my own work towards the MSc. and that, to the best of my knowledge, it contains no material previously published by another person nor material which has been accepted for the award of any other degree of the University, except where due acknowledgment has been made in the text.

PADDY JONATHAN
(PG4070210)
……………………… ……………………………… ………………………………

Student Name & ID Signature Date

Certified by:

DR. F.T ODURO
…………………… ……………………………….. ………………………………

Supervisor Name Signature Date

Certified by:

MR. F. ADARKWA
………………………… ……………………………… ……………………………
Head of Dept. Name Signature Date

DEDICATION

This thesis is dedicated to my wife Sarah Paddy. It is also dedicated to the memory of the founders and pioneers of mathematics whose fundamental theorems have made mathematics what it is today and lastly to every individual (both alive and yet to be born) who may have interest in the study of mathematics.

ABSTRACT

The Business of lending is gradually becoming a major target for many banks, as a result there is high competition among the financial institutions in Ghana leading to default of most loans. In order to raise the quality of giving loans and reduce the risk involve in giving loans, credit scoring models have been developed by banks and researchers to improve the process of assessing credit worthiness during the credit evaluation process. This study uses historical data on payments, demographic characteristics and statistical techniques to construct logistic regression model (credit scoring models) and to identify the important demographic characteristics related to credit risk. The logistic regression model was used to design a logistic regression model calculator which was used to calculate the probability of default. Customers' age, sex, occupation, number of dependent, marital Status and amount of loan collected were used. The results showed that default rate is higher in males than in females, 30—39 year olds have the highest rate of default. Married customers defaulted more than the customers who are not married (single) and the higher the number of dependents, the higher the default rate. The self employed clients defaulted more than salary earners. It was found out that the higher the amount of loan collected, the higher the probability of default. The predicting power of the model is 70%.The model has 70.5% accuracy rate of distinqiushing defaulters from non-defaulters. If one was identified as defaulter, he/she had 84% chance of actually defaulting and if a customer was identified as non-defaulter, he/she had 54% chance of actually not defaulting.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACRONYMS	MEANING
ADB	Agricultural Development Bank
CUA	Credit Union Association
FAO	Food and Agricultural Organization
KYC	Knowing Your Customer
MASLOC	Micro Credit and Small Loan Centre
DT	Decision Tree
LDA	Linear Discriminate Analysis
RBF	Radial Basis Function
SVM	Support Vector Machine
MCDM	Multiple criteria Decision Making
MLE	Maximum Likelihood Estimation
BIS	Bank of international Settlement
BCBS	Basel committee on Banking Supervision
SPSS	Statistical package for the Social Sciences
FD	False Default
ND	Non-Default
TP	True Positive
TN	True Negative
FP	False Positive

ACKNOWLEDGMENT

My primary debt of gratitude goes to Almighty God for his grace, mercy and protection all these years of my educational life. Glory be unto his name.

I am also profoundly grateful to my eminent supervisor Dr. F .T Oduro of Department of Mathematics, Kwame Nkrumah University of Science and technology for spending substantial part of his time reading my entire manuscript, carefully pointing out errors and enthusiastically offering many useful suggestions for improvement. I extend my appreciation to the head of Department, Mr. F. Adarkwa and entire lectures of the Department of Mathematics for the knowledge they have impacted to me.

I wish to express my heartfelt thanks to Mr. Sarfo Kantanka (supervisory manager) Atwima kwanwoma Rural Bank, Mr. Oduro Prempeh (Head of loans Atwima kwanwoma Rural Bank) and all the branch Managers for given me the permission to use their Bank as a case study.

I am also thankful to my parents, relatives and friends who in diverse ways contribute to the success of our education. I wish them in returns abundance of love, peace, success and prosperity.

CHAPTER ONE

INTRODUCTION

1.0 BACKGROUND OF THE STUDY

This chapter relating to the background can be classified to three main domains such as: overview of Ghanaian banking system, the need for rural credit and the history relating to Atwima Kwanwoman Rural Bank.

1.1 OVERVIEW OF GHANAIAN BANKING SYSTEM

Even though banking is as primitive as human society, it has gone through changes in many ways world-wide throughout the years. Most banks today offer a wide range of products and services than ever before, and their central functions of putting the community's surplus funds (deposits and investment) to work by lending to people remain unchanged as it has always been. These changes came as a result of government policies, globalization, economic deregulation and information, communication technology (Mohammed and Robert, 2006).

In fact, banks are vital to the health of a nation's economy; monies collected by the bank from the community are given back to the community in the form of loans to buy houses and cars, to start and expand businesses, to pay children's school fees and for other numerous purposes.

In Ghana, the banking industry has not only increased in products and services, but also increased in terms of numbers. The first bank, the Bank of British West Africa (now StanChart Bank) was established in 1896. Vidal et al., (1999) testified that within a short time the bank was able to acquire the business of maintaining the Government accounts and introduced the use of cheques in settlement of Government accounts which helped to educate the public on the usefulness of banking. Vidal et al., (1999) further testified that due to successful operation of the above mentioned bank, another bank, the Colonial Bank now Barclays Bank Ghana was also establish in 1918. The largest indigenous financial institution in the country, Ghana Commercial Bank was established in 1953.

The Bank of Ghana which is the central Bank of the Republic of Ghana was formed in 1957.The Agricultural Development Bank (ADB) was set up by an Act of Parliament (Act 286) in 1965 to promote and modernize the agricultural sector through appropriate but profitable financial intermediation. Its original name then was the Agricultural Credit and Co-operative Bank and the establishing Act gave its main objective as "to provide credit facilities to agriculturists and persons for connected purposes" (www.agricbank.com). Essel and Michael (2011) stated that the ADB which was created to service the rural sector began to concentrate on traditional urban-based banking activities.

The first credit union in Africa was established at Jirapa in the Northern Region (now Upper West) in 1955 by Canadian Catholic missionaries. By 1968, when credit unions were brought under legislation, the Credit Union Association (CUA) was formed as an apex body. There were 254 Credit Unions (64 of them rural) with some 60,000 members (Quainoo, 1997). The number of CUs continued to grow to nearly 500 by the mid-1970s, but their financial performance was not particularly strong. High inflation in the late 1970s and high number of defaulters eroded their capital, and by the early 1990s, the number of CUs had fallen by half (Andoh and Steel, 2003).

1.1.1 THE NEED FOR RURAL CREDIT

The 1992 Constitution of the Republic of Ghana makes a commitment to rural development as part of a national strategy to improve the living conditions in rural areas. Traditionally, rural development credit has been provided by two types of sources: institutional and non-institutional. In rural communities, non-institutional credit is provided by moneylenders, relatives, friends, traders, commission agents, cooperatives, consumers, distributors of farm inputs, and processors of agricultural products (FAO, 1994). This credit market is small; however, the total credit from these non-institutional sources is insufficient to implement rural development programs. For rural development to proceed at a smooth pace, larger institutional sources of credit need to be created. In Ghana, institutional sources of credit are the commercial banks, the Agricultural Development Bank, the National Investment Banks, and the Bank of Ghana Rural Banks (Essel and Michael, 2011).

He testified that institutional sources of credit; commercial banks, the Agricultural Development Bank, the National Investment Banks, and the Bank of Ghana Rural Banks were created to enable development to proceed at a smooth pace. Most Farmers felt neglected and abandoned by the banks that were originally established to take care of their financial needs; since only wealthy people and businessmen, had access to credit from these sources. He point out that due to the high cost involved in given small rural credit, in addition to high level of default that has often accompanied small credit ,the ADB, and the commercial banks lost interest in rural lending and concentrated on lending to people in the urban centers who could afford collateral.

Essel and Michael (2011) said, the Ghana government, through the Bank of Ghana, introduced the idea of rural banking into the country in 1976 to overcome many of these difficulties. According to the association of rural banks (1992), 'the aims of rural banks are:

i. to stimulate banking habits among rural dwellers;
ii. to mobilize resources locked up in the rural areas into the banking systems to facilitate development; and
iii. to identify viable industries in their respective catchment [areas] for investment and development."

As at February, 2011, Ghana can boast of 135 licensed rural banks with Ashanti Region leading with 24 banks.(refer to Appendix 1.0)

1.1.2 HISTORY RELATING TO ATWIMA KWANWOMA RURAL BANK

The Atwima Kwanwoma Rural Bank established on 6th September, 1983 at Pakyi No.2 in the Ashanti Region as the 68th of its kind to be established in Ghana and places 13th on the list in the Region, currently provides the following product and services: Current Accounts, Fixed Deposit Account, Susu Deposit Account, Golden Deposit, Group Savings Accounts, Susu loan, Dwetire Group Loan, Small and Medium Enterprise Loan, Salary Loan, Church Development Loan, Intuitional Loan, Travel Finance Loan and Motor Vehicle loan, has been one of the favorite borrowers last resort. The bank in its policy of bringing banking to the doorsteps of small and medium scale customers, it has opened six (6) branches throughout Ashanti Region at Pakyi (Head Quarters), Santasi, Old-Tafo, New-Tafo, Ayigya and Atonsu-Agogo.

As at the 1990s, the Ghana banking industry specialized in corporate finance, advisory services, and money and capital market activities. The money and capital market were not largely patronized by the banking customers as a result of high demand of collaterals from their clients and hence few defaulters were recorded in the 1990's. The collateral or asserts of the credit subscribers were confiscated bringing excessive fears into those who wished to patronized in the loans.

With the emergence of new banks since the year 2000, Ghana can currently boast of 27commercial banks, numerous Rural Banks, Credit Unions, Micro Finance, and Money and Capital Markets as at 2011, giving the Ghanaian populace easy access to loans and other credit products with little or no collateral. As a result of this, smaller financial institutions especially the emerging credit Unions and Micro Finance Companies and at times even the so-called commercial banks witness high cost of bad debts leading to some of them collapsing. This has made credit risk a critical area in the banking industry and is of concern to a variety of stakeholders: institutions, consumers and regulators, hence the need for it to be studied.

1.2 STATEMENT OF PROBLEM

Credit risk has been the subject of considerable research interest in Banking and Financial communities, and has recently drawn the attention of statistical researchers. The exposure to credit risk continues to be the leading source of problems in the banking industry and as a result needs to be managed. Credit risk is identified as a core pillar for the viability of banks and credit institutions (Michael et al., 2011).

Credit risk is the risk of loss due to a debtor's non-payment of a loan or other line of credit. (Wikipedia.org, as of March 2009). Credit risk is defined as the potential that a bank and borrower or counterparty will fail to meet its obligations in accordance with agreed terms. Default occurs if the debtor is unable to meet its legal obligation according to the debt contract. The examples of default event include the bond default, the corporate bankruptcy, the credit card charge-off, and the mortgage foreclosure. Other forms of credit risk include the repayment delinquency in retail loans, the loss severity upon the default event, as well as the unexpected change of credit rating (Aijun, 2009).

Credit risk does not necessarily occur in isolation, because financial risk is not mutually exclusive. The same source that endangers credit risk for the institution may also expose it to other risk. For instance, a bad portfolio may attract liquidity problem.

Ghana credit market is highly competitive, because of that many financial institution have designed ways to enable them compete with the credit market .Most of them especially the banks who used to lend to enterprises that are able to offer immovable assets as collateral are now lending to individuals, corporate bodies and enterprises without any form of collateral. As a result of this, most banks ended up with high default rate and yet even with collateral; some banks are unwilling to provide lending due to the risk that collateral may have been used in multiple borrowing.

In 2006, the commonest practice of most financial institutions was the door to door banking; people were asked to open accounts and come for loans. Other banks like Stanchart, Barclays Bank and later Fidelity Bank gave loans to government workers who were not their respective customers. It was just by Knowing Your Customer (KYC), collecting their pay- in slips and amortizing the principal and interest to be deducted from their salaries by the

Controller and Accountant General Department. Even though these strategies went a long way to help the banks in generating money, but also contributed to high default rate in Ghana's credit market. According to a survey (www.ghanabusinessnews.com, June 16th, 2009), Barclays Bank Ghana Ltd. last year wrote-off GH¢46.9 million from its books, due to activities of fraudsters and bad loans.

The loss to the bank constitutes 33.3% of the total write-offs made by all the banks in the industry, which was GH¢150.68 million as at 2008.This was the first time in over 100 years history of the banking industry in Ghana. (Suleiman, 2011) highlighted that loan default had reduced the share of the top five banks in the industry from 49.5 per cent in 2009 to 45 per cent in 2010. The paper blamed the government for a fifth of bad loans at the country's banks because of arrears owed to contractors, suppliers, and municipalities. (Suleiman, 2011) wrote that the government's failure to repay its loans has led to road projects being halted and has left schools and clinics struggling to pay for supplies, Government arrears amount to Gh¢2 billion ($1.38 billion), for which it released nearly 600 million Ghana cedis in February,2011 to settle its indebtedness to contractors. United Bank for Africa, (UBA) also wrote off GH¢3 million in the same year (Suleiman, 2011).Ghana Commercial Bank, one of West Africa's biggest banks by number of customers, in February 2010 said it will make a $50 million provision that year to cover bad loans after suffering the fallout of a huge unpaid debt from state owned Tema Oil Refinery (Suleiman, 2011).

Ghana rural banks have also become victims of credit risk problem; as at September 2004 the total loans and advance of the rural banking system was GH ¢55.7 billion. Out of this amount 5.57 billion cedis of the loans were labeled as bad loans, constituting 10% of the loans (Ghana news agency, 2005). This problem has compelled some of the rural banks to issue treat of sending their loan defaulters to court as a result of failing to repay their loan in time. Since many of the borrowers of the rural banks have defaulted in honoring their obligations, the non-performing level of the banks gross loan and advances are on increase. Most rural banks situated in the urban centers that issue Small and Medium Loans lost huge sum of money from defaulters as a result of decongestion exercises; most of the defaulters went back to their respective villages where the bank could not trace them. The problem of credit risk particular in our rural and community banks need to be given greater attention. To allow loans and other source of credit granted by rural banks, whose capital were mobilized largely from the poor and deprived go bad was untenable and defeats the objectives of rural banking system (Ghana News Agency,2005).

The paper argues that, in Ghana 60% population dwell in the rural areas.75% of rural dwellers depend on Agriculture for living. Out of this less than 10% of the rural Banks portfolio goes into Agriculture, because banks do not want to lend to risky sector.

The Micro Credit and Small Loan Centre (MASLOC) was established in 2006 to manage microfinance schemes introduce under the second phase of Ghana's poverty reduction strategy to grow the private sector .Higher default in loan repayment is crippling the scheme and denying others small- scale entrepreneurs access to credit; five million Ghana cedis are yet to be pay by loan defaulters (Domteh, 2010).

Credit risk management is an important aspect of a bank's success and ensures that a lending institution will not take on more risk than it can handle. Credit risk management is the lending institution's primary line of defense to protect itself against customers who fail to meet the terms of the loans or other credit that was extended to them. The objective of credit risk management is to minimize the risk and maximize bank's risk adjusted rate of return by assuming and maintaining credit exposure within the acceptable parameters.

In a financial institution, the risk lost that resulted in the default of payment of the debtors must be expected. Banks and other financial institutions are exposed to many different type of risk, because of these, it is meaningful for a bank to keep substantial amount of capital to protect its solvency and to maintain its economic stability. Basel II is the second of the Basel Accords which was initially published in June 2004. It contains recommendations on banking laws and regulations issued by the Basel Committee on Banking Supervision. The Basel II provides statements of its rules regarding the regulation of the bank's capital allocation in connection with the level of risks the bank is exposed to; it creates an international standard that banking regulators can use when creating regulations about how much capital banks need to put aside to guard against the types of financial and operational risks banks face while maintaining sufficient consistency so that this does not become a source of competitive inequality amongst internationally active banks. According to the Basel II community, the international standard created can help protect the international financial system from problems that might arise should a major bank or a series of banks collapse by setting up risk and capital management requirements well planned to ensure that a bank holds capital reserves appropriate to the risk the bank exposes itself to through its lending and investment practices (www.wikipedia.org, 2011). The more the bank is exposed to risks, the greater the amount of capital must be when it comes to its reserves, so as to maintain its solvency and stability. To determine the risks that come with lending and investment practices, banks must assess the risks. Credit risk management must play its role so as to help banks be in compliance with Basel II Accord and other regulatory bodies.

Atwima Kwanwoma Rural Bank would be used as a case study. Logistic Regression Model would be adopted in this study.

1.3 OBJECTIVES OF THE STUDY

This study is aimed at achieving three main objectives such as:

1. To determine the creditworthiness of a borrower based on a number of quantifiable borrowers characteristics using the logistic regression model.
2. To identify the predicting power of the model and its implications for use in the rural banking sector.
3. To make policy recommendation based on the findings from the thesis.

1.4 METHODOLOGY

The source of data was collected from Atwima Kwawoma Rural Bank at Pakyi Kumasi in Ashanti Region. The data is a secondary data which consist of both demographic and behavioral characteristics of past (5 years ago) and prospective customers of the bank. Computations and Analysis were performed using excel and SPSS. Logistic Regression Model was use to create the credit risk model.

1.5 JUSTIFICATION

Effective credit risk management is essential to the long-term success of any banking organization; since it protect the shareholders money and prevents most financial institutions from going on bankruptcy. If not properly managed, it may lead to the collapsing of the credit market and most companies. It is therefore justifiable because it helps financial institutions to verify whether demographic and behavioral characteristics such as age, sex, occupation, marital status and amount collected among others determines the creditworthiness of a borrower. Hence the logistic regression model can be used by managers of banking industries to plan ahead to prevent people from defaulting.

1.6 LIMITATION

There were several challenges relating to the success of the study. The following are few among others;

1. Unwillingness on the part of agencies concerned to give private information relating to their customers.

2. Limited access to extensive data set of variables.

3. Non- availability of some powerful mathematical and econometric software's for data analysis

1.7 ORGANIZATION OF THE THESIS

Chapter one highlights on the background of study, the objectives, some of the challenges faced during the study and the organization of the study. Chapter two shows the origin of credit risk and also highlights on the works of some of its contributors that lead to the study of this thesis topic. In chapter three, the methodology used to achieve the objectives under study would be clearly stated. Analysis and interpretation of data collected would be done in chapter four. Summary of findings, recommendation and conclusion would also appear in chapter five.

CHAPTER TWO

LITERATURE REVIEW

2.0 INTRODUCTION

This chapter shows origin of credit risk management and also highlights on the works of some of its contributors.

2.1 ORIGIN AND EVOLUTION OF CREDIT RISK MANAGEMENT

The origin and Evolution of Credit risk can be traced back thousands of years ago. Credit is much older than writing. Krimsky and Plough (1988) trace the origin of risk awareness back to Ancient Mesopotamia, based on the work by Covello and Mumpower (1985), as do Golding (1992) and Thompson, et al.(2005). According to Covello and Mumpower (1985), an optimistically early dating of the practice of risk analysis is that of the time of ancient Mesopotamia, as well as ancient Greece and Rome, the former relating to sacerdotal practice and the latter to the history of philosophy (Covello and Mumpower, 1985). The main consideration of Covello and Mumpower's statements about the history of risk are that they are generic.

Althaus (2005) traces a more linear and detailed postulated development through linguistics of the term 'risk'. Even though the code which codified legal thinking from 4,000 years ago in Mesopotamia (Hammurabi's Code) fails to give basic rules of borrowing, it did emphasize that failure to pay a debt is a crime. The code boldly said that one who fails to pay a debt at the right time should be treated identically to theft and fraud. Hammurabi's Code did not address concepts such as interest, collateral and default as use by the present day financial institutions but also set some limits to penalties for a defaulter (Aaron, 2004). Aaron (2004) said a defaulter could face a penalty of been sold by his creditor into slavery, but his wife and children could only be sold for a three-year term.

Tracing the history of most places, credit default was a crime and the defaulters were punished. The punishment varies from place to place. In some areas defaulters of credit risk were punishable by death, mutilation, torture, imprisonment or enslavement. The punishments could be visited upon debtors and their dependents. At times debt could be transfer to relative or political entities. Looking at the penalties involve in credit and pains lenders faced in collecting their moneys one has to be surprise why anyone borrowed or lent money in ancient times.

2.2 THE BANK FOR INTERNATIONAL SETTLEMENT (BIS) AND THE BASEL ACCORDS:

2.2.1 THE BANK FOR INTERNATIONAL SETTLEMENT (BIS):

The Bank for International Settlements (or BIS) is an international organization of central banks which exists to "foster cooperation among central banks and other agencies in pursuit of monetary and financial stability" (Wikipedia online, September 2011). BIS is not accountable to any government. It carries out its work through subcommittees, the secretariats it hosts, and through its annual General Meeting of all members. The BIS also provides banking services, but only to central banks, or to international organizations like itself.

Based in Basel, Switzerland, the BIS was established by the Hague agreements of 1930. As an organization of central banks, the BIS seek to make monetary policy more predictable and transparent among its 58 member central banks. While monetary policy is determined by each sovereign nation, it is subject to central and private banking scrutiny and potentially to speculation that affects foreign exchange rates and especially the fate of export economies. Two aspects of monetary policy have proven to be particularly sensitive, and the BIS therefore has two specific goals: to regulate capital adequacy and make reserve requirements transparent. Capital adequacy policy applies to equity and capital assets. These can be overvalued in many circumstances. Accordingly the BIS requires bank capital/asset ratio to be above a prescribed minimum international standard, for the protection of all central banks involved. The BIS' main role is in setting capital adequacy requirements. From an international point of view, ensuring capital adequacy is the most important problem between central banks, as speculative lending based on inadequate underlying capital and widely varying liability rules causes economic crises as "bad money drives out good"(Gresham's Law).

The BIS sets "requirements on two categories of capital, Tier 1 capital and Total capital. Tier 1 capital is the book value of its stock plus retained earnings. Tier 2 capital is loan loss reserves plus subordinated debt. Total capital is the sum of Tier 1 and Tier 2 capital. Tier 1 capital must be at least 4% of total risk-weighted assets. Total capital must be at least 8% of total risk-weighted assets. When a bank creates a deposit to fund a loan, its assets and liabilities increase equally, with no increase in equity. That causes its capital ratio to drop. Thus the capital requirement limits the total amount of credit that a bank may issue. It is important to note that the capital requirement applies to assets while the bank reserve requirement applies to liabilities.

2.2.2 THE BASEL ACCORDS:

The Basel Accord(s) refers to the banking supervision accords (recommendations on banking laws and regulations), Basel I (first published in 1988 and enforced by law in 1992 by the G-10 countries) and Basel II (published in June 2004) issued by the Basel Committee on Banking Supervision (BCBS). They are called the Basel Accords as the BCBS maintains its secretariat at the Bank of International Settlements in Basel, Switzerland and the committee normally meets there. The Basel Committee consists of representatives from central banks and regulatory authorities of the G10 countries, plus others (specifically Luxembourg and Spain).

The committee does not have the authority to enforce recommendations, although most member countries (and others) tend to implement the Committee's policies. This means that recommendations are enforced through national (or EU-wide) laws and regulations, rather than as a result of the committee's recommendations - thus some time may pass between recommendations and implementation as law at the national level. Tier 1 capital is the core measure of a bank's financial strength from a regulator's point of view. It consists of the types of financial capital considered the most reliable and liquid, primarily Shareholders' equity. Examples of Tier 1 capital are common stock, preferred stock that is irredeemable and non-cumulative, and retained earnings. Capital in this sense is related to, but different from, the accounting concept of shareholder's equity. Both tier 1 and tier 2 capital were first defined in the Basel I capital accord.

The new accord, Basel II, has not changed the definitions in any substantial way. Each country's banking regulator, however, has some discretion over how differing financial instruments may count in a capital calculation. This is appropriate, as the legal framework varies in different legal systems. Tier 2 capital is a measure of a bank's financial strength with regard to the second most reliable form of financial capital, from a regulator's point of view. The forms of banking capital were largely standardized in the Basel I accord, issued by the Basel Committee on Banking Supervision and left untouched by the Basel II accord. Tier 1 capital is considered the core capital and more reliable form of capital.

2.3 CREDIT SCORING MODELS

The decision making process of accepting or rejecting a client’s credit by banks is commonly executed through Judgmental Techniques and/or Credit Scoring models. The Judgmental approach used by most banks and financial institution are based on 3c’s,4c’s or 5C’s which are character (reputation), capital (leverage), collateral, capacity (volatility of earnings) and condition. Credit scoring models are very useful for many practical applications especially in banks and financial institution. Credit scoring model is a system creditors used to assign credit applicant to either a “good credit’’ one that is likely to repay financial obligations or “a bad credit” one who has a high probability of defaulting on financial obligation.

Credit Scoring was used for other purposes such as aiding decision in approving personal applications. Although credit scoring model are widely used for loan applications in financial and banking institutions, it can be used for other type of organizations such as insurance, real estate, telecommunication and recreational clubs for predicting late payments. Credit scoring was first introduced in the 1940s and over the years had evolved and developed significantly. In the 1960s, with the creation of credit cards, banks and other credit card issuers realized the advantages of credit scoring in the credit granting process.

Beaver (1967) and Altman (1968), developed univariate and multivariate models to predict business failures using a set of financial ratios. Beaver(1967) used a dichotomous classification test to determine the error rates a potential creditor would experience if he classified firms on the basis of individual financial ratios as failed or non-failed. He used a matched sample consisting of 158 firms (79 failed and 79 non-failed) and he analyzed 14 financial ratios. Altman (1968) used a multiple discriminant analysis technique (MDA) to solve the inconsistency problem linked to the Beaver's univariate analysis and to assess a more complete financial profile of firms. His analysis drew on a matched sample containing 66 manufacturing firms (33 failed and 33 non-failed) that filed a bankruptcy petition during the period 1946-1965. Altman examined 22 potentially helpful financial ratios and ended up selecting five as providing in combination the best overall prediction of corporate bankruptcy. The variables were classified into five standard ratios categories, including liquidity, profitability, leverage, solvency and activity ratios.

In the 1980s, credit scoring was used for other purposes such as aiding decision in approving personal loan applications. Geske (1977) extends the original single debt maturity assumption to various debt maturities by using compound option modeling. Merton (1974) assumed that the default occurs only at the maturity date, another group of structural models is developed by Black and Cox (1976) and often referred to as "first-passage-time model". In this class of models, default event can happen not only at the debt's maturity, but also can be prior to that date, as long as the firm's asset value falls to the "pre-specified barrier" (that is, default trigger value). Thus, the model does not only allows valuation of debt with an infinite maturity, but, more importantly, allows for the default to arrive during the entire life-time of the reference debt or entity.

In Leland and Toft (1996) work, firms allow to continuously issue debts of a constant but infinite time to maturity. Earlier works in credit risk modeling were characterized by a dominant focus on credit scoring and static assessment of default probabilities,

(Altman and Saunders,1998). However, this situation does not hold in recent articles any longer, since the focus on the studies of credit risk modeling has now changed focus from individual level to the loan-portfolio level and from static model to the dynamic model. Selwyn piramuthu (1997) analyze the beneficial aspects of using both neurofuzzy systems as well as neural networks for credit-risk evaluation decisions. David (2000) investigated the credit scoring accuracy of five neural network models: multilayer perceptron, mixture-of-experts, radial basis function, learning vector quantization, and fuzzy adaptive resonance.

The neural network credit scoring models were tested using 10-fold cross validation with two real world data sets. He results were benchmarked against more traditional methods under consideration for commercial applications including linear discriminant analysis, logistic regression, k nearest neighbor, kernel density estimation, and decision trees. Results demonstrated that the multilayer perceptron may not be the most accurate neural network model, and that both the mixture-of-experts and radial basis function neural network models should be considered for credit scoring applications. Logistic regression was found to be the most accurate of the traditional methods.

William et al., (2000) wrote an article which describes the internal rating systems presently in use at the 50 largest US banking organizations: since internal credit risk rating systems are becoming an increasingly important element of large commercial bank's measurement and management of the credit risk of both individual exposures and portfolios. They used the diversity of current practice to illuminate the relationships between uses of ratings, different options for rating system design, and the effectiveness of internal rating systems. They concluded that Growing stresses on rating systems make an understanding of such relationships important for both banks and regulators

Huang and Huang (2003) model postulates the asset risk premium which was a stochastic process with a negative correlation. They concluded based on the empirical finding that risk premiums of a security tend to move reversely against the returns of stock index in it.

In recent years, credit scoring has been used for home loans, small business loans and insurance applications and renewals (Koh, Tan et al., 2004; Thomas, 2000). Conversely, Liu and Schumann (2005) studied how four feature selection methods - 'ReliefF', 'Correlation-based', 'Consistency-based' and 'Wrapper' algorithms help to improve three aspects of the performance of scoring models: model simplicity, model speed and model accuracy. They realized their results using four classification algorithms - 'model tree (M5)', 'neural network (multi-layer perceptron with back-propagation)', 'logistic regression', and 'k-nearest-neighbours'.

Hsieh (2005), in their submission came with a proposed method or model which demonstrates two real world credit data sets. They also proposed the hybrid mining approach which is used to build effective credit scoring models. A credit scoring model provides an estimate of a borrower's credit risk , i.e. the likelihood that the borrower will repay the loan as promised, based on a number of quantifiable borrower characteristics (Dinh and Kleimeier, 2007). Lean et al., (2007) used a multistage neural network ensemble learning model to evaluate credit risk at the measurement level. They proposed six models consisting of six stages. The first stage, a bagging sampling approach was used to generate different training data subsets especially for data shortage. In the second stage, the different neural network models are created with different training subsets obtained from the previous stage. In the third stage, the generated neural network models are trained with different training datasets and accordingly the classification score and reliability value of neural classifier can be obtained. In the fourth stage, a decorrelation maximization algorithm was used to select the appropriate ensemble members.

In the fifth stage, the reliability values of the selected neural network models (i.e., ensemble members) were scaled into a unit interval by logistic transformation. In the final stage, the selected neural network ensemble members are fused to obtain final classification result by means of reliability measurement. Frydman and Schuermann (2008) discovered two kinds of databases used in building and checking credit risk model. They categoried them as, Standard and Poor (S and P)'s CreditPro database and Moody's KMV Credit Monitor database. These databases have long time series (S and P's built from the year 1981 and Moody's from the early 1990s) and large number of observations. Their conclusion was that the data used was associated with North American banking system and the loan concerned was for large companies.

Historically, discriminant analysis and linear regression have been the most widely used techniques for building scorecards. Other techniques include logistic regression, probit analysis, nonparametric smoothing methods especially k-nearest neighbors, mathematical programming, Markov chain models, recursive partitioning, expert systems, genetic algorithms and neural networks(Hand and Henley, 1997).Artificial neural networks (ANNs) have been critized for its ‘black box’ approach and interpretative difficulties. Multivariate adaptive regression splines (MARS), classification and regression tree (CART), case based reasoning (BR), and support vector machine (SVM) are some recently developed techniques for building credit scoring models. Huang et al., (2004) investigated the performance of the SVM approach in credit rating prediction in comparison with back propagation neural networks (BNN).

However, only slight improvement of SVM over BNN was observed. Lin (2009) use Bank Scope database to target 37 listed banks in Taiwan over the time period of 2002-2004. Huang, Chen, and Wang (2007) reported that compared with neural networks, genetic programming and decision tree classifiers, the SVM classifier achieved identical classification accuracy with relatively few input variables.

Lee et al., 2006 demonstrated the effectiveness of credit scoring using CART and MARS. Their results revealed that, CART and MARS outperform traditional discriminant analysis, logistic regression, neural networks, and support vector machine (SVM) approaches in terms of credit scoring accuracy. Recently, with the development of data mining software the process involved in building credit scoring model is made much easier for credit analysts. Despite the development of new novel techniques, for practical applications the popular techniques for banking and business enterprises are credit scorecards, logistic regression and decision trees as it is relatively easy to identify the important input variable, interpret the results and deploy the model. Evelyn et al., (2009) evaluated the relative performance of logistic credit risk models that were selected by means of standard stepwise model selection methods, and average models obtained by Bayesian model averaging (BMA). Their bootstrap analysis shows that BMA should be considered as an alternative to stepwise model selection procedures.

Xiaolin et al., (2009) used mixed logit model to predict credit risk of listed companies in China. In order to reduce the difficulty in dealing with the facts of correlation and multi dimension of the financial indexes of listed companies and meanwhile to ensure that the data are not lost, they introduced factor analysis to the mixed logit equation and constructed a factor analysis mixed logit model. Fifteen factors were extracted from original financial indexes, and four main factors were selected to substitute the original financial indexes as explanatory variables. Their results show that the new approach is reliable, and general prediction accuracy is higher than 80%.

Amir et al., (2010) applied machine-learning techniques to construct nonlinear nonparametric forecasting models of consumer credit risk. By combining customer transactions and credit bureau data from January 2005 to April 2009 for a sample of a major commercial bank's customers, they were able to constructed out-of-sample forecasts that significantly improve the classification rates of credit-card-holder delinquencies and defaults, with linear regression R2's of forecasted/realized delinquencies of 85%. Using conservative assumptions for the costs and benefits of cutting credit lines based on machine-learning forecasts, they estimated the cost savings to range from 6% to 25% of total losses. Moreover, the time-series patterns of estimated delinquency rates from this model over the course of the recent financial crisis suggests that aggregated consumer-credit risk analytics may have important applications in forecasting systemic risk.

Ming et al., (2010) proposed a model of Discrimination for non-performing loans recovery based on support vector machines (SVM) and wavelet transform integrated according to the complexity of financial system, in order to improve the accuracy and reliability of risk assessment. First they performed model input optimization using the wavelet transform, and then selected the radial basis function (RBF) as the kernel function of wavelet-SVM, At last, they compared with the original SVM, and the experimental results shown that the method is a feasible and effective recognition method with higher generalization performance. Yi et al.,(2010) developed a two-step approach to evaluate classification algorithms for financial risk prediction.They constructs a performance score to measure the performance of classification algorithms and introduces three multiple criteria decision making (MCDM) methods (i.e., TOPSIS, PROMETHEE, and VIKOR) to provide a final ranking of classifiers. An empirical study was designed to assess various classification algorithms over seven real-life credit risk and fraud risk datasets from six countries. The results show that linear logistic, Bayesian Network, and ensemble methods are ranked as the top-three classifiers by TOPSIS, PROMETHEE, and VIKOR. Their work also discusses the construction of a knowledge-rich financial risk management process to increase the usefulness of classification results in financial risk detection. Bee et al.,(2011) used data mining to improve assessment of credit worthiness using credit scoring models. Due to privacy concerns and unavailability of real financial data from banks they used data of payment history of members from a recreational club for the credit scoring model. The club has been facing a problem of rising number in defaulters in their monthly club subscription payments. The management would like to have a model which they can deploy to identify potential defaulters.

The classification performance of credit scorecard model, logistic regression model and decision tree model were compared. The classification error rates for credit scorecard model, logistic regression and decision tree were 27.9%, 28.8% and 28.1%, respectively. They came to a conclusion that although no model outperforms the other, scorecards are relatively much easier to deploy in practical applications. Gang Wang and Jian Macwe (2011) also proposed an integrated ensembled approach, called RS-Boosting, which is based on two popular ensembled strategies, i.e., boosting and random subspace, for corporate credit risk prediction. They realized that there were two different factors encouraging diversity in RS-Boosting, which will be advantageous to get better performance. Therefore two corporate credit datasets were selected to demonstrate the effectiveness and feasibility of the proposed method.

Experimental results reveal that RS-Boosting gets the best performance among seven methods, i.e., logistic regression analysis (LRA), decision tree (DT), artificial neural network (ANN), bagging, boosting and random subspace. They concluded that RS-Boosting can be used as an alternative method for corporate credit risk prediction. Yao Ping and Lu Yongheng (2001) developed a credit scoring model by constructing a hybrid SVM-based credit scoring models to evaluate the applicant's credit score according to the applicant's input features: (1) using neighborhood rough set to select input features; (2) using grid search to optimize RBF kernel parameters; (3) using the hybrid optimal input features and model parameters to solve the credit scoring problem with 10-fold cross validation; (4) comparing the accuracy of the proposed method with other methods.

Their Experimental results demonstrate that the neighborhood rough set and SVM based hybrid classifier has the best credit scoring capability compared with other hybrid classifiers.

It also outperforms linear discriminant analysis, logistic regression and neural networks. Mileris (2011), focused on a credit rating model development which is related to credit ratings for Lithuanian companies. He went further to describe the steps of model's development and improvement processes. He explained that model's development begins with the selection of initial variables (financial ratios), characterizing default and non-default companies. 20 financial ratios of 5 years were calculated according to annual financial reports.

A discriminant analysis, logistic regression and artificial neural networks (multilayer perceptron) were applied and the conclusion was that developed model is a valid tool for the estimation of credit risk. Jian et al,.(2011) used the Decision tree (DT) which happens to be one of the most popular classification algorithms in data mining and machine learning. They proposed of two dual strategy ensemble trees: RS-Bagging DT and Bagging-RS DT, which are based on two ensemble strategies: bagging and random subspace, to reduce the influences of the noise data and the redundant attributes of data and to get the relatively higher classification accuracy. Two real world credit datasets are selected to demonstrate the effectiveness and feasibility of proposed methods.

Experimental results reveal that single DT gets the lowest average accuracy among five single classifiers, i.e., Logistic Regression Analysis (LRA), Linear Discriminant Analysis (LDA), Multi-layer Perceptron (MLP) and Radial Basis Function Network (RBFN). Moreover, RS-Bagging DT and Bagging-RS DT get the better results than five single classifiers and four popular ensemble classifiers, i.e., Bagging DT, Random Subspace DT, Random Forest and Rotation Forest.

They concluded that RS-Bagging DT and Bagging-RS DT can be used as alternative techniques for credit scoring. Chun-Ling Chuang and Szu-Teng Huang (2011) presented a research proposal on a hybrid system which combines fuzzy clustering and MARS. They said both models are suitable for the bankruptcy prediction problem, due to their theoretical advantages when the information used for the forecasting is drawn from company financial statements. They tested the accuracy of their approach in a real setting consisting of a database made up of 59,336 non-bankrupt Spanish companies and 138 distressed firms which went bankrupt in 2007.

As benchmarking techniques, they used discriminant analysis, MARS and a feed-forward neural network. They concluded that the hybrid model outperforms the other systems, both in terms of the percentage of correct classifications and in terms of the profit generated by the lending decisions. By the approach of Logistic Regression, Meng- lung et al,. (1999) took samples from one domestic credit card center. About 300 normal card users and 300 abnormal card users were chosen respectively. Key considerations are: gender, education degree, the type of his job, how many years in the company, how many cards he owns when he applies a new card, and whether he is one of the clients asking for house loans in this bank.

Mei-hui Lu (2000) with the use of Logistic Regression found that some vital elements stood out if the factor of debtors' corresponding areas is not taken into account. These are: marital status, academic degree, financial history, the length of the loan, the relationship between a debtor and a guarantor, and the comparative relationship between mail address of the debtor and the location of his collateral. Besides, one can use the final mode to get a probability of turning into a bad debt. This last measure can be used as another criterion of examination.

Rock (1984) thought the organizations issuing credit cards put more focus on the relationship between other creditors, the ratio of debt to income, income, vocation, ownership of the house, the length of working in the same company, and whether the debtor has a check account or deposit account.

Updegrave (1987) held the view that the organizations issuing credit cards valued eight variables: income, employment, age, the number of creditors, the length of working in the same company, whether he has a check account or deposit account, having been declared bankruptcy before, and the record of past payment by credit cards.

Steenacker and Goovaerts (1989) employed the mode of Logistic-Regression and got the conclusion that the variables influencing credit loans are: age, having a telephone or not, how long the debtor has lived in the current residence, how long he has worked in the company, the standard of his living area, job, if he is a civil servant, monthly income, ownership of the house, the number of past loans, and the length of loans.

Ang et al., (1979) investigated the profiles of late-paying consumer loan borrowers using variables such as gross amount of loan, age, sex, marital status, number of dependents, years lived at residence, monthly take home pay, monthly take home pay of spouse, own or rent residence, other monthly income, total monthly payments on all debts, type of bank accounts, number of credit references listed, years on job, total family monthly income per month, debt to income ratio, total number of payments on the loan, and annual percentage interest on the loan.

Koh et al., (2004) used age, annual income, gender, marital status, number of children, number of other credit cards held and whether the applicant has an outstanding mortgage loan to construct a credit scoring model to predict credit risk of credit card applicants as bad loss, bad profit and good risk.

Abdou et al., (2008) used twenty variables some of which were loan amount, loan duration, sex, marital status, age, monthly salary, additional income, house owned or rent, and education level for building credit scoring models to evaluate credit risk (paid or default) for personal loan. Gschwind (2007) concluded that mining basic tenant data, accounts receivable data, and government-published data can generate predictions of late payments of rental. Mavri et al,. (2008) used variables such as gender, age, education, marital status and monthly income to estimate the risk level of credit card applicants. Vojtek and Kocenda (2006) provided a table of indicators that are typically important in retail credit scoring models. They classify the indicators as demographic, financial, employment and behavioural indicators.

CHAHPTER THREE

METHODOLOGY OF THE STUDY

3.0 INTRODUCTION

In this chapter, we would discuss basic concepts on regression, linear regression and logistic regression.

3.1 REGRESSION

Regression methods have become an integral component of any data analysis concerned with describing the relationship between a response variable and one or more explanatory variables. It is often the case that the outcome variable is discrete taking on two or more possible values. In simplest terms, regression is a statistical procedure which attempts to predict the values of a given variable, (termed the dependent, outcome, or response variable) based on the values of one or more other variables (called independent variables, predictors, or covariates). The result of a regression is usually an equation (or model) which summarizes the relationship between the dependent and independent variable(s).

Typically, the model is accompanied by summary statistics describing how well the model fits the data, the amount of variation in the outcome accounted for by the model, and a basis for comparing the existing model to other similar models. By comparing these statistics across multiple models the user is able to determine a combination and order of independent variables that most satisfactorily predict the values of the outcome. Numerous forms of regression have been developed to predict the values of a wide variety of outcome measures. Before beginning a study of logistic regression it is important to understand that the goal of any analysis using this method is the same as that of any model-building technique used in statistics: to find the best credit risk model to describe the relationship between an outcome (dependent or respond) variable and a set of independent (predictor or explanatory) variables. These independent variables are often called covariates. The most common example of modeling usually used is the linear model, where the outcome variable is assumed to be continuous.

Over the last decade the logistic regression model has become, in many fields, the standard method of analysis in this situation. What distinguishes a logistic regression model from the linear regression model is that the outcome variable in logistic regression is binary or dichotomous. This difference between logistic and linear regression is reflected both in the choice of a parametric model and in the assumptions. Once this difference is accounted for, the method employed in an analysis using logistic regression followed the same general principles used in linear regression. Thus a techniques used in linear regression analysis will motivate our approach to logistic regression.

The first difference concerns the nature of the relationship between the outcome and independent variables. In any regression problem the key quantity is the mean value of the outcome variable, given the value of the independent variable. This quantity is called conditional mean and will be expressed as "$E(y/x)$" where y denotes the outcome variable and x denotes a value of the independent variable. The quantity $E(y/x)$ is read "the expected value of y, given the value x."In linear regression we assume that this mean may be expressed as equation linear in x (or some Transformation of x or y) such as $E(y/x)=\beta_0+\beta_1 x$.

Thus expression implies that it is possible for $E(y/x)$to take on any value as x ranges between -$\infty$and ∞.With dichotomous data the conditional mean must be greater than or equal to zero and less or equal to 1, $0 \leq E(y/x) \leq 1$This mean approaches Zero and 1 gradually. The change in the $E(y/x)$per unit change in x becomes progressively smaller as the conditional mean gets to zero or 1. In order to simplify notation, we use the quantity $\pi(x) = E(y/x)$ to represent the conditional mean of y given x when the logistic regression model is $\pi(x) = \frac{e^{B_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$

A transformation of $\pi(x)$ that is central to our study of logistic regression is the logit transformation. The logit transformation is defined, in terms of $\pi(x)$, as:

$$g(x) = In\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 x.$$

The importance of this transformation is that $g(x)$ has many of the desirable properties of a linear regression model. The logit, $g(x)$, is linear in its parameters, may be continuous, and may range from - $\infty$and ∞ depending on the range of x.

The second important difference between the linear and logistic regression models concerns the conditional distribution of the outcome variable. In linear regression model we assume that an observation of the outcome variable may be expressed as $y = E(y/x) + \varepsilon$. The quantity ε is called the error and expresses an observation's deviation from the conditional mean. The most common assumption is that ε follows a normal distribution with mean zero and some variance that is constant across levels of the independent variable. It follows that the conditional distribution of the outcome variable given x will be normal with mean $E(y/x)$, and a variance that is constant. This is not the case with a dichotomous outcome variable. In this situation we may express the value of the outcome variable given x as $y = \pi(x) + \varepsilon$. Here the quantity ε may assume one of two possible values. If $y = 1$ then $\varepsilon = 1 - \pi(x)$ with probability $\pi(x)$, and if $y = 0$ then $\varepsilon = -\pi(x)$ with $1 - \pi(x)$.Thus, ε has a distribution with mean zero and variance equal to $\pi(x)[1 - \pi(x)]$.That is, the conditional distribution of the outcome variable follows a binomial distribution with probability given by the conditional mean, $\pi(x)$.

3.2 LOGISTIC REGRESSION

The most commonly used statistical methodology is the logistic regression method. Logistic regression is a mathematical modeling approach that can be used to describe the relationship of several *X'*s (independent variables) to a dichotomous dependent variable. Even though other modeling approaches such as discriminant analysis are possible also, but logistic regression is by far the most popular modeling procedure used to analyze credit risk management where the payment status is dichotomous. By definition Logistic Regression is defined as a technique for analyzing problems in which there are one or more independent variables, which determine an outcome that is measured with a dichotomous variable in which there are only two possible outcomes. In logistic regression, the dependent variable is binary or dichotomous. The goal of logistic regression is to find the best fitting model to describe the relationship between the dependent variable – probability to default and a set of independent (predictive) variables. This section involves the construction of the logistic regression model. The data consist of defaulters and non-defaulters. The dependent variable which is payment status is code as defaulter =1 and Non-defaulter =0. The objective of Logistic regression model in credit scoring is to determine the conditional probability of a specific applicant belonging to a class (defaulter or non-defaulter).For this study logistic regression would be used to model the event Y=1 (defaulter) and the predicted values are always between 0 and 1, and correspond to the probability of Y being one (1).

Unlike ordinary linear regression, logistic regression does not assume that the relationship between the independent variables and the dependent variable is a linear one, Nor does it assume that the dependent variable or the error terms are distributed normally .Logistic regression is commonly used to obtain predicted probabilities that a unit of the population under analysis will acquire the event of interest as a linear function of one or more: Continuous-level variables, Dichotomous (binary) variables or, a combination of both continuous and binary independent variables.

3.2.1 LOGISTIC FUNCTION

Logistic function describes the mathematical form on which the logistic model is based. This function, called ***f***(***z***), is given by 1 over 1 plus *e* to the minus *z*. When plotted the values of this function as *z* varies from $-\infty$ to $+\infty$, as shown in fig1, the range of *f*(*z*) is between 0 and 1, regardless of the value of *z***.**

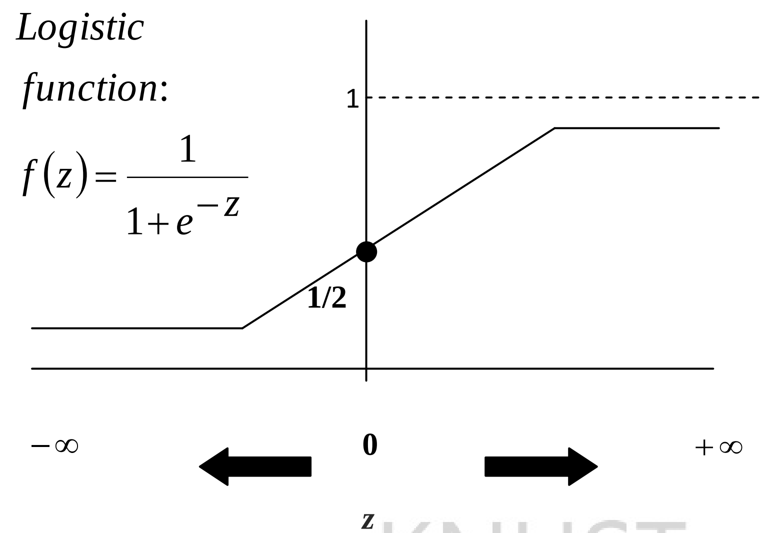
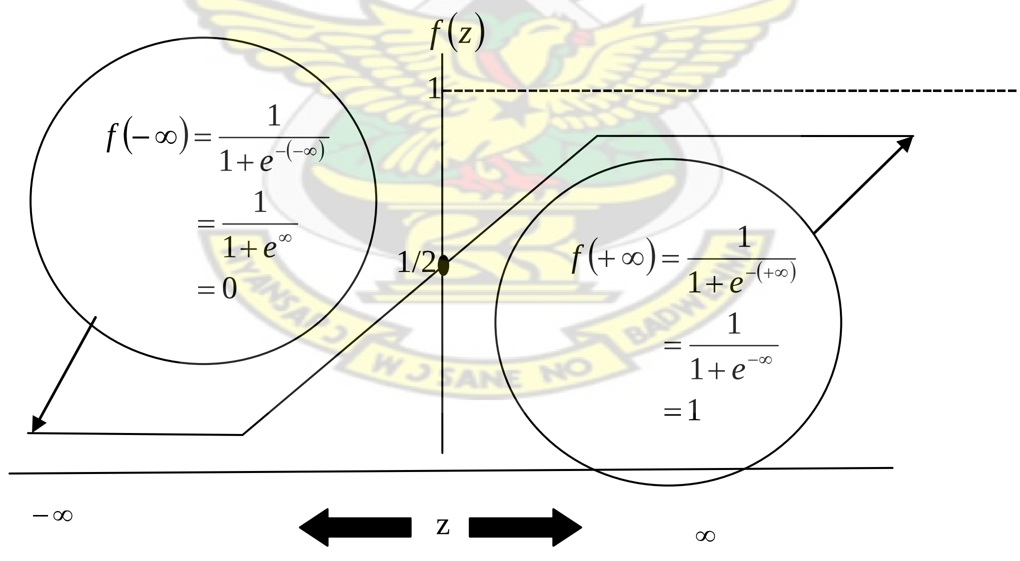


Fig 3.1 plot of logistic function 1

Considering the circle on the left side of the graph of figure 2, when z is $-\infty$, the logistic function ***f*(*z*)** equals 0. On the right side, when z is ∞, then ***f*(*z*)** equals 1.



$Range: 0 \leq f(z) \leq 1$

Fig 3.2 plot of logistic function 2

The fact that the logistic function $f(z)$ ranges between 0 and 1 is the primary reason the logistic model is so popular. The model is designed to describe a probability, which is always some number between 0 and 1. The logistic model, therefore, is set up to ensure that whatever estimate of risk we get, it will always be some number between 0 and 1. Thus, for the logistic model, we can never get a risk estimate either above 1 or below 0. This is not always true for other possible models, which is why the logistic model is often the first choice when a probability is to be estimated.

The shape of the logistic function is S shaped. As shown in the graph below, if we start at $z = -\infty$ and move to the right, then as z increases, the value of $f(z)$ hovers close to zero for a while, then starts to increase dramatically toward 1, and finally levels off around 1 as z increases toward $+\infty$. The result is an elongated, S shaped picture.

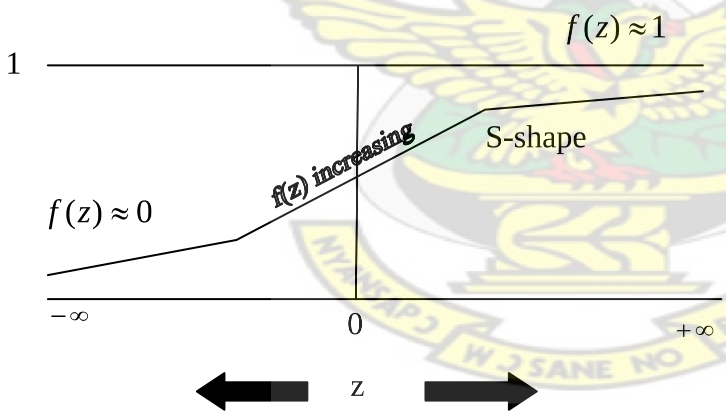


Fig 3.3 plot of logistic function 3

3.2.2 ASSUMPTIONS ABOUT LOGISTIC REGRESSION

Logistic regression does not make any assumptions of normality, linearity, and homogeneity of variance for independent variables. It therefore prefer to other types of analysis such as discriminant analysis when the data does not satisfy these assumptions.

3.2.3 RELATIONSHIP BETWEEN THE ODDS RATIO AND REGRESSION COEFFICIENT OF DICHOTOMOUS INDEPENDENT VARIABLE

In the logistic regression model, the slope coefficient represents the change in the logit corresponding to a change of one unit in the independent variable, that is, $\beta_1 = g(x+1) - g(x)$ The independent variable, x, is coded as either zero or one.

The difference in the logit for subject x=1 and x=0 is

$g(1) - g(0) = [\beta_0 + \beta_1] - [\beta_0] = \beta_1$.The first step in the interpreting the effect of a covariate in the model is to express the desired logit difference in terms of the model. In this case the logit difference is equal to β_1.The odds of an outcome being present with x=1 is defined as $\frac{\pi(1)}{1-\pi(1)}$ and odds of the outcome being present among x=0 is define as $\frac{\pi(0)}{1-\pi(0)}$.The odd ratio denoted by OR is defined as ratio of the odds for x=1 to x=0 and is given $OR = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]}$.........3.1

Table 3.1 Values of the Logistic Regression Model

When the Independent Variable Is Dichotomous

Outcome Variable(Y)	Independent Variable(X) $x = 1$	$x = 0$
$y = 1$	$\pi(1) = \frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}}$	$\pi(0) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$
$y = 0$	$1 - \pi(1) = \frac{1}{1 + e^{\beta_0 + \beta_1}}$	$1 - \pi(0) = \frac{1}{1 + e^{\beta_0}}$
Total	1.0	1.0

Substitute the expression for the logistic regression model in table 3.1 into equation 3.1

$$OR = \frac{\left(\frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}}\right) \Big/ \left(\frac{1}{1 + e^{\beta_0 + \beta_1}}\right)}{\left(\frac{e^{\beta_0}}{1 + e^{\beta_0}}\right) \Big/ \left(\frac{1}{1 + e^{\beta_0}}\right)} \text{..........3.2}$$

Simplify Equation 3.2

$$= \frac{e^{\beta_0 + \beta_1}}{e^{\beta_0}}$$

$$OR = \frac{e^{\beta_0 + \beta_1}}{e^{\beta_0}}$$

$$\Rightarrow OR = e^{\beta_1}$$

Hence relationship between odds Ratio and the regression coefficient of a dichotomous independent variable is $OR = e^{\beta_1}$

3.2.4 MATHEMATICAL APPLICATION OF LOGISTIC REGRESSION

$$Logit(Y) = \ln(odds) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \ldots\ldots\ldots\ldots\ldots\ldots 3.3$$

Odds are defined as the probability of belonging to one group divided by the probability of belonging to the other.

$odd = \dfrac{P(Y=1)}{1-P(Y=1)} \ldots\ldots\ldots 3.4$, Where $P(Y=1)$ is the probability of outcome of interest.

Equation equations 3.3 and equation 3.4

$$\log\left[\frac{P(Y=1)}{1-P(Y=1)}\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k$$

Let $Z = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k \ldots\ldots\ldots\ldots 3.5$

$$In\left[\frac{P(Y=1)}{1-P(Y=1)}\right] = Z$$

taking the exponent of each side of the equation, we get

$$e^{In\left[\frac{p(Y=1)}{1-(Y=1)}\right]} = e^{Z}$$

$$\frac{P}{1-P} = e^{Z} \qquad\qquad P = e^{Z}[1-P]$$

$$P = e^{Z} - e^{Z} p$$

$$p + e^{Z} p = e^{Z}$$

$$P\left[1 + e^{Z}\right] = e^{Z}$$

$$P = \frac{e^{Z}}{1 + e^{Z}} \quad \ldots\ldots\ldots\ldots\ldots 3.6$$

$$P = \frac{1}{1 + e^{-Z}}, \text{ Where } z = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_k x_k$$

Equation for P(y = 1) – the proportion of subjects with y = 1 and substituting the value z back into equation 3,

$$P = \frac{e^{\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_k x_k}}{1 + e^{\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots \beta_k x_k}} \quad \ldots\ldots\ldots 3.7$$

e is the “natural constant” ≈ 2.718 , p = probability (proportion) of $y = 1$

Where α is the Y intercept, βs are regression coefficients, xs are set of predictors.

α and β are typically estimated by the maximum likelihood (ML) Method

It can be generalized from the above deductions that

$$0 < e^{\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_k x_k} \leq \infty$$

And so it follows that

$$0 < \frac{e^{\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_k x_k}}{1 + e^{\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots \beta_k x_k}} \leq 1$$

$0 < p \leq 1$. This formulation for p ensures that our estimates of the probability of having the condition "y" is between 0 and 1

By the Bayes rule:

$$\overset{\Lambda}{G}(X) = \arg\max \Pr(G = k/X = x).$$

Decision boundary class k and l is determined by the equation:

$$\Pr(G = k/X = x) = \Pr(G = l/X = x).$$

Divide both sides by $\Pr(G = l/X = x)$ and take the log. The above equation is equivalent to

$$\log \frac{\Pr(G = k/X = x)}{\Pr(G = l/X = x)} = 0$$

Since we enforce linear boundary, we can assume

$$\log \frac{\Pr(G = k/X = x)}{\Pr(G = l/X = x)} = a_0^{(k,l)} + \sum_{j=1}^{p} a_j^{(k,l)} x_j.$$

For logistic regression, there are restrictive relations between $a^{(k,l)}$ for different pair of (k,l)

$$\log \frac{\Pr(G = 1/X = x)}{\Pr(G = k/X = x)} = \beta_{10} + \beta_1^T x$$

$$\log \frac{\Pr(G = 2/X = x)}{\Pr(G = k/X = x)} = \beta_{20} + \beta_2^T x$$

.

$$\log\frac{\Pr(G=k-1/X=x)}{\Pr(G=k/X=x)}=\beta_{(k-1)0}+\beta_{k-1}^{T}x$$

For any pair (k,l):

$$\log\frac{\Pr(G=k/X=x)}{\Pr(G=l/X=x)}=\beta_{ko}-\beta_{10}+(\beta_k-\beta_l)^{T}x$$

Number of Parameters: $(k-1)(p+1)$

Denote the entire parameter by

$$\theta=\{\beta_{10},\beta_1,\beta_{20},\beta_2,\ldots\beta_{(K-1)0,}\beta_{(K-1)}\}$$

The log ratio of posterior probabilities are called log-odds or logit transformations

Under the assumptions, the posterior probabilities are given

by:

$$\Pr(G=k/X=x)=\frac{\exp(\beta_{k0}+\beta_k^{T})}{1+\Sigma_{l=1}^{K-1}\exp(\beta_{10}+\beta_l^{T}x)}$$

For $k=1,\ldots\ldots.,K-1$

$$\Pr(G=K/X=x)=\frac{1}{1+\Sigma_{l=1}^{K-1}\exp(\beta_{10}+\beta_l^{T}x)}$$

For $\Pr(G=k/X=x)$ give above, obviously

Sum up to 1: $\sum_{k-1}^{K}\Pr(G=k/X=x)=1$

A simple calculation shows that the assumptions are satisfied.

Similarities:

Both attempt to estimate $\Pr(G=k/X=x)$

Both have linear classification boundaries

Difference:

Linear regression on indicator matrix: approximate

$\Pr(G=k/X=x)$ by a linear function of x.

$\Pr(G=k/X=x)$ is not guaranteed to fall between 0 and 1

and to sum up to 1.

Logistic regression: $\Pr(G=k/X=x)$ is a nonlinear function

of x. It is guaranteed to range from 0 to 1 and to sum up to 1.

Criteria: find parameters that maximize the conditional likelihood of G given X using the training data.

Denote $P_k(x_i;\theta)=\Pr(G=k/X=x_i;\theta)$.

Given the first input x_1, the posterior probability of its class

Being g1 is $\Pr(G=g_1/X=x_1)$

Since samples in the training data set are independent, the

Posterior probability for the N samples each having class g_i,

$i = 1, 2, ..., N$, given their inputs $x_{1,}x_2,\ldots\ldots x_N$ is:

$$\prod_{i=1}^{N} \Pr(G = g_i / X = x_i)$$

The conditional log-likelihood of the class labels in the training data set is

$$L(\theta) = \sum_{i=1}^{N} \log \Pr(G = g_i / X = x_i)$$

$$= \sum_{i=1}^{N} \log P g_i (x_i ; \theta)$$

For binary classification, if $g_i = 1$, denote $y_i = 1$; if $g_i = 2$,

Denote $y_i = 0$.

Let $P_1(x;\theta) = P(x;\theta)$, then

$P_2(x;\theta) = 1 - P_1(x;\theta) = 1 - P(x;\theta)$.

Since $K = 2$,the parameter $\theta = \{\beta_{10}, \beta_1\}$

We donates

$$\beta = (\beta_{10}, \beta_1)^T$$

If $y_i = 1$ i.e., $g_i = 1$,

$$\log P_{gi}(x;\beta) = \log P_1(x;\beta)$$

$$= 1.\log P(x;\beta)$$

$$= y_i.\log P(x;\beta)$$

If $y_i = 0$, i.e., $g_i = 2$,

$$\log P_{gi}(x;\beta) = \log P_2(x;\beta)$$

$$=1.\log(1 - P(x;\beta))$$

$$=(1 - y_i)\log(1 - p(x;\beta)).$$

Since either $y_i = 0$ or $1 - y_i = 0$, we have

$$\log P_{gi}(x;\beta) = y_i \log P(x;\beta) + (1 - y_1)\log(1 - p(x;\beta))$$

The conditional likelihood

$$L(\beta) = \sum_{i=1}^{N} \log P_{gi}(x_i;\beta)$$

$$= \sum_{i=1}^{N}[y_i \log P(x_i;\beta) + (1 - y_i)\log(1 - p(x_i;\beta))]$$

There are $p+1$ parameters in $\beta = (\beta_{10}, \beta_1)^T$

Assume a column vector form for β:

$$\beta = \begin{pmatrix} \beta_{10} \\ \beta_{11} \\ \beta_{12} \\ . \\ . \\ . \\ \beta_{1,p} \end{pmatrix}$$

Here we add the constant terms 1 to x accommodate the intercept

$$\beta = \begin{pmatrix} x,1 \\ x,2 \\ . \\ . \\ . \\ . \\ x,p \end{pmatrix}$$

By assumption of logistic regression model :

$$P(x;\beta) = \Pr(G = 1/X = x) = \frac{\exp\left(\beta^T x\right)}{1 + \exp\left(\beta^T x\right)}$$

$$1 - P(x;\beta) = \Pr(G = 2/X = x)\frac{1}{1 + \exp\left(\beta^T x\right)}$$

Substitute the above in $L(\beta)$:

$$L(\beta)=\left[\,y_i\beta^T x_i-\log\left(1+e^{\beta^T x_i}\right)\right]$$

To maximize $L(\beta)$,we set the first order partial derivatives of $L(\beta)$ to zero.

$$\frac{\partial L(\beta)}{\beta_{1j}}=\sum_{i=1}^{N} y_i x_{ij}-\sum_{i=1}^{N}\frac{x_{ij}e^{\beta^T x_i}}{1+e^{\beta^T x_i}}$$

$$=\sum_{i=1}^{N} y_i x_{ij}-\sum_{i=1}^{N} P(x;\beta)x_{ij}$$

$$=\sum_{i=1}^{N} x_{ij}\left(y_i-P(x_i;\beta)\right)$$

For all $j=0,1......,p$

In matrix form, we write

$$\frac{\partial L(\beta)}{\partial\beta}=\sum x_i\left(y_i-p(x;\beta)\right)$$

To solve the set of $p+1$nonlinear equations $\frac{\partial L(\beta)}{\partial\beta_{ij}}=0$,

$j=0,1......,p$, use the Newton-Raphson algorithm.

The Newton-Raphson algorithm requires the second-derivatives or Hessian matrix:

$$\frac{\partial^2 L(\beta)}{\partial\beta\partial\beta^T}=-\sum_{i=1}^{N} x_i x_i^T P(x_i;\beta)(1-p(x_i;\beta))$$

The element on the jth row and nth column is (counting from 0):

$$\frac{\partial L(\beta)}{\partial\beta_{1j}\partial\beta_{1n}} = -\sum_{i=1}^{N} \frac{\left(1+e^{\beta^T x_i}\right)e^{\beta^T x_i} x_{ij} x_{in} - \left(e^{\beta^T x_i}\right)^2 x_{ij} x_{in}}{(1+e^{\beta^T x_i})^2}$$

$$= -\sum_{i=1}^{N} x_{ij} x_{in} p(x_i;\beta) - x_{ij} x_{in} p(x_i;\beta)^2$$

$$= -\sum_{i=1}^{N} x_{ij} x_{in} p(x_i;\beta)(1 - p(x_i;\beta)).$$

Starting with β^{old} , a single Newton-Raphson update is

$$\beta^{new} = \beta^{old} - \left(\frac{\partial L(\beta)}{\partial\beta\partial\beta^T}\right)^{-1} \frac{\partial L(\beta)}{\partial\beta},$$

Where the derivatives are evaluated at β^{old} .

The iteration can be expressed compactly in matrix form.

Let y be the column vector of y_i .

Let x be the N × (p + 1) input matrix.

Let p be the N-vector of fitted probabilities with ith element

P($x_i;\beta^{old}$).

Let W be an N × N diagonal matrix of weights with ith

Element $p\left(x_i;\beta^{old}\right)\left(1-p\left(x_i;\beta^{old}\right)\right)$.

Then

$$\frac{\partial L(\beta)}{\partial \beta} = X^T (y - p)$$

$$\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta^T} = -X^T WX$$

The Newton-Raphson step is

$$\beta^{new} = \beta^{old} + \left(X^T WX\right)^{-1} X^T (y - p)$$

$$= \left(X^T WX\right)^{-1} X^T W\left(X\beta^{old} + W^{-1}(y - p)\right)$$

$$= \left(X^T WX\right)^{-1} X^T W_Z$$

Where $z \overset{\Delta}{=} X\beta^{old} + W^{-1}(y - p).$

If z is viewed as a response and X is the input matrix, β^{new} is the solution to a weighted least square problem:

$$\beta^{new} \leftarrow \arg\min_{\beta} (z - X\beta)^T W(z - X\beta).$$

Recall that linear regression by least square is to solve

$$\arg\min_{\beta} (z - X\beta)^T (z - X\beta).$$

z is referred to as the adjusted response. The algorithm is referred to as iteratively reweighted least squares or IRLS

3.2.5 GENERATING THE LOGISTIC REGRESSION MODEL

The first task in model estimation is to transform the independent variable and determine the coefficients of the independent variables. The basic logistic regression analysis begins with logit transformation of the dependent variable through utilization of maximum likelihood estimation. This is done using the odds ratio. The odds ratio for an event is represented as the probability of the event outcome / (1 - probability of event outcome).

The odds ratio can be described as

$$Odds_i = \left[\frac{P_i}{1-P_i}\right] = e^{\alpha+\beta_1 x_1+.....+\beta_k x_k}$$

Where P i is the probability of an event i,

$\alpha+\beta_1 x_1+......$ represents the regression model

It represents all event probabilities, relationships and their exponential nature. The odds ratio has numerous advantageous properties. It clearly portrays the increased or decreased likelihood of an event outcome occurrence. If the odds ratio is less than one there is a decreased likelihood of an event occurring and if the odds ratio is greater than one then there will be an increased likelihood of the event occurring. The odds ratio provides an intuitive foundation for any sensitivity analysis of interest between the dependent and independent variable. The odds ratio is based on the probabilities that a specific binary outcome will occur when using particular model estimation. It is converted to a continuous function through the logit transformation. The new plot of the transformation of the independent data into probabilities versus the dichotomous dependent data will be continuous ranging from infinity to negative infinity. The log of the odds ratio is known as the logit.

For each data point, $\log it_i$ is represented by

$$\log it_i = In\left[\frac{P_i}{1-P_i}\right]$$

The maximum likelihood estimation (MLE) is now used to estimate the coefficients $(\alpha, \beta_1, \beta_2 ....\beta_p)$ from the logit transformation. MLE is similar to the ordinary least squares used in multiple regression analysis. The likelihood is the probability that the observed values of the dependent variable will be predicted by the observed independent variable data. The log likelihood (LL) is the log of that likelihood and is in the range of infinity to negative infinity. The logistic curve simplifies the coefficient estimation. The maximum likelihood estimate seeks to maximize the LL value and estimate the coefficient found at that maximum point. It is determined through an iterative process that is normally handled by computer software such as SAS, SPSS or Minitab. One point worth mentioning is that MLE is extremely accurate for large sample sizes. Since the LL is the log probability that the dependent variables will be predicted by the observed independent variables, we should seek to maximize that probability. The coefficient estimate where the log likelihood is maximized will represent the best probability that the observed dependent variable is predicted by the observed independent variables. SPSS or some other statistical package is then used to compute the log likelihood and logit transformations to estimate the coefficients for the initial model.

CHAPTER FOUR

DATA ANALYSIS AND DISCUSSION OF RESULTS

4.0 INTRODUCTION

This chapter consists of data analysis (preliminary and regression analysis) and discussion of results.

4.1 PRELIMINARY ANALYSIS

This section gives preliminary analysis of the data and it involves exploration of the data by means of summary statistic and charts.

The purpose is to find out the association between independent variables and the dependent variable. We want to know the credit worthiness of a client base on number of quantifiable borrower's characteristics. The data is a secondary data which consist of both demographic and behavioral characteristics of past (5 years ago) and prospective customers of the bank. The source of data was collected from Atwima Kwawoma Rural Bank at Pakyi Kumasi in Ashanti Region. SPSS Statistical software was used to undertake the sample analysis which included descriptive statistic and analysis of association. The response or dependent variable is called Default, which describes the customer's repayment ability. The bank definition of default is identical to the Basel II framework: the borrower is in default if he is more than 90 days overdue with any payment connected with the loan.

1. SEX

In table 4.1, it was observed that, out of 600 customers who borrowed money from the bank, 330 were men accounting for 55% of the total borrowers, while 270 were women representing 45% of the total borrowers. Number of men and women who paid money normally (non-defaulters) were 249 and 234 respectively. These account for 52.0% (for men) and 48.8% (for women) of the total number of non-defaulter. In all 117 customers defaulted accounting for 19.5% of the total borrowers with men leading with 69.2% of the total borrowers while 30.8% of them were women. Most men defaulted than women since within the Sex, 24.6% of the men who borrowed money defaulted as compare with women 13.3% .

Table 4.1 Cross Table - Payment Status versus Sex

			sex		Total
			female	male	
Payment Status	non-default	Count	234	249	483
		% within payment Status	48.4%	52%	100.0%
		% within sex	86.7%	75.5%	80.5%
		% of Total	39.0%	41.5%	80.5%
	Default	Count	36	81	117
		% within payment Status	30.8%	69.2%	100.0%
		% within sex	13.3%	24.6%	19.5%
		% of Total	6.0%	13.5%	19.5%
Total		Count	270	330	600
		% within payment Status	45.0%	55.0%	100.0%
		% within sex	100%	100%	100%
		% of Total	45.0%	55.0%	100.0%

2. OCCUPATION

Two occupational types were considered; salary earners and self employed. In Figure 4.1, we found out that most of the people who collected loans were salary earners (67.7%). In all 330 out of 600 customers paid their debt normally, with salary earners leading with 39.5% of the total borrowers. 45% of customers defaulted. Out of 406 salary earners who collected loan from the bank, 169 of them defaulted, accounting for 28.2% of the total borrowers. Default rate within self employed is higher than that of salary earners.52.1% of the self Employers defaulted while 41.6% of the salary earners defaulted (Appendix 3.1).

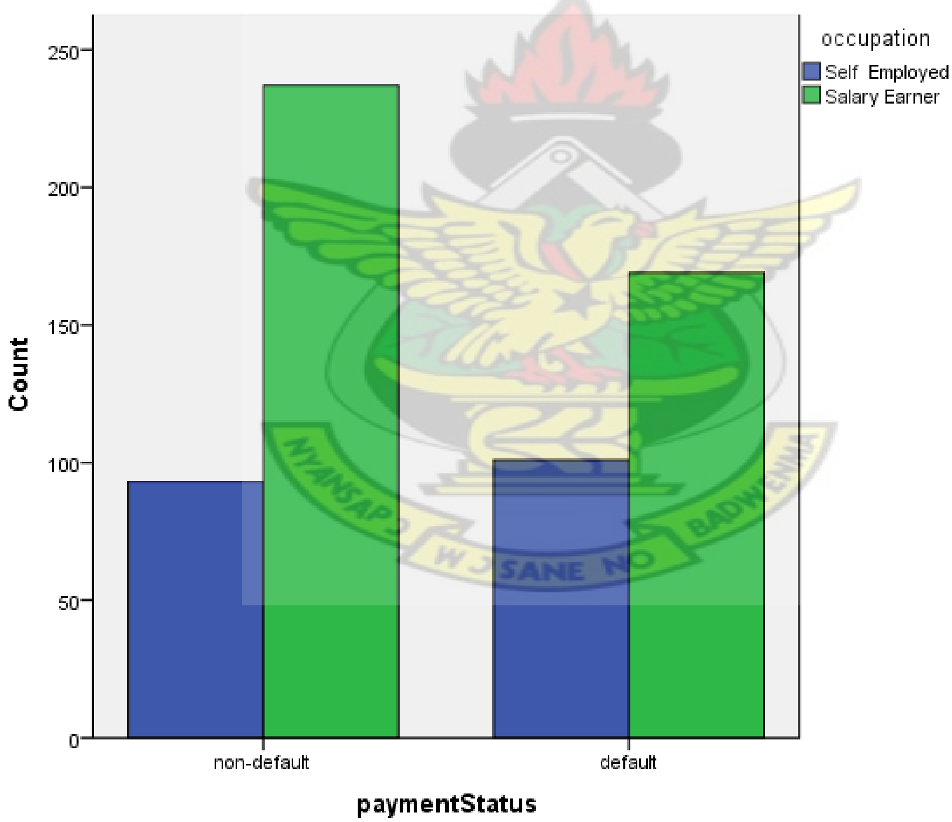


Figure 4.1 payment Status versus occupation

3. MARITAL STATUS

Levels of marital Status are Single and marriage. Marriage borrowers are the most (80.3%) comparing with single borrowers (19.7%).The proportion of defaulters accounts for 46.3% among married borrowers. It is more than proportion of people who defaulted (45.0%). The proportion of default account for 39.8 % among the clients who are single . It is less than the proportion of people who default. Most of the borrowers who defaulted were married; 233 marriage customers (86.2%) out of 270 customers defaulted. Although marriage customers account for (78.5%) of total non-defaulters, only 53.7% of the married customers pay their loan normal (did not default) as shown in Table 4.2.

Marriage people defaulted than Single customers since 46.3% of married defaulted while 39.8% of Single customers defaulted.

Table 4.2 Cross table-payment Status versus marital Status

			Marital Status		Total
			single	married	
Payment Status	non-default	Count	71	259	330
		% within payment Status	21.5%	78.5%	100.0%
		% within marital Status	60.2%	53.7%	55.0%
		% of Total	11.8%	43.2%	55.0%
	default	Count	47	223	270
		% within payment Status	17.4%	82.6%	100.0%
		% within marital Status	39.8%	46.3%	45.0%
		% of Total	7.8%	37.2%	45.0%
Total		Count	118	482	600
		% within payment Status	19.7%	80.3%	100.0%
		% within marital Status	100.0%	100.0%	100.0%
		% of Total	19.7%	80.3%	100.0%

4. NUMBER OF DEPENDENTS

Customers with no dependents were most (156), which accounted for 26.0% of borrowers, followed by those with more than three dependents and then customers with exactly three dependents (150 and 137 customers respectively).customers with only one dependents (69 customers) which accounted for only 11.5% were less than those with two dependents of 88 borrower accounting for 14.7% as show in (Appendix 3.2). Customers with more than three dependent defaulted most, since 29.3% of customers within that number of dependents defaulted. The percentage of customers who defaulted for three, two, one and no dependents are 25.5% ,21.6%,18.6% and 16% respectively.

5. AMOUNT OF LOAN COLLECTED

Most customers (55.2%) collected a loan within the range of ¢1001-¢5000. (refer to figure 4.3 Below). The next range of amount of loan collected by customers is ¢5001-¢10000 accounting for 40.5%. Only 3 people collected less than 500 accounting for 0.5% .Out of 331 customers who collected the loan, 58.3% of them defaulted. All the customers who collected an amount of loan less than ¢500 and amount ranging from ¢501-¢1000 paid their debt normally (non-default) as shown in Appendix 3.3.

Customers who borrowed ¢5,001.00-¢10,000.00 had highest percentage of default since 58.8% of its members who collected the loan defaulted. People who collected loans ¢1,001.00-¢5,000.00 defaulted more than those who collected above ¢10,000.00. Customers who borrowed less than ¢500 and ¢501-¢1000 did not default.

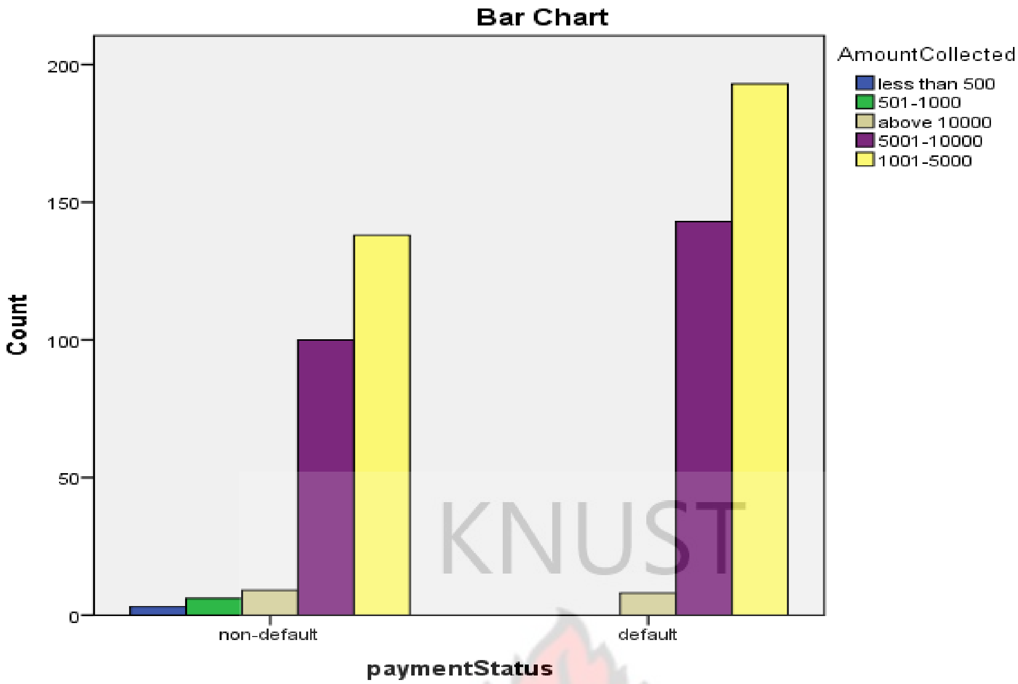


Fig 4.2 payment Status versus Amount Collected

6. AGE

Out 600 customers who collected the loan, 200 of them were 30-39 years old constituting higher number of borrowers (33.3% of the total number of borrowers). Only 33 customers were 20-29 years. In all, 165 defaulters were recorded with age 30-39 years leading with 38.8% followed by customers 40-49 years (25.5%). Age 20-29 years had the least number of defaulters (3.1%). 32% of the customers within the age 30-39 years defaulted. Out of 61 customers above 60 years, 18 of them defaulted accounting for 29.5%.With in age 40-49 years,25.6% of them defaulted. Age 20-29 years had least default rate since only 15.2% of its members defaulted.(refer to Table 3.4)

Table 4.3 Cross table - payment Status versus age

			Age in years					Total
			20-29	30-39	40-49	50-60	Above 60	
Payment Status	non-default	Count	28	136	122	106	43	435
		% within payment Status	6.4%	31.3%	28%	24.4%	9.9%	100.0%
		% within age	84.4%	68%	74.4%	74.7%	70.5%	72.5%
		% of Total	4.7%	22.7%	20.3%	17.7%	7.2%	72.5%
	default	Count	5	64	42	36	18	165
		% within payment Status	3.1%	38.8%	25.5%	21.8%	10.9%	100.0%
		% within age	15.2%	32%	25.6%	25.3%	29.5%	27.5%
		% of Total	0.8%	10.7%	7%	6%	3%	27.5%
Total		Count	33	200	164	142	61	600
		% within payment Status	5.5%	33.3%	27.3%	23.7%	10.2%	100.0%
		% within age	100.0%	100.0%	100.%	100.0%	100.0%	100.0%
		% of Total	5.5%	33.3%	27.3%	23.7%	10.2%	100.0%

4.2 REGRESSION ANALYSIS

In this section, behavioral and demographic factors that contributed immensely in determining characteristics that are indicative of people who are likely to default on a loan were obtained. Observations made in the preliminary analysis were further analyzed with the help of statistical tools such as chi-square goodness -of-fit test and test independent. The logistic regression model was use to determine the credit worthiness of a borrower.

4.2.1 Test for Multicollinearity Among Independent Variables

Multicollinearity in the logistic regression solution is detected by examining the standard error for coefficients (β_i). A standard error larger than 2.0 indicates numerical problems, such as multicollinearity among the independent variables.

Independent variables that indicate numerical problems were dropped before analysis. Hence, none of the independent variables in this analysis had a standard error larger than 2.0.

4.2.2 Over all Relationship Between independent and dependent variables

Let the Hypothesis be stated as follows.

H_0 : there exist no difference between the model with only a constant and the model with independent variables.

H_1 : a relationship exist between the dependent and independent variables.

Table 4.4 Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	116.376	15	.000
	Block	116.376	15	.000
	Model	116.376	15	.000

The presence of a relationship between the dependent variable and combination of independent variable is based on the statistical significance of the model chi-square at step1 after the independent variable have been added to the analysis (Table 4.4)

In the Analysis, the probability of the model chi-square(116.376) was <0.0001,which is less than or equal to the level of significant of 0.05.The null hypothesis that there is no difference between the model with only a constant and the model with independent variables was rejected. The existence of a relationship between the independent variables and the dependent variable was supported, indicating that there was a statistically significant in overall relationship between the independent variables and the dependent variable.

4.2.3 Relationship of individual independent variables to Dependent Variable

4.2.3.1 Significant tests for coefficients of the independent variables (β_I)

Hypothesis

$$H_0: \begin{pmatrix} \beta_1 \\ \beta_2 \\ \\ \\ \\ \beta_I \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \qquad H_1: \begin{pmatrix} \beta_1 \\ \beta_2 \\ \\ \\ \\ \beta_I \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$\beta_1, \beta_2, ..., \beta_i$ represent coefficients of independent variables in the model. Since the p –value for the variable Age(2) was 0.00,less than or equal to the level of significance of 0.05,the null hypothesis that the coefficient for Age(2) was equal to zero was rejected. This supports the relationship that, given the odd ratio (0.288) for Age(2),one unit (one year) increase in Age(2) decrease the odd ratio that a customer will default on a loan by 0.288 times.

Since the p –value for the variable Age(3) was 0.00,less than or equal to the level of significance of 0.05,the null hypothesis that the coefficient for Age(3) was equal to zero was rejected as shown in Appendix 3.4.

This supports the relationship that, given the odd ratio (0.236) for Age(3),one unit (one year) increase in Age(3) decrease the odd ratio that a customer will default on a loan by 0.236 times.

Also the p –value for the variable Age(4) was 0.00,less than or equal to the level of significance of 0.05,the null hypothesis that the coefficient for Age(4) was equal to zero was rejected. This supports the relationship that, given the odd ratio (0.104) for Age(4),one unit (one year) increase in Age(4) decrease the odd ratio that a customer will default on a loan by 0.104 times.

Since the p –value for the variable occu(1) was 0.012,less than or equal to the level o significance of 0.05,the null hypothesis that the coefficient for occu(1) was equal to zero was rejected. This supports the relationship that, given the odd ratio (0.618) for occu(1),one unit increase in occ(1) decrease the odd ratio that a customer will default on a loan by 0.618 times.

The value of 0.012 associated with Occ(1) indicates that we would expect our model's result to deviate significantly from reality only about 12 times out of a thousand if we repeated our model-building process over and over again on new data samples.

Since the p –value for the variable mar(1) was 0.033,less than or equal to the level of significance of 0.05,the null hypothesis that the coefficient for mar(1) was equal to zero was rejected. This supports the relationship that, given the odd ratio (2.092) for mar(1),one unit (one year) increase in mar(1) increase the odd ratio that a customer will default on a loan by 2.092 times.

The value of 0.033 associated with mar(1) indicates that we would expect our model's result to deviate significantly from reality only about 33 times out of a thousand if we repeated our model-building process over and over again on new data samples.

Since the p –value for the variable sex(1) was 0.003,less than or equal to the level o significance of 0.05,the null hypothesis that the coefficient for sex(1) was equal to zero was rejected. This supports the relationship that, given the odd ratio (1.77) for sex(1),one unit increase in sex(1) increase the odd ratio that a customer will default on a loan by 1.77 times.

The value of 0.003 associated with sex(1) indicates that we would expect our model's result to deviate significantly from reality only about 3 times out of a thousand if we repeated our model-building process over and over again on new data samples.

Since the p –value for the variable nud(2) was 0.001,less than or equal to the level o significance of 0.05,the null hypothesis that the coefficient for nud(2) was equal to zero was rejected. This supports the relationship that, given the odd ratio (0.397) for nud(2),one unit increase in nud(2) decrease the odd ratio that a customer will default on a loan by 0.397 times. The value of 0.001 associated with nud(2) indicates that we would expect our model's result to deviate significantly from reality only about 1 times out of a thousand if we repeated our model-building process over and over again on new data samples.

Since the p –value for the variable nud(4) was 0.029,less than or equal to the level of significance of 0.05,the null hypothesis that the coefficient for nud(4) was equal to zero was rejected. This supports the relationship that, given the odd ratio (3.314) for nud(4),one unit increase in nud(4) increase the odd ratio that a customer will default on a loan by 3.314

times. The value of 0.029 associated with nud(4) indicates that we would expect our model's result to deviate significantly from reality only about 29 times out of a thousand if we repeated our model-building process over and over again on new data samples.

Since the p –value for the variable am(2) was 0.036,less than or equal to the level o significance of 0.05,the null hypothesis that the coefficient for am(2) was equal to zero was rejected. This supports the relationship that, given the odd ratio (0.658) for am(2),one unit increase in am(2) decrease the odd ratio that a customer will default on a loan by 0.658 times. The value of 0.036 associated with am(2) indicates that we would expect our model's result to deviate significantly from reality only about 36 times out of a thousand if we repeated our model-building process over and over again on new data samples.

Since the p –value for the variable am(3) was 0.018,less than or equal to the level o significance of 0.05,the null hypothesis that the coefficient for am(3) was equal to zero was rejected. This supports the relationship that, given the odd ratio (4.920) for am(3),one unit increase in am(3) increases the odd ratio that a customer will default on a loan by 4.920 times.

The value of 0.018 associated with am(3) indicates that we would expect our model's result to deviate significantly from reality only about 18 times out of a thousand if we repeated our model-building process over and over again on new data samples.

However, the p-values for remaining independent variables are greater than 0.05 level of significance, hence we fail to reject the null hypothesis that their coefficients of regression are equal to zero, therefore they are not statistically significant. This independent variables are excluded from the model. (refer Appendix 3.4).

4.2.4 STRENGTH OF MODEL RELATIONSHIP

The Nagelkerke R square (0.697) shows in table 4.5 below indicates a relatively strong correlation between the independent variables and the dependent variable (payment Status).It gives an indication that about 70% of the variation of payment status is explained by the independent variables. The model is therefore useful in predicting default.

Table 4.5 Model Summary

-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
359.496	.370	.697

4.2.5 Classification of the model

Classification of the model, that is, how well the model distinguishes Default and Non-Default of a loan, can be assessed using the classification table. As show in Appendix 3.5, the model is able to predict Default and Non-Default adequately (83.9% and 54.1% respectively).Overall, the model has 70.5% accuracy rate of distinqiushing Defaulters from Non-defaulters.

4.2.6 Goodness of fit

Let the Hypothesis be stated as follows.

H_0: The model fits data well

H_1: The model does not fit data (that is, the numbers of observed Defaulters are different from those predicted from the model)

The Hosmer and Lemeshow Test $(P = 0.456)$ is greater than $\alpha = 0.05$ shown in Table 4.6 below.

Therefore we fail to reject the null hypothesis and conclude that number of observed defaulters was not significantly different from those predicted by the model and therefore the overall model is good.

Table 4.6 Hosmer and Lemeshow Test

	Chi-square	df	Sig.
	9.176	8	.456

The Hosmer and Lemeshow Test results ($\chi^2 = 9.176$, 8 degrees of freedom, $P = 0.456$) show n in table 4.6 above indicate that the goodness of fit is satisfactory.

4.2.7 Final model

The best-fit logistic regression model for individual characteristics is represented below

$$Z = In\left(\frac{p}{1-p}\right) = 0.611 - 1.245X_1 - 1.446X_2 - 2.263X_3 + 0.481X_4 + 0.738X_5$$
$$+ 0.571X_6 - 0.924X_7 + 1.198X_8 - 0.419X_9 + 1.593X_{10}$$

Where X_1=Age(2)= $30-39$ years　　X_2=Age(3)= $40-49$ years

X_3=Age(4)= $50-59$ years.　　X_4= Occu(1)=self employed

$X_5 = Mar(1)$ =married　　$X_6 = Sex(1)$ =Male

$X_7 = nud(2)$ = 2 dependents　　$X_8 = nud(4)$ = more than 3 dependents

$X_9 = am(2)$ =Amount collected = ¢501.00-¢1000.00

$X_{10} = am(3)$ =Amount collected = ¢5001.00-¢10000.00

$X_{11} = am(4)$ =Amount collected= ¢1001.00-¢5000.00

$$P(Default = 1) = \frac{1}{1+e^{-z}}$$

where

$$Z = 0.611 - 1.245X_1 - 1.446X_2 - 2.263X_3 + 0.481X_4 + 0.738X_5 +$$

$$0.571X_6 - 0.924X_7 + 1.198X_8 - 0.419X_9 + 1.593X_{10}$$

4.2.8 Logistic Regression Model Calculator

Logistic regression model calculator is designed to calculator the probability of default. Substituting the values of the independent variables which are in the model into the logistic regression model, the probability of default can be calculated.

Example: A customer whose age is 32years,a self employed, married, male, has more than three (3) dependents and collected ¢ 4000.00 will have 0.91 probability of default. Hence need not be giving a loan. (Refer to Appendix 4.1)

4.2.9 Respondent Logistic Regression Model Calculator

It is designed to calculate probability of default and also gives the average probability of default of customers who collected a loan. Appendix 4.2 gives the result of 80 current customers who borrowed money from the bank with their respective probability and the average probability of defaulters

4.3.0 Type I and Type II Error

Let the Hypothesis be stated as follows.

H_O: The borrower will default on a loan

H_1: The borrower will not default on a loan

From Appendix 3.5, the number of defaulters and non-defaulters predicted by the model were 277 and 146 respectively. The number of predicted defaulters that actually did not default was 53 while the number of predicted non-defaulters that defaulted was 124.The total number of defaulters and non-defaulters predicted by the model were 330 and 270 respectively.

False positive Rate= $\frac{FD}{ND} = \frac{53}{199} = 0.266 \approx 0.27$

True positive Rate= $\frac{TD}{D} = \frac{277}{401} = 0.6977 \approx 0.70$

Type I error=1-0.6977=0.30 Type II error=0.266=0.27

Sensitivity= $\frac{TP}{TP+FN} = \frac{277}{277+124} \approx 0.70$

Specification= $\frac{TN}{FP+TN} = \frac{146}{146+53} \approx 0.73$

Positive Predictive value= $\frac{TP}{TP+FP} = \frac{277}{277+53} \approx 0.84$

Negative Predictive value= $\frac{TN}{TN+FN} = \frac{146}{146+124} \approx 0.54$

CHAPTER FIVE

SUMMARY, RECOMMENDATIONS AND CONCLUSION

5.0 INTRODUCTION

This chapter concludes the analysis and gives recommendation that would be necessary to reduce default in the banking industry.

5.1 SUMMARY OF FINDINGS

Sex is an important element in assessing the possibility of breaking contracts. There is clear evidence that women default less frequently on loans possibly because they are more risk averse. This research and other studies show that Sex default rate of women is lower than that of men, but most banks do not normally consider this statement in their score sheets. May be this is due to the notion of equality of men and women. It can be seen that 24.6% of male customers defaulted while only 13.3% of female customers had payment troubles.

The result of this thesis shows that 52.1% of the self employers defaulted while 41.6% of the salary earners defaulted. The default of self employers may due to the fact that the loan collected was not use for the purpose of which it was collected; the amount collected might have used for payment of child school fees, performing of marriage etc. which has a long term retards. Salary earner seems to have lower chance of breaking a contract than a self employed borrower.

This thesis is to investigate whether different marital status can predict default as it is often seen as sign of responsibility, reliability or maturity of borrowers. Convention wisdom holds that the married have lower credit risk than the single since the family bond will lower the chance of breaking a contract due to the sense of responsibility.

In most default literatures the rate of default of Singles is higher than that of the married. The result of this thesis is different from the existing literature. Married people defaulted than Single customers since 46.3% of married Defaulted while 39.8% of Single customers defaulted. This deviation may not be seen as a sign of irresponsible and lack of maturity in the side of married customers, but might cause by the fact that most of the marriage customers are in most default ages groups of 30-39 and 40-49 years.

Number of dependents represents the number of people that the borrower has to support. As the number of dependent increases, so does the pressure on the borrower's income due to higher expenses such as food and day care fees. For example, moving from zero to one dependent increases the default percentage from 16% to 18.6%.

The result of this thesis shows that, the high the number of dependent, the higher the likelihood for one to default. The percentage of default within each number of dependent for zero, one, two, three and more than three are 16% ,18.6% ,21.6% ,25.5% and 29.3% respectively.

It is often assumed that older borrowers are more risk averse and will therefore be less likely to default. The probability to perform well is greater for older people. I expect the risk of default to decline in the later stage of life but did not happen so. It can be seen that 30--39year-olds have the most defaults. Almost 32% of customers in this group will default on their payment. The oldest group also seems to have more payment troubles since 29.5% of customers in this group will default on their payment. The risk of default declines in the later stage of life from 30—39 years, 40—49 years to 50—59 years since their respective rate of default within their group are 32%, 25.6% and 25.3%.

Amounted of loan collected can determine the credit worthiness of a borrower. Customer who collected amount less than ¢1000.00 did not default since 100% of them paid their money normally. Even though most people collected amount within the range above ¢10000.00, greater percentage (58.8%) of them did not pay their money normally. people who collected ¢5001-¢10000 defaulted less than whose who collected above ¢10000.00.

The model has 70.5% accuracy rate of distinqiushing defaulters from non-defaulters. If one was identified as defaulter, he/she has 84% chance of actually defaulting and if a customer is identified as non-defaulter, he/she has 54% chance of actually not defaulting. Type I and Type II errors are 0.30 and 0.27 respectively.

5.2 CONCLUSION

Although the sample is only a small percentage of the Bank's total customer base, the results are promising. Out of 600 customers who borrowed money from the bank, 330 of them were male while 270 were female. The default rate of men is higher than that of the women. The married borrower defaulted more than singles. Default rate increase with increasing number of dependent. People who are self employed have the highest rate of default than salary earners. In this study, model is proposed for credit risk assessment. With the knowledge of logistic Regression, it is possible to derive credit Risk models (logistic Regression Model) to determine the credit worthiness of a borrower.

5.3 RECOMMENDATION TO STAKEHOLDERS

Based on the conclusion of this thesis, the following recommendations are suggested for efficient credit Risk management and improvement of quality of service to customers in Atwima kwawnoma Rural Bank, and the banking industry as a whole.

1. Currently loans application has been reviewed by the system of centralization. It is therefore necessary for the bank to hire qualified personnel to analyze credit risk. They can gather information of clients in the entire bank, of changes in financial environment, of consumption trends and of consumer loans. The personnel can periodically analyze and adjust the weight of every risk factor and control the rate of overdue payment loan in the right moment.
2. The bank should periodically organize entrepreneurship training for clients especially to self employers. This training will help to educate customers on how to use their loan so that the loan given will not be diverted from the purpose of which it was collected.
3. The bank can combine Logistic Regression Model used in this Research and the result from the analysis of cross tabulation to set up a standard of operating for examining credit risks. This can go a long way in determining the creditworthiness of a client.
4. A complete database of all personal loan cases can be set up by the bank. By employing this database to adapt the future policy of personal consumer loans, the strategies of credit extension are improved and the result can be a reference for producing relative new financial products. The bank can keep on seeking characteristic factors influencing the credit risk of applying for consumer loans.

5. Putting them in the model, the error will be reduced and accuracy rates of the model will be improved.
6. After forming a standardized model of assessing credit risks, the bank can simplify the application forms according to characteristic factors influencing credit risk of applicants waiting for approval of loans. Application of loans can even be done online, this process can shorten application time of personal loans and the effectiveness of serving customers and gaining more business can also be achieved.

5.4 RECOMMENDATION FOR FUTURE RESEARCH

Credit companies often use scoring models that tend to classify customers to be either “good” or “bad”. The problem here is that the scoring system does not take into account the society and political environment and also the state of the economy. In following papers it could be interesting to study how default behavior changes when there are hits in the economy and societal and political environment that the research is been conducted

This thesis used logistic regression model for the credit scoring Model. Other method such as linear discriminant analysis, k nearest neighbor, kernel density estimation, and decision trees are also worthy topics for future studies.

Apart from the independent variables (sex, Age, Occupation, marital Status, number of dependents, Amount of loan Collected) used in the thesis, other independent variables such as level of education, number of years with current employer, number of years at current address and Household income could also be used by future researcher to see if they (the variables) can determine the creditworthiness of a borrower.

This research only studies practical data of the case bank and the rationality of weight of variable is not discussed. Further studies can go on in this direction and thus might lead to a better predictive power in a model of assessing credit risks.

BIBLIOGRAPHY

Abdou H., Pointon J., El-Masry A., (2008), "Neural nets versus conventional techniques in credit scoring in Egyptian banking", pp. 1275–1292.

Arminger, G., Enache, D. & Bonne, T. (1997). Analyzing Credit Risk Data: A Comparison of Logistic Discrimination, Classification Tree Analysis, and Feed forward Network. *Computational Statistics*, Vol. 12, Issue 2, p. 293-310.

Altman, E. I. and Saunders, A., (1998), "Credit risk measurement: Developments over the last 20 years", Journal of Banking and Finance, Vol. 21, pp. 11-12 and 1721-1742.

Altman, E. I., (1068), "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy", Journal of Finance, vol. 4, p. 23.

Ang, J. S., Chua, J. H., and Bowling, C. H., (1979). "The profiles of late paying consumer loan borrowers", Journal of Money, Credit and Banking, Vol. 11 p. 2.

Beaver, W., (1967), "Financial ratios predictors of failure", Journal of Accounting Research, Supplement to Vol. 4.

Bee W. Y., Seng H. O., and Huselina M. H., (2011), "Expert systems with Applications" Vol. 38, pp 13274-13283.

Black, F., Cox, J., (1976), "Valuing corporate securities: Liabilities: Some effects of bond indenture provisions", Journal of Finance, Vol. 31, pp. 351–367.

Daily Guide, (2011), "Loan Defaulters To Be Prosecuted", Daily Guide - Business and Finance, Thu, 04 Aug 2011.

David G .K., Mitcel K., (2002). “Logistic Regression: A Self-learning Text”, Ed. 2, Springer-verlay Inc, New York, pp. 7-20.

Dinh, T. H. T., and Kleimeier, S., (2007). “A credit scoring model for Vietnam’s retail banking market”. Journal: International Review of Financial Analysis, Vol. 16, pp 471–495.

Duffie, D., (2005), “Credit risk modeling with affine processes”, Journal of Banking and Finance, Vol. 29, pp. 2751 - 2802.

Dunn, L. F. & Kim, T. (1999). An Empirical Investigation of Credit Card Default. Working Paper, Ohio State University, Department of Economics, 99-13.

Essel T. T, Michael A. N., (2001) “Effectiveness of institutional credit for rural development in africa:A control-case study of rural banks in Ghana”, p. 2-10.

FAO, (1994), "Rural Poverty in the 1990s, in Rural Development”, SpeciaL No. 15, Universities Press, Accra.

Frydman, H., Schuermann, T., (2008). “Credit rating dynamics and Markov mixture models”, Journal of banking and finance, vol. 32, pp. 1062-1075.

Gang W, Jian M., (2011), “Study of corporate credit risk prediction”, Hefei, Anhui , pp 22-31.

Geske, R., (1977), “The valuation of corporate liabilities as compound options”, Journal of Financial Economics, vol. 7, pp. 63–81.

Ghana News Agency, (2005), “National Rural Bank Managers' conference opens”, GNA-Business and Finance News, Fri, 22 Apr 2005.

Golding, D. (1992), "'A Social and Programmatic History of Risk Research' In *Social Theories of Risk"*, Praeger, pp. 23 – 52.

Hand, D. J. and Henley, W. E., (1997), "Statistical classification methods in consumer credit scoring: A review", Journal of the Royal Statistical Society: Series A. Statistics in Society, Vol. 160(3), pp. 523–541.

Ho, G-C., (2002), "Credit Risks of Small Amount Consumer Credit Loans", National Sun Yat-sen University.

Huang, C.-L., Chen, M.-C., and Wang, C.-J., (2007), "Credit scoring with a data mining approach based on support vector machines. Expert System with Applications", vol. 37, pp. 847–856.

Huang, Z., Chen, H., Hsu, C.-J., Chen, W.-H., & Wu, S. (2004), "Credit rating analysis with support vector machines and neural networks: A market comparative study", pp. 543–558.

Jian M, Lihua H., and Kaiquan X.,(2012), "Knowledge-Based Systems", Vol. 26, pp. 61-68.

Journal of the Operational Research Society, (2005), "Data mining feature selection for credit scoring models", vol. 56, pp. 1099-1108.

Koh, H. C., Tan, W. C., and Goh, C. P., (2004), "Credit scoring using data mining Techniques", Journal: Singapore Management Review, vol. 26, pp. 25–47.

Krimsky, S., and PloughS., (1988), "*Environmental Hazards : Communicating Risks as a Social Process*. Dover", Mass. : Auburn House.

Lean Y., Shouyang W., and Kin K. L., (2008), "Credit risk assessment with a multistage neural network ensemble learning approach. Expert Systems with Applications," vol. 34. Pp. 1434–1444.

Lee, T.-S., Chiu, C.-C., Chou, Y.-C., and Lu, C.-J. (2006), "Mining the customer credit using classification and regression tree and multivariate adaptive regression splines", Computational Statistics and Data Analysis, p. 50, pp. 1113–1130.

Lin, S.L., (2009). "A new two-stage hybrid approach of credit risk in banking industry. Expert Systems with Applications," vol. 36(4), pp. 8333-8341.

Liu Y., and Schumann M., (2005), "Hybrid mining approach in the design of credit scoring models. Expert Systems with Applications," Vol. 28, pp 655-665.

Lu M.-H., (2000), "Rating Models of Credit Extension from Banks-Using Logistic Regression," National Chengchi University.

Jacobson, T. & Roszbach. K. (2003). Bank lending policy, credit scoring and value-at-risk. *Journal of Banking & Finance*, Vol. 27, Issue 4, p. 615-633.

Mavri, M., Angelis, V., and Loannou, G., (2008), "A two-stage dynamic credit scoring model based on customers profiles and time horizon". Journal of Financial Services Marketing, vol. 13(1), pp.17–27.

Merton, R.C., (1974), "On the pricing of corporate debt: The risk structure of interest rates", The Journal of Finance, vol. 29 (2), pp. 449 – 470.

Mileris, R., (2011), "Engineering Economics", Kaunas University of Technology, vol. 22, Kaunas, pp. 126 -133.

Mohammed A., Robert, H., (2006), "Credit risk, capital structure and lending decisions of banks in Ghana. Journal: Banks and Bank Systems", Vol. 1, pp. 1-10.

Quainoo, A.A., (1997). "A Strategy for Poverty Reduction through Micro-Finance: Experience, Capacities and Prospects," Accra: UNDP, African Development Bank, World Bank.

Rock, A., (1984), “Sure Ways to Score with Lenders”, pp. 8-10.

Steenackers, A., and Goovaerts, M. J., (1989), “A Credit Scoring Model for Personal Loans, Insurance and Mathematics Economics,” pp. 31-34.

Suleiman M., (2011), “Heading Bad Loans Choke Banks”, Daily Graphic, Business News, 16th March 2011.

Thompson, K.M., Deisler, P.F., and Schwing, R.C., (2005), “‘Interdisciplinary Vision: The First 25 Years of the Society for Risk Analysis (Sra), 1980 - 2005)’.” *Risk Analysis, vol.* 25, no. 6, pp. 1333 - 1386.

Thompson, P.B., (1986), “The Philosophical Foundations of Risk.” Southern Journal of Philosophy, vol. 24, pp. 273 - 86.

Updegrave, W. L., (1987), “How Lenders Size You up, Money (Apr.)”, p. 15.

Vojtek, M., and Kocenda, E., (2006), “Credit scoring methods”, Czech Journal of Economics and Finance, Vol. 56 (3–7), pp. 152–167.

West, D., (2000), “Neural network scoring models. Computer and Operations Research,” Journal: vol. 27, pp. 1131–1152.

William F. S., and David O. A., (2003), “Rural and Micro Finance Regulation in Ghana: Implications for Development and Performance of the Industry”, Africa Region Working Paper Series No. 49, World Bank.

William F. T., Mark C., (2000), “Credit risk rating systems at large US banks”, Journal of Banking and Finance, vol. 24, pp. 167-201.

Xiaolin S., Xuezhi Q., and Bo C., (2009), “International conference on management on e-commerce", Sch. of Manage., Dalian Univ. of Technol., Dalian, China, pp. 258-261

Yao P. and Lu Y., (2011), "Neighborhood rough set and SVM based hybrid credit scoring classifier. Expert Systems with Applications" vol. 38, pp. 11300–11304.

Yi, P., Guoxun W., Gang K. and Yong S., (2010), "An empirical study of classification algorithm evaluation for financial risk prediction", Beijing, p 21.

.

APPENDICES

Appendix A: LICENCED RURAL BANKS IN GHANA AS AT FEBRURY 2011

LIST OF RURAL BANK

REGIONAL DISTRIBUTION

ASHANTI REGION	LOCATION
1. Atwima Rural Bank Ltd	Foase
2. Sekyedomase Rural Bank Ltd	Sekyedomase
3. Adansi Rural Bank Ltd	Fomena
4. Asokore Rural Bank Ltd	Asokore
5. Kwamaman Rural Bank Ltd	Kwaman
6. Asante Akyem Rural Bank Ltd	Juansa
7. Kumawuman Rural Bank Ltd	Kumawu
8. Akrofoum Area Rural Bank Ltd	Akrofuom
9. Ahafo Ano Premier Rural Bank Ltd	Wioso
10. Bosomtwe Rural Bank Ltd	Kuntanase
11. Okomfo Anokye Rural Bank Ltd	Wiamoase
12. Odotobiri Rural Bank Ltd	Jacobu
13. Atwima Kwanwoma Rural Bank Ltd	Pakyi No.2
14. Sekyere Rural Bank Ltd	Jamasi
15. Amanano Rural Bank Ltd	Nyinahin
16. Amansie West Rural Bank Ltd	Antoakrom
17. Juaben Rural Bank Ltd	Juaben
18. Atwima Mponua Rural Bank Ltd	Toase

19. Nwabiagya Rural Bank Ltd	Barrekese
20. Otuasekan Rural Bank Ltd	Kofiase
21. Nsutaman Rural Bank Ltd	Nsuta
22. Offinso Rural Bank	Offinso
23. Ejuraman Rural Bank	Ejura
24. Bosome- Freho Rural Bank	Nsuaem No. 2

CENTRAL REGION LOCATION

1. Nyakrom Rural Bank Ltd	Agona Nyakrom
2. Mfantseman Community Bank Ltd	Biriwa
3. Enyan Denkyira Rural Bank Ltd	Enyan Denkyira
4. Gomoa Rural Bank Ltd	Apam
5. Kakum Rural Bank Ltd	Elmina
6. Nyankumase Ahenkro Rural Bank Ltd	Fante-Nyankumase
7. Union Rural Ban k Ltd	Senya Bereku
8. Assinman Rural Bank Ltd	Assin Manso
9. Brakwa Breman Rural Bank Ltd	Brakwa
10. Ayanfuri Rural Bank Ltd	Ayanfuri
11. Eastern Gomoa Assin Rural Bank Ltd	Gomoa-Dominase
12. Akatakyiman Rural Bank Ltd	Komenda
13. Ekumfiman Rural Bank Ltd	Essuehyia
14. Gomoa Ajumako Rural Bank Ltd	Afransi
15. Agona Rural Bank Ltd	Kwannyaku

16. Akyempim Rural Bank Ltd	Gomoa-Dawuranpong
17. Akoti Rural Bank Ltd	Assin-Akropong
18. Twifu Rural Bank Ltd	Twifo-Agona
19. Awutu Emasa Rural Bank Ltd	Awutu Bereku
20. Awutu Bawjiase Rural Bank Ltd	Bawjias
21. Odupon Kpehe Rural Bank Ltd	Kasoa

EASTERN REGION	LOCATION
1. Asuopra Rural Bank Ltd	Afosu
2. Manya Krobo Rural Bank Ltd	Odumase-Krobo
3. Akwapim Rural Bank Ltd	Mamfe
4. Kwahu Rural Bank Ltd	Pepease
5. Anum Rural Bank Ltd	Anum
6. South Birim Rural Bank Ltd	Achiase
7. Upper Manya Kro Rural Bank Ltd	Asesewa
8. Kwahu Praso Rural Bank Ltd	Kwahu Praso
9. Atiwa Rural Bank Ltd	Kwabeng
10. Mumuadu Rural Bank Ltd	Osino
11. Afram Rural Bank Ltd	Tease
12. Mponua Rural Bank Ltd	Amuana Praso
13. Akim Bosome Rural Bank Ltd	Akim Swedru
14. Kwaebibirim Rural Bank Ltd	Asuom
15. Akyem Mansa Rural Bank Ltd	Ayirebi

16. South Akim Rural Bank Ltd	Nankese
17. Odwen-Anoma Rural Bank Ltd	Hweehwee
18. Dumpong Rural Bank Ltd	Asakraka
19. Adonten Community Bank Ltd	New Tafo
20. Asuogyaman Rural Bank	Akosombo
21. Citizens Rural Bank Ltd	Nsawam
22. Fanteakwa Rural Bank	Begoro

BRONG AHAFO REGION	**LOCATION**
1. Kintampo Rural Bank Ltd	Kintampo
2. Wamfie Rural Bank Ltd	Wamfie
3. Suma Rural Bank Ltd Suma	Ahenkro
4. Baduman Rural Bank Ltd	Badu
5. Asutifi Rural Bank Ltd	Acherensua
6. Nkoranza Kwabre Rural Bank Ltd	Akuma
7. Fiagya Rural Bank Ltd	Busunya
8. Bomaa Rural Bank Ltd	Bomaa
9. Nsoatreman Rural Bank Ltd	Nsoatre
10. Derma Area Rural Bank Ltd	Derma
11. Yapra Rural Bank Ltd	Prang
12. Nkoranman Rural Bank Ltd	Seikwa
13. Amantin & Kasei Rural Bank Ltd	Amantin
14. Ahafo Community Bank Ltd	Kukuom

15. Drobo Community BanK Ltd	New Drobo
16. Nafana Rural Bank Ltd	Sampa
17. Capital Rural bank Ltd	Abesim
18. Atweaban Rural Bank	Duayaw Nkwanta
19. Nkrankwanta Rural Bank	Nkrankwanta
20. Wenchi Rural Bank	Wenchi

WESTERN REGION	**LOCATION**
1. Esiama Rural Bank Ltd	Esiama
2. Amenfiman Rural Bank Ltd	Wassa-Akropong
3. Nzema Manle Rural Bank Ltd	Awiebo
4. Jomoro Rural Bank Ltd	Tikobo No.1
5. Asawinso Rural Bank Ltd	Sefwi-Asawinso
6. Lower Pra Rural Bank Ltd	Shama
7. Fiaseman Rural Bank Ltd	Bogoso
8. Lower Amenfi Rural Bank Ltd	Manso Amenfi
9. Ahantaman Rural Bank Ltd	Agona-Nkwanta
10. Upper Amenfi Rural Bank Ltd	Ankwanso
11. Kaaseman Rural Bank Ltd	Kaase
12. Bia Torya Rural Bank Ltd.	Bonsu Nkwanta
13. Western Rural Bank Ltd	Sekondi
14. Sefwiman Rural Bank Ltd	Bibiani

VOLTA REGION	LOCATION
1. North Tongu Rural Bank Ltd	Adidome
2. Asubonteng Rural Bank Ltd	Worawora
3. Avenor Rural Bank Ltd	Akatsi
4. Unity Rural Bank Ltd	Ziope
5. North Volta Rural Bank Ltd	Guaman
6. Weto Rural Bank Ltd	Kpeve
7. Agave Rural Bank Ltd	Dabala
8. Mepe Area Rural Bank Ltd	Mepe
9. Anlo Rural Bank Ltd	Anloga
10. Butawu Rurak Bank Ltd	Tsito
11. Gbi Rural Bank Ltd	Hohoe
12. Kpassa Rural Bank	Kpassa

GREATER ACCRA REGION	LOCATION
1. Shai Rural Bank Ltd	Dodowa
2. Ada Rural Bank Ltd	Kasseh
3. Dangbe Rural Bank Ltd	Prampram
4. Ga Rural Bank Ltd	Amasaman
5. Abokobi Rural Bank Ltd	Abokobi
6. La Community Bank Ltd	Labadi

UPPER EAST REGION	LOCATION
1. Bessfa Rural Bank Ltd	Garu
2. Naara Rural Bank Ltd	Paga
3. Builsa Community Bank Ltd	Sandema
4. Toende Rural Bank Ltd.	Zebilla
5. Bongo Rural Bank	Bongo

UPPER WEST REGION	LOCATION
1. Nandom Rural Bank Ltd	Nandom
2. Sonzele Rural Bank Ltd	Jirapa
3. Sissala Rural Bank Ltd	Tumu
4. Lawra Area Rural Bank Ltd	Lawra

NORTHERN REGION LOCATION

1. Bonzali Rural Bank Ltd	Kumbugu
2. Bangmarigu Community Bank Ltd	Walewale
3. East Mamprusi Rural Bank Ltd	Gambaga
4. Borimanga Rural Bank Ltd.	Savelugu
5. Buwulonso One Stop RB Ltd	Damongo
6. Tizaa Rural Bank Ltd	Gushegu
7. Zabzugu Rural Bank	Zabzugu

SUMMARY REGIONAL DISTRIBUTION OF RURAL BANKS AS AT AUGUST 2010

REGION	NUMBER
1. ASHANTI	24
2. CENTRAL	21
3. EASTERN	22
4. BRONG AHAFO	20
5. WESTERN	14
6. VOLTA	12
7. GREATER ACCRA	6
8. UPPER EAST	5
9. UPPER WEST	4
10 NORTHERN	7
TOTAL	**135**

APPENDIX 2.0

Research Questionnaire

KWAME NKRUMAN UNIVERSITY OF SCIENCE AND TECNOLOGY

DEPARTMENT OF MATEMATICS

RESEARCH QUESTIONNAIRE

TOPIC: CREDIT RISK MANAGEMENT IN BANKING INDUSTRY

CASE STUDY: ATWIMA KWANWOMA RURAL BANK

The following Questionnaire has been produced as part of a research thesis .All information received will be treated with strict confident .

CUSTOMERS DEMOGRAPH AND BEHAVORIAL CHARCTERISTICS

1. SEX:

[ ] Male [ ] Female

2. AGE: [ ] 20 years-29years [ ] 30years—39years

[ ] 40-49years [ ] 50years-59years [ ] > 60years

3. OCCUPATION:

[ ] Salary Earner [ ] self employed

4. MARITAL STATUS:

[ ] Single [ ] Married

5. NUMBER OF DEPENDENTS

[ ] None [ ] 1 [ ] 2 [ ] 3 [ ] 4

[ ] more than 4

6. AMOUNT COLLECTED

[ ] less than GH¢500.00 [ ] GH ¢500.00 - GH ¢1000.00

[ ] GH ¢1001.00-GH ¢5,000.00 [ ] GH ¢5,001.00 - GH¢10,000.00

[ ] more than GH¢10,000

7. PAYMENT STATUS

[ ] Default [ ] Non-Default

Thank you for your cooperation

APPENDIX 3.0

Appendix 3.1 Cross table- payment Status versus occupation

			occupation		Total
			Self Employed	Salary Earner	
Payment Status	non-default	Count	93	237	330
		% within payment Status	28.2%	71.8%	100.0%
		% within occupation	47.9%	58.4%	55.0%
		% of Total	15.5%	39.5%	55.0%
	default	Count	101	169	270
		% within payment Status	37.4%	62.6%	100.0%
		% within occupation	52.1%	41.6%	45.0%
		% of Total	16.8%	28.2%	45.0%
Total		Count	194	406	600
		% within payment Status	32.3%	67.7%	100.0%
		% within occupation	100.0%	100.0%	100.0%
		% of Total	32.3%	67.7%	100.0%

Appendix 3.2 Cross table - payment Status versus number of Dependent

			Number of Dependent					Total
			0	1	2	3	4	
Payment Status	non-default	Count	131	56	69	102	106	464
		% within payment Status	22.0%	28.2%	14.9%	12.1%	22.8%	100.0%
		% within numberof Dependent	84.0%	81.2%	78.4%	74.5%	70.7%	77.3%
	default	Count	25	13	19	35	44	136
		% within payment Status	18.4%	9.6%	14%	25.7%	32.4%	100.0%
		% within numberofDependent	16%	18.6%	21.6%	25.5%	29.3%	22.7%
Total		Count	156	69	88	137	150	600
		% within payment Status	26.0%	11.5 %	14.7%	22.8%	25.0%	100.0%
		% within numberofDependent	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

0-no dependent 1-one Dependent 2- Two Dependent 3- three Dependent 4 –more than three Dependent.

Appendix 3.3 Cross table- payment Status versus Amount of money

			Amount Collected					Total
			less than ¢500	¢501-¢1000	¢1001-¢5000	¢5001-¢10000	Above ¢10000	
Payment Status	non-default	Count	3	6	9	100	138	256
		% within payment Status	1.2%	2.3%	3.5%	39.1%	53.9%	100.0%
		% within Amount Collected	100.0%	100.0%	52.9%	41.2%	41.7%	42.7%
	default	Count	0	0	8	143	193	344
		% within payment Status	0.0%	0.0%	2.3%	41.6%	56.1%	100.0%
		% within Amount Collected	0.0%	0.0%	47.1%	58.8%	58.3%	57.3%
Total		Count	3	6	17	243	331	600
		% within payment Status	0.5%	1.0%	2.8%	40.5%	55.2%	100.0%
		% within Amount Collected	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Appendix 3.4 SPSS Analysis output

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I .for EXP(B)	
							Lower	Upper
age			39.716	4	.000			
age(1)	-.277	.382	.527	1	.468	.758	.359	1.602
age(2)	-1.245	.332	14.102	1	.000	.288	.150	.551
age(3)	-1.446	.297	23.710	1	.000	.236	.132	.421
age(4)	-2.263	.624	13.140	1	.000	.104	.031	.354
occ(1)	-.481	.191	6.356	1	.012	.618	.425	.898
mar(1)	.738	.346	4.555	1	.033	2.092	1.062	4.121
sex(1)	.571	.195	8.595	1	.003	1.77	.386	.828
nud			27.153	4	.000			
nud(1)	.530	.647	.671	1	.413	1.699	.478	6.041
nud(2)	-.924	.267	11.958	1	.001	.397	.235	.670
nud(3)	-.452	.332	1.857	1	.173	.636	.332	1.219
nud(4)	1.198	.550	4.743	1	.029	3.314	1.127	9.744
Am			14.389	4	.006			
am(1)	-.045	1.294	.001	1	.972	.956	.076	12.084
am(2)	-.419	.200	4.377	1	.036	.658	.444	.974
am(3)	1.593	.672	5.630	1	.018	4.920	1.319	18.347
am(4)	2.108	1.170	3.249	1	.071	8.233	.832	81.492
Constant	0.611	.307	27.441	1	.000	1.842		

Appendix 3.5 Classification Table

Observed		Predicted		
		Payment Status		Percentage Correct
		Default	Non-Default	
Payment Status	Default	277	53	83.9
	Non-Default	124	146	54.1
Overall Percentage				70.5

APPENDIX 4.0

Appendix 4.1 Logistic Regression Model Calculator

X1 = Age 30-39 yrs X3 = 50-59 yrs X5=Married X7 = 2 dependents X9= Amount collected = ¢501.00-¢1000.00	X2= 40-49 yrs X4= Self Employed X6= Male X8= more than 3 Dependents X10= Amount collected = ¢5001.00-¢10000.00

Note

Enter
ONE **(1)** if the person is within the variable range given above, else
ZERO (**0**) If he doesn't fall within the category
All those who are not captured in this model are classified to be
ZERO (0)

X1 =	1	X2=	0
X3 =	0	X4=	1
X5=	1	X6=	1
X7 =	0	X8=	1
X9=	0	X10=	0

P(Default= 1) =	0.913251644

Enter (1) ones in each category like Age, Employment, etc. ONE **(1)** if the person is within the variable range given in the Table below,else ZERO (**0**) If she/he doesn't fall within the category

AVERAGE P(DEFAULT = 1) =	**0.494990905**

Respondent Number	Age			Employment		Marital Status		Gender		Dependents		Amount to be collected			P(Default=1)
	30 -39	40 -49	50 -59	Self	Employee	Single	Married	Male	Female	2	> 3	501-1000	1001-5000	5001-10001	
	X1	**x2**	**X3**	**X4**			**X5**	**X6**		**X7**	**X8**	**X9**		**X10**	
1	1				1	1		1				1			0.381779966
2		1		1			1	1			1	1		0	0.849923349
3			1		1	1			1						0.160838827
4			1		1						1				0.388410152
5			1		1		1		1		1		1		0.570526603
6	1			1		1		1		1		1			0.283940206
7		1			1	1			1		1			1	0.876099408
8		1		1										1	0.775389902
9		1		1		1		1		1			1		0.330262069
10			1		1		1		1	1			1		0.137288001
11	1			1			1	1		1		1			0.453385759
12				1		1			1			1			0.662174584
13	1			1			1	1			1		1	0	0.913251644
14			1	1			1	1			1	1			0.714430441
15								1							0.765307219
16				1		1			1				1		0.748758148
17		1		1			1		1			1			0.491250893
18			1				1	1		1					0.219771235
19		1		1		1		1		1		1			0.244900042
20	1			1									1		0.461824442
21		1												1	0.680919355
22			1												0.160838827
23	1										1				0.637377557
24		1		1		1		1			1	1			0.730271404
25			1	1		1			1	1		1			0.074882537

Respondent Number	Age			Employment		Marital Status		Gender		Dependents		Amount to be collected			P(Default=1)
	30 - 39	40 - 49	50 -59	Self	Employee	Single	Married	Male	Female	2	> 3	501-1000	1001-5000	5001-10001	
	X1	x2	X3	X4			X5	X6		X7	X8	X9		X11	
56	1				1	1			1		1		1		0.637377557
57		1		1			1		1	1		1			0.277078456
58		1		1		1					1	1			0.604679085
59			1		1			1		1		1			0.027491552
60	1								1		1				0.637377557
61					1		1					1			0.717075285
62						1			1	1			1		0.422382642
63			1		1	1		1					1		0.097704011
64			1	1		1			1	1					0.109583745
65			1	1		1			1		1		1		0.50674959
66			1	1			1				1		1		0.682438302
67		1			1	1						1			0.222008488
68		1		1			1	1			1	1			0.643823907
69			1		1		1	1				1			0.129656423
70					1		1	1			1		1		0.878254114
71			1		1		1	1				1			0.129656423
72		1			1	1			1		1		1		0.589766455
73	1				1	1			1					1	0.722921544
74	1			1		1		1			1		1		0.616330065
75	1			1			1		1				1		0.642217093
76	1			1			1		1		1		1		0.85606691
77			1	1		1		1		1			1	1	0.254832986
78			1		1		1	1				1			0.129656423
79		1			1		1	1			1	1			0.527721543
80	1				1		1		1						0.525976591
81	1			1			1		1		1	1			0.796409033
82		1		1		1			1				1		0.412412769
83			1	1		1					1				0.50674959
84					1	1		1			1			1	0.944328198
85		1		1			1		1		1		1		0.82948758

Respondent Number	Age			Employment		Marital Status		Gender		Dependents		Amount to be collected			P(Default=1)
	30 - 39	40 - 49	50 -59	Self	Employee	Single	Married	Male	Female	2	> 3	501-1000	1001-5000	5001-10001	
	X1	**x2**	**X3**	**X4**			**X5**	**X6**		**X7**	**X8**	**X9**		**X10**	
26		1			1		1		1	1			1		0.264832658
27		1						1							0.434380675
28	1							1					1		0.484255207
29	1			0	1	1			1		1			1	0.896321091
30			1	1			1	1				1			0.43020863
31		1		1		1		1		1					0.330262069
32			1	1			1		1		1		1		0.682438302
33			1		1		1	1							0.415080926
34					1	1		1		1			1		0.564144584
35		1		1		1			1			1			0.315830504
36		1		1		1			1		1	1			0.604679085
37		1		1			1		1	1			1		0.368187582
38					1	1			1				1		0.648168882
39			1	1		1		1		1			1		0.178873254
40		1			1	1		1							0.434380675
41	1			1		1			1	1					0.254074159
42		1			1		1	1			1	1			0.777818743
43	1			1			1		1				1		0.642217093
44		1		1		1		1		1			1		0.330262069
45	1				1	1		1					1		0.484255207
46		1			1		1	1		1				1	0.758230017
47			1		1		1	1		1			1		0.219771235
48					1		1		1		1			1	0.984326712
49	1				1	1			1				1		0.346604104
50		1		1			1	1			1				0.895948783
51			1	1		1		1				1			0.265222234
52		1			1	1		1			1				0.717886094
53			1	1			1		1	1			1		0.204728314
54			1		1		1		1	1		1			0.094747367
55	1			1		1			1	1			1		0.254074159

Appendix 4.2 Respondents Logistic Regression Model Calculator