ESTIMATION OF THE PROBABILITY OF DEFAULT OF CONSUMER CREDIT IN GHANA

CASE STUDY OF AN INTERNATIONAL BANK

BY

EDWARD KOFI AWOTWI

A THESIS SUBMITTED TO THE DEPARTMENT OF MATHEMATICS KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY KUMASI, IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTERS OF SCIENCE

INDUSTRIAL MATHEMATICS

INSTITUTE OF DISTANCE LEARNING

JULY 2011

DECLARATION

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

Edward Kofi Awotwi (20104464)	…….……	……
Student's Name and ID	Signature	Date

Certified by

Mr. K. F. Darkwa	…….	……
Supervisor	Signature	Date

Certified by

Dr. S. K. Amponsah	…….	……
Head Mathematics Department	Signature	Date

Certified by

Prof. I. K. Dontwi	…….	……
Dean IDL	Signature	Date

ABSTRACT

This thesis uses empirical data on customer credit information to model probability of loan default in Ghana. We have constructed the logistic regression model using a dataset from an international bank in Ghana, Bank A. 9939 observations of customers were recorded of which 14% turned out to default their loan. The analyses are performed using logistic regression, with SPSS program. Six variables were found to be highly significant in the model. These are Marital Status, Number of months the applicant has been in current employment, interest rate, tenure of loan, income level and loan amount. The model was used to predict successfully the probability of default of an applicant. Applicants who are not married are 1.24 times more likely to default than those who are married. Lower income earners are more likely to default compared to higher income earners. Those who have been in their current employment for longer period are more likely to repay their loans. A unit increase in the number of months in current employment reduces the probability of default by 0.998.

TABLE OF CONTENTS

DECLARATION……..II

ABSTRACT…….....III

TABLE OF CONTENTS…….....IV

LIST OF TABLES…….....VII

LIST OF A BBREVIATIONS……VIII

DEDICATION…….....IX

ACKNOWLEDGEMENT…….....X

CHAPTER ONE…… …….....1

INTRODUCTION……..1

1.1. Background of Study…….....1

1.2. Problem Statement…….....3

1.3. Objective……4

1.4. Methodology……...4

1.5. Justification...…….....4

1.6. Thesis Organisation……5

CHAPTER TWO ……6

LITERATURE REVIEW……...6

2.1. Introduction……6

2.2. Review of works on Default in Loans……6

2.3 Variables commonly used in loan default models……..16

2.4. PD Modelling Techniques……21

CHAPTER THREES …….....24

METHODOLOGY…….....24

3.0 Introduction……...24

3.1 Logistic regression……24

3.2. The Logistic Function……...25

3.3. The Logistic Model…….....26

3.3.1. Generalized Least Squares……....26

3.3.2. Odds……..28

3.3.4 The Model in Logit Form…….....29

3.4. Estimating α and β…….....30

3.5. Test of correlation between the selected predicted and response variables……....32

3.6. Test of predictive ability of the model……33

3.6.1 Deviance…….....34

3.6.2 Pearson Chi-Square Goodness of Fit Statistic……...34

3.6.3 G Likelihood Ratio Chi-Square Statistic…….....35

3.6.4 Pseudo R^2…….....35

3.6.5. Wald Statistic……36

CHAPTER FOUR…….....37

DATA ANALYSIS……....37

4.1. Introduction…….....37

4.2 Data Description and Summary Statistics……37

4.3. Descriptive Statistics of Variables used in the model……..41

4.4. Computation procedure…….....45

4.5 Results……...46

4.6. Predicting the Probability of default of an applicant……52

CONCLUSION AND RECOMMENDATIONS…….....54

5.1 Conclusion…….....54

5.2 Recommendations…….55

REFRENCES…….....56

LIST OF TABLES

Table 4.1. Definition of Variables used in the Study ….…….39

Table 4.2. Data used in the analysis …….……...40

Table 4.3: Descriptive Statistics for all Applicants……..41

Table 4.4a: Descriptive Statistics for non-defaulted applicants…….....42

Table 4.4b. Descriptive Statistics for defaulted applicants…….....43

Table 4.5: Default rate among, Sex, Marital Status Residence and Loan Amount…….....44

Table 4.6. Loan Default Model……..…..……47

Table 4.7. Variables Selected for the model……48

Table 4.8. Omnibus Tests of Model Coefficients…….....49

Table 4.9. Hosmer and Lemeshow Test……..…….....50

Table 4.10a Block 0 Classification Table…….....50

Table 4.10b Block 1 Classification Table…….....51

Table4.11 Customer Profile…….....52

Table 4.12. Model Illustration. ……..53

LIST OF A BBREVIATIONS

CSM	Credit Scoring Model
E L	Expected Loss
EAD	Exposure at Default
GHC	Ghana Cedis
IRB	Internal Rating Based
LDA	Linear Discriminant Analysis
LGD	Lost Given Default
LR	Logistic Regression
NN	Neural Network
OLS	Ordinary Least Square
PD	Probability of Default

DEDICATION

I dedicate this work to my dear mother MADAM EELIZABETH AWOTWI and the entire family

ACKNOWLEDGEMENT

My sincere appreciations first and foremost go to Almighty God who has seen us through this thesis. To my supervisor Mr. F. K. Darkwa for his useful guidance which enabled me, produce this thesis.

I will also like to thank Dr. S. K. Amposah for his encouragement.

Finally to my families and all others who have contributed in one way or the other to make the writing of this thesis possible.

CHAPTER ONE

INTRODUCTION

1.1.Background of Study

A loan officer at a bank wants to be able to identify characteristics that are indicative of people who are likely to default on loans, and then use those characteristics to discriminate between good and bad credit risks. Lopez and Saidenberg (2000) define credit risk as the degree of value fluctuations in debt instruments and derivatives due to changes in the underlying credit quality of borrowers. Credit-scoring models examine the credit-worthiness of customers by assigning them to various risk groups. These models provide predictions of default probabilities by using statistical classification techniques, and they group them by risk classes. Among the several concepts that help analyze credit risk, Probability of Default is the most critical.

The "building block" for quantifying credit risk is Expected Loss (EL); the loss can be expected from holding an asset. This is calculated as the product of three components: the probability of default (PD), the loss given default (LGD), and the exposure at default (EAD).

EL is defined as follows: EL = PD*LGD*EAD

The probability of default (PD) is defined as the frequency that a loan will default and is expressed in percentage terms. The loss given default (LGD) measures the cost to the financial institution when the loan defaults. It is expressed in percentage terms. The

exposure at default (EAD) is the amount of money outstanding when the default occurs. The ultimate goal is to provide a measure of the loss expected for booking a credit and the capital required to support it. Czuszak (2002) confirms the importance of the probability of default stating that credit risk measurement and management is found in the probability and financial consequences of obligator default.

In recent years, financial institutions have devoted important resources to build statistical models to measure the potential losses in their loan portfolios. The New Basel Capital Accord allows banks to compute the minimum capital requirements using an internal ratings based (IRB) approach which is founded on the most sophisticated credit risk internal models.

The Basel II accord directs international credit system to pay closer attention to measuring and managing credit risk (Hertig 2005). This is true, in particular, for those banks that adopt an Internal Rating Based Approach (IRB). This revision impacts extending credit facility of the mass market.

Much of the academic research on credit risk is focused on the large corporate credit market where data were more easily available to researchers. Research on risk measurement and probability of default modeling for consumer credits has increased in recent years, but this area still remains relatively underdeveloped.

1.2. Problem Statement

Consumer credit and default prediction have been studied relatively little - if at all - in Ghana despite it increases in popularity. Sudden change in income level, unemployment, increases in prices of goods and services and other unexpected occasions are some reasons to apply for a consumer loan to maintain the consumption at the same level. There has also been intense conversation about the nature and morality of consumer credit due to the high costs related to it.

The need of consumer credit today is at its highest, but at the same time the default rate has risen and from the banks' perspective the riskiness of these loans is usually higher than that of a regular bank loan. In this regards, applicants are scrutinized before the credit facility is granted them.

Traditional methods of deciding whether to grant loan to an individual are based on human judgment and experience of previous decisions. These methods are not objective but very subjective. However, to consider every small loan as a separate loan is time consuming and expensive. Usually the lender doesn't have information about the solvency or credit behaviour of a new potential customer and especially in consumer credit business customers are often persons who are applying for a loan for the first time. Thus, to determinate the customer's expected probability of default the lender must estimate his ability to pay back from his current characteristics, as default can only be observed afterwards. Using a statistical approach in estimating the probability of default gives an objective and straight forward approach.

1.3.Objective

The objectives of this study are to

- Model loan default as a logistic regression
- Predict the probability of default for a given customer.

1.4. Methodology

Binomial Logistic Regression will be used to model the loan default. Secondary data on loans (both those who have defaulted and those in good standing) will be taken from an international bank in Ghana BANK A.

SPSS software version 16 will be used to analyze the data. The resources centres for the study are my personal laptop, the internet, the Accra Polytechnic library, University of Ghana library and Kwame Nkrumah University of Science and Technology Library

1.5.Justification

The study is significant in the following ways:

It will help banks and financial institutions to have system that can effectively predict their PD in order to access the needed capital required to book credit.

It will help credit professionals to make accurate, on-the-spot credit decisions within a limited time.

This will also serve as a tool necessary to assist the loan officer in making loan decisions, controlling and monitoring loan portfolio risk and isolating loans that need additional attention.

The model will also assist in the risk evaluation and management process of customers and loan portfolios.

By developing an accurate PD model, banks will be able to identify loans that have lower probability of default versus loans that have a higher probability of default. Thus, they will better rate the loans, price the loans, and may benefit from capital savings.

It will help future researchers who want to research into consumer credit.

1.6.Thesis Organisation

The structure of this thesis is as follows.

Chapter one gives a background introduction on credit risk management and measurement. Chapter two will review relevant literature on the subject matter. Chapter three explains the methodology employed to examine the factors influencing the default of a loan. The empirical analysis is captured in chapter four. Chapter five gives the conclusion and recommendations from the results.

CHAPTER TWO

LITERATURE REVIEW

2.1. Introduction.

Consumer credit markets have been studied relatively little due to the confidential nature of the customer data and the difficulty of measuring the risk appropriately. Most of the studies conducted uses U.S data but the literature and research is evolving also in Asia.

According to Straka (2000) default risk is extremely important due to the automation of decision-making process and the easiness of applying for the loan. A study of automated credit evaluations, development of CSMs has proven to reduce defaults.

2.2. Review of works on Default in Loans

Weizhuo Wang [2001] did a research on default of mortgage loans in China and observed that borrower's demographic characteristics impact banks' lending decision process. His results suggest the Loan amount and Interest rate range are positively correlated with the probability of loan default; which means an increase in the loan amount or interest rate level increases the probability of loan default. Seven variables were identified as having negative impact on the probability of loan default, these are age group between 22 to 59 years old, annual income greater than 36,000 RMB, good bank rate, Occupation such as general manager (occupation type (1)), general staff (occupation type (2)), professional

employee (occupation type (3)), borrower resides within the same district with the bank. He affirmed that a good credit scoring model has the ability to detect bad loans and this could help the bank to reduce the loan losses from loan default. Consequently, it can improve the profitability and the financial stability of the bank. Therefore, the credit scoring model should be developed and used to support credit officers in monitoring the Chinese mortgage loan applications.

In their study on Probability of Default in Agric loans, Amilie and Allen [2006] identified three major financial ratios that significantly influence the probability of default. These are leverage, profitability and liquidity. Length of loan also tested to be highly significant in the default prediction. They observed that differences exist between default models based on the type of farming activity. Thus, concluded it is preferential and more accurate to develop a model for each type of activity though this requires more data to estimate.

Erdem [2008] studied factors affecting the probability of default in credit card in Turkey. Out of 474 credit card users studied, it was found that the number of children, level of education, subjective norm, perceived behavioural control and attitude toward the behaviour were found to be effective in the formation of the behavioural intention.

Wilson et al. (2000) studied payment behaviour prediction of 7034 UK companies with logistic regression. They found that history of payment behaviour is more predictive than accounting data. Their evaluation was implemented with two aspects; that of predicting future payment behaviour and that of corporate failure prediction.

Agarwal et al (2009) assessed the role of individual social capital information characteristics on household default and bankruptcy outcomes. They used monthly panel data set of more than 170 000 credit cardholders for a period of over 24 months. With the observations of each borrower's default and bankruptcy filing status they were able to find distress factors such as riskiness, spending, debt, income, wealth, economic conditions, legal environment and socio-demographical characteristics to significantly affect default. Their results showed that borrowers who migrate from their state of birth default more. Another finding was that a borrower who is married and owns a house of his own has a lower risk of default.

Vasanthi and Raja (2006) estimated the likelihood of default risk associated with income and other factors with Australian data (Australian Bureau of Statistics, ABS 2001) in a sample of 3431 households. The goal was to establish the relationship between the default risk of homeowners and their socio-economic and housing characteristics. The repayment rate is substantially high compared to consumer credit, amounting to 93.03%.

Vasanthi and Raja found out that the age of the head of the household is significant: the younger households tend to be adversely affected by the increasing burden of mortgage payments. Income as socio-demographic variable show to have predictive power: lower income is one of the major contributory factors for default. Another important factor was the loan to value ratio indicating that higher loan to value ratio would increase the probability of default. Also the educational level of the head of household and marital

status had significance impact on default. Vasanthi and Raja drew a conclusion that the probability of default is higher with an uneducated, younger and divorced as head of the family compared to others.

Autio et al. (2009) conducted a comprehensive study of the use of small instant loans in Finland among 1 951 young adults. An open online survey for 18- to 29-year olds included questions about age, gender, financial situation, such as income, employment and occupational status, and family structure. The results showed that the 18- to 23-year-olds use small instant loans more than the 24- to 29-year-olds. The latter group, on the other hand, use consumer credit more, because of their higher income and occupational status, Gender does not seem to have an effect on the number of loans taken, but occupational status, income and household structure do.

Laitinen and Kankaanpää (1999) assessed six alternative methods (LDA, LR, RPA, survival analysis, and HIP) that have been applied to financial failure prediction. The main objective was to study whether the results stemming from the use of alternative methods differ from each other. They used only three financial ratios (total debt to total assets, the ratio of cash to current liabilities and the operating income to total assets) due to methodological issues. The results of 76 randomly selected from Finnish small and medium sized failed firms indicate that no superior method has been found but the predictive power of logistic analysis was best resulting 89.5% prognostic accuracy.

Laitinen 2000) continued the work and tested whether Taylor's series expansion can be used to solve the problem associated with the functional form of bankruptcy prediction models. To avoid the problems associated with the normality of variables, the logistic model which describes the insolvency risk was applied. Several financial ratios were employed with estimation sample including 400 firms and the results suggest that the cash to total assets, cash flow to total assets, and shareholder's equity to total assets ratios operationalize the factors affecting the insolvency risk. The usefulness of Taylor's model in bankruptcy prediction was evaluated applying the logistic regression model to the data.

Sumit Agarwal etal [2008] researched in to the determinant of automobile loan default and observed that automobiles are highly visible consumption goods that are often purchased on credit. In their article, they used a unique proprietary data set of individual automobile loans to assess whether borrower consumption choice reveals information about future loan performance. The result was that an increase in income raises the probability of prepayment, whereas a rise in unemployment increases the probability of default. A decrease in the market rate (the three year Treasury note rate) increases both the probabilities of prepayment and default. They also find that loans on most luxury automobiles have a higher probability of prepayment, while loans on most economy automobiles have a lower probability of default.

According to Marjo H (2010) both socio-demographical and behavioural variables have a notable effect on default. The most significant socio-demographical variables are income, time since last moving, age, possession of credit card, education and nationality. Some behavioural variables seemed to have even more predictive power. Those are the amount of scores the customer obtained, loan size and the information if customer has been granted a loan earlier from the same company. Interestingly, the results have variation to some extent when excluding few of the variables outside the model. The predictive power of all three models is adequate and thus can be employed as a reliable credit scoring model for the credit institutions.

Hayen in 2003 searched univariate regression based on rating models driven for three different default definitions. Two are the Basel II definitions and the third one is the traditional definition. The test results show that there is not much prediction power is lost if the traditional definition is used instead of the alternative once.

In 2000, Hurdle and Muller used a semi parametric regression model called generalized partially linear model and showed that performed better than logistic regression.

In 1980's new method for classifying was introduced by Breiman et al. Which split data into smaller and smaller pieces? Classification and regression tree is an appropriate method for classification of good and bad loans. It is also known as recursive partitioning.

In 1985, Altman, Frydman and Kao presented recursive partitioning to evaluate the predicatively and compared with linear discriminant analysis and concluded that performs better than linear discriminant analysis. In 1997, Pompe compared classification trees with linear discriminant analysis and Neural Network.

The 10-fold cross validation results indicates that decision trees outperform logistic regression but not better than neural networks. Xiu in 2004 tried to build a model for consumers' credit scoring by using classification trees with different sample structure and error costs to find the best classification tree. When a sample was selected one by one, this means that the proportion of good loans is equal to the proportion of bad loans and type I error divided by type II error is equals to the best results were obtained.

According to Adel Lahsasna et al. (2008), during the last fifteen years, soft computing methods have been successfully applied in building powerful and flexible credit scoring models and have been suggested to be a possible alternative to statistical methods.

Kiviloto (1980) used self organizing maps (SOM): a type of neural network, and it was compared with the other two neural network types learning vector quantization and radial basis function and with linear discriminant analysis. As a result like in previous researches, neural network algorithm performed better than discriminant analysis especially the self organizing maps and radial basis functions. Charalombous et al. aimed to compare neural network algorithms such as radial basis function, feed forward

network, learning vector quantization and backpropogation with logistic regression. The result is similar as Kivilioto's study, the neural networks has superior prediction results.

According to Chao-Ying J et al., since 1988, research using logistic regression has been published with increasing frequency in three leading higher education journals: Research in Higher Education, The Review of Higher Education, and The Journal of Higher Education. Yet, there is great variation in the presentation and interpretation of results in these publications, which can make it difficult for readers to understand and compare the results across articles. A systematic review of articles that have used logistic regression not only promotes the learning about this method, but also helps suggest new guidelines for principled applications of this versatile technique. Logistic regression, being one special class of regression models, is well suited for the study of categorical outcome variables, such as staying in or dropping out from college. This technique is increasingly applied in educational research.

Robert M and Thomas S. Y (2006) researched on Valuing Fixed Rate Mortgage Loans with Default and Prepayment Options. They showed that the prepayment default model has significant explanatory power. Using the mortgage loan prices at origination, the model shows that OAS and duration depend on the FICO score, original loan-to-value ratio, the loan size and the recovery ratio. Lastly, a model of the economic value of a loan default guarantee is specified and the model shows that the price elasticity of the guarantee with respect to the loan size and the borrower's FICO score are -0.46 and -11.89 respectively.

According to Lewis (1992) consumer credit has been around for 3000 years since the time of the Babylonians. For the last 750 years of that time there has been an industry in lending to consumers, beginning with the pawn brokers and the usurers of the middle Ages, but the lending to the mass market of consumers in the non-Islamic world is a phenomenon of the last fifty years. In the 1920s, Henry Ford and A.P.Sloan had recognised that it was not enough to produce products, like cars, for the mass market but one also had to develop ways of financing their purchase. This led to the development of finance houses, e.g. GE Capital, GM Finance. The advent of credit cards in the 1960s meant that consumers could finance all their purchases from hair clips to computer chips to holiday trips by credit.

Updegrave (1987) found that there were eight variables that affected consumer credit risk: the number of variables, the historic repayment record, bankruptcy history, work and resident duration, income, occupation, age and the state of savings account. Similar results were found by Steenackers and Goovaerts (1989) who collected data on personal loans in Belgian credit company and found out that age, resident and work duration, the number and duration of loans, district, occupation, phone ownership, working in the public sector or not, monthly income and housing ownership have a significant relationship with repayment behaviour.

Jacobson and Roszbach (2003) contributed to the existing literature by taking into account the sample-selection bias that credit scoring models are suffering from. Therefore the basic value at- risk measure is not reliable enough but they suggest using unbiased scoring model such as bivariate probit approach that also takes into account rejected loans. In their work they used a data set consisting of 13 338 applications for a loan at a major Swedish lending institution between September 1994 and August 1995. All loans were granted in stores where potential customers applied for instant credit to finance the purchase of a consumer good. They had 57 variables available but employed only 16 because they lacked a univariabte relation with the variables of interest of displayed extremely high correlation with another variable. Income, age, change in annual income and amount of collateral-free credit facilities had significant impact on default.

Apilado et al,(1974) applied discriminant analysis to construct their credit scoring models and state that "discriminant analysis firstly distinguishes among group and identifies group differences; secondly, it classifies existing and new observation into predetermined groups, and finally it identifies the key variables that contribute the most to the discrimination among groups" Discriminant analysis was used as a credit scoring tool first by Durand (1941) to produce good predictions of credit repayment. Extensive use of discriminant analysis to build credit scoring models for general banks and credit card sectors has been carried out by Eisenbeis (1983), Martel and Fitts (1981), Grablowsky and Talley(1981), Reichart et al., (1983), Titterington (1992), Desai et al., (1996), Bardos (19980) and Lee et al., (1999).

2.3 Variables commonly used in loan default models

The pragmatism and empiricism of credit scoring implies that any characteristics and environments of the borrower that has obvious connections with default risk should be used in the scoring system (Lewis, 1992). The variables should be sequentially added or deleted to maximise the model's predictive accuracy (Henley and Hand, 1997).

According to Dinh and Kleimeier, (2007), there are two important standards for variable selection; first, the variables should have significant coefficients and contribute to explanation of the dependent variable's variance. Second, the variables should have close correlation with included variables. Lewis (1992) suggests that there is no need to justify the case for any variable. If it helps the predictions, it should be used. However, the major factors commonly used in credit scoring models include the borrowers' income, age, gender, education, occupation, employer type, region, time at present address, residential status, marital status, home phone, collateral value, loan duration, time with bank, number of loans, and current account (Dinh and Kleimeier, 2007; Roszbach, 2004; Jacobson and Roszbach, 2003; Martinelli, 1997; Crook, Hamilton, and Thomas, 1992; Boyes, Hoffman, and Low, 1989; Capon, 1982;)

Income is a commonly used proxy of the borrower's financial wealthy and his/her ability to repay (Dinh and Kleimeier, 2007). There is a positive relationship between income and the borrowers' default rate; higher income is associated with lower default risk (Jacobson and Roszbach, 2003). Occupation is a common variable used in credit scoring model and is highly correlated with the borrowers' income level. Education enhances the borrowers' ability to repay. The better educated borrowers are deemed to have more stable and higher income employment and thus a lower default rate.

The borrowers' education level distinguished from post-graduate to non-high school graduate. Borrowers with high level of education are more likely to repay their loan since they occupy higher positions and with high income levels.

According to Dinh and Kleimeier, (2007), employer refers to the type of company for which a borrower works such as stated-owned, joint-stock company, etc. The type of company a borrower works in could be a proxy for income level and stability. Missing values of this variable are also very informative as borrowers who do not answer this question show the highest probability of default.

Cook et al., (1992) noted that time with employer measures the number of years that the borrower has been working for the current employer. It reflects the satisfaction of the borrower with the current job. The higher the borrowers' job satisfactions, the more stable their employment will be and the higher their ability to repay their loans

According to Capon (1982) length of time with employer may discriminate against

women, since women's length of employment reduces due to pregnancy and childbearing

Age measures the borrower's age in years. Thomas (2000) and Boyle et al. (1992) confirm that older borrowers are more risk adverse, and therefore the less likely to default. Thus banks are more hesitant to lend to younger borrowers who are more risk averse.

Arminger et al., (1997) noted that GENDER in addition to age is one of the most used socio-demographical variables to differentiate the predictive power between men and women. There is clear evidence that women default less frequently on loans possibly because they are more risk averse. According to Coval et al., (2000) gender is a fair discriminatory base on the statistical default rates of men versus women. There are ample evidences that women default less frequently on loans because women are more risk adverse

Region means the area of the country that borrower lives. As people of similar wealth tend to live in the same location, the geographic criterion can indicate a borrower's level of financial wealth. Some suburb might attract richer residents and this could result increase in housing and property prices. This also affects the collateral value and probability of default.

According to Marjo M (2010) the residential variable measures whether borrowers own their home, rent, or live with their parents. This could indicate the borrowers' financial wealth in the case of home ownership. Residential status also indicates financial pressure on borrowers' income, for example rental cost

According to Crook et al.'s (1992) the default risk drops with an increase in time at present address; it might be a proxy for the borrowers' maturity, stability, or risk aversion. Changing address might be a signal that a borrower's financial wealth is high or improving rapidly. Time addresses the number of years that the borrowers have been living at their current address.

According to Dinh and Kleimeir (2007), marital status affects the borrower's level of responsibility, reliability, or maturity. The probability of default is higher for married than single borrowers. They discover that the marital status is typically related to number of dependants which in turn reflects financial pressure on the borrower and borrower's ability to repay a loan.

Gup and Kolari, (2005) Collateral is a form of guarantee to support the loan. Borrowers' collateral can be a single of default risk, such as, if the loans that the house serves as collateral, the probability of default is very low. This is because the borrowers are risk adverse and fear of losing their house. Collateral reduces the bank's risk when it makes a loan. The higher the collateral value the higher the incentive for the borrowers to repay the loan since they do not want to lose their collateral. The collateral value could also be

a proxy for the borrowers' financial wealth since it is significantly positive correlated with the borrowers' income (Dinh and Kleimeier, 2007).

Loan duration indicates the maturity of loans in months. Loan duration reflects the borrowers' intention, risk aversion, or self-assessment of repayment ability. Time with the bank indicates the borrowers' length banking relationship in years. It can be assumed that the longer a borrower stays with the bank, the more the bank knows about this borrower, and it could lower the probability of default. But this variable should be updated regularly due to adverse and unexpected changes in the borrowers' situation.

Number of loans measures the total number of loans a borrower has received from the bank during the whole relationship with the bank. Today, most borrowers have more than one loan from the same bank. This variable reflects the difficulty for a defaulted borrower to receive further loans from the same bank.

Jacobson and Roszbach (2003) explained that Loan Size the amount of credit the applicant is granted. The customer may have applied for larger amount but has been denied the loan. He is able to try lower amount for maximum of three times. Several studies use loan size as a predictor variable but the overall results are ambiguous and thus no clear expectations can be formed. Jacobson and Roszbach (2003) show that loan size has no significant influence on default risk. In the study of Kocenda and Vojtek (2009) small loans appear to be more risky if variable 'own resources' is included. However, if this information is not used, the regression identifies that the larger loans as more risky.

Current account indicates whether the borrower holds a current account with the bank. It partly represents the borrowers' financial wealth, and relationship between the borrower and the bank. The borrowers who hold current accounts with their banks have a lower default risk.

However, Boyes et al. (1989) recognised that if banks were minimising default risk, one should find the above variables with positive (negative) effect on the probability of granting a loan and a negative (positive) effect on default risk.

2.4. Probability of Default (PD) Modelling Techniques

According to Weizhuo(2010), there are several statistical methods used to estimate credit scoring models in assessing borrowers' credits, such as discriminate analysis (Dunn and Frey, 1976), linear probability models (Turvey, 1991), probit models (Lufbuttow et al., 1984) and logit models (Mortensen et al., 1988). The last three methods estimate the default rate based on the historical data on loan performances and the borrowers' characteristics. The idea of linear probability is to look up for a linear combination of explanatory variables. It assumes there is a linear relationship between the default rate and the factors. The probit model assumes the probability of default follows the standard cumulative normal distribution function. The probability of default is logistically distributed in the logit model and discriminant analysis divides borrowers into high and low default-risk classes (Mester, 1997).

Weizhuo(2010), explained that, discriminant analysis presents the critical assessment of the use of discriminant analysis in business. However, Hand et al. (1996b) show that the discriminant function obtained by segmenting a multivariate normal distribution into two classes' optimal discriminant function. Problems also arise in testing for the significance of individual variables when the assumption of normality does not hold and therefore we cannot perform statistical inferences (Rosenberg and Gleit, 1994).

According to Collins and Green (1982) the linear probability model could present reasonable prediction results compared to discriminant analysis and logit models. However, Pyndick and Rubinfeld (1998), Greene (1997), and Judge et al. (1985) indicate that the linear probability model could predict the default rate, but the predictive value might not necessary lie between zero and one. Moreover, because the variance of the models are generally heteroscedasticity, it leads to inconsistent estimation problem and invalid conventional measure of fit such as the R^2.

According to Hand and Henley (1997), the logistic approach is a more appropriate statistical tool than linear regression, when there are two discrete classes (good and bad risks) defined in the model. This gives the logistic approach superior classification rate. The probit model is very similar to the logit model. The logit model is generally preferred to the probit model because of its simplicity (Barney et al., 1999; Novak and LaDue, 1999; Lee and Jung, 1999)

Clarke, (2005) explained that, logistic modelling approach is commonly used to model the bank's lending decision. According to Collins and Green (1982), the logit model can increase the overall classification rate, and substantially reduce the error rate. The logistic approach also gives superior classification compare to discriminant analysis (Wiginto, 1980). According to the literature, there is no best method for estimating credit scoring models and new methods continue to evolve. However, the logit models and neural networks have been applied frequently in previous research.

CHAPTER THREES

METHODOLOGY

3.0 Introduction

This chapter develops a suitable methodology for modeling the probability of default in loan. Relevant mathematical techniques will be presented.

3.1 Logistic regression

In some regression situations, the response variable y has only two possible outcomes, for example, high blood pressure or low blood pressure, developing cancer of the oesophagus or not developing it, whether a crime will be solved or not solved, and whether a bee specimen is a "killer" bee or a domestic honey bee. In such cases, the outcome y can be coded as 0 or 1 and we wish to predict the outcome (or the probability of the outcome) on the basis of one or more independent variables $x's$

Logistic is a mathematical modelling approach that can be used in describing the relationship of several independent variables to a dichotomous dependent variable.

3.2. The Logistic Function

The logistic function describes the mathematical form on which the logistic model is based. The logistic function $f(z)$ is given by

$$f(z) = \frac{1}{1+e^{-z}} \quad \text{.....(3.1.)}$$

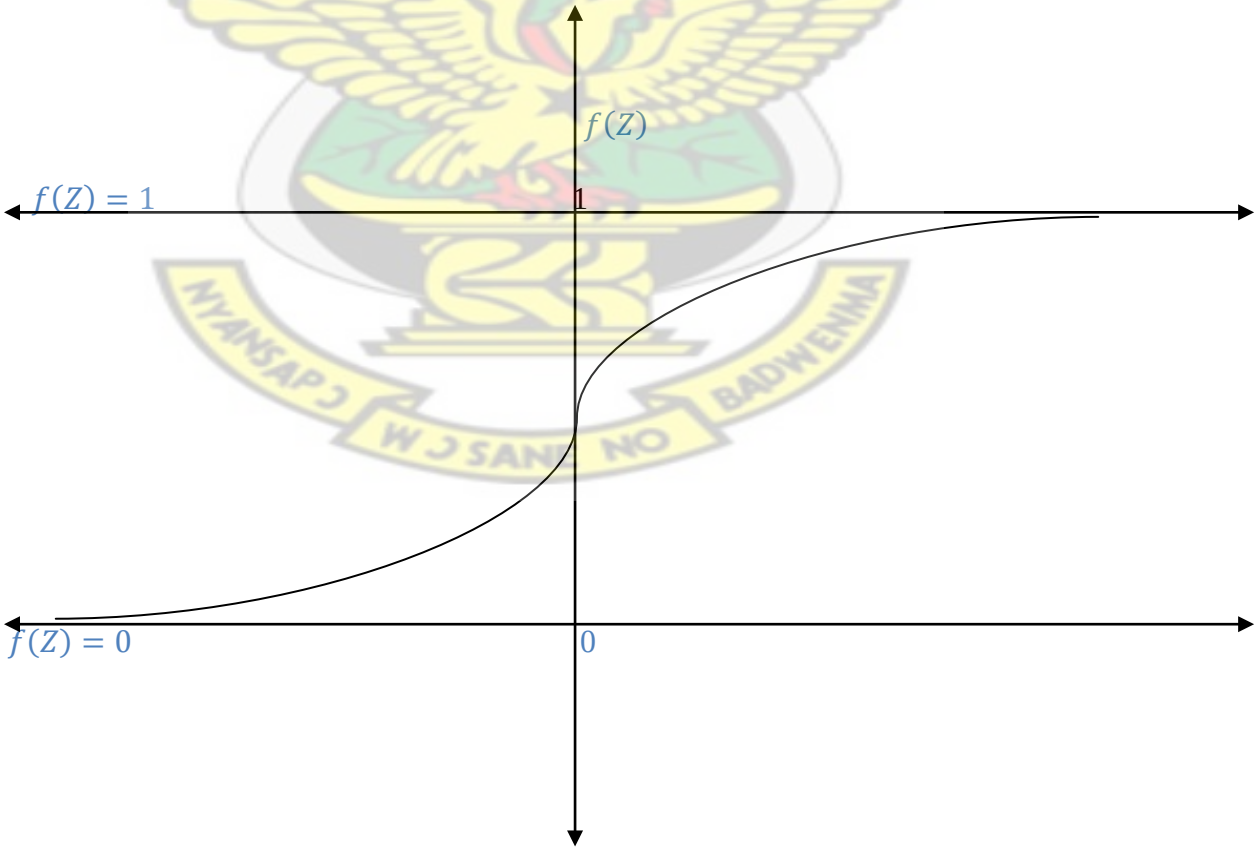
Where

$$z = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots\ldots \beta_n X_n \quad \text{.....(3.2)}$$

And

$x_1, x_2, x_3 \ldots\ldots x_n$, are the independent variables and α, β_2, β_3 ... are the constant term.

The function $f(z)$ can be represented graphically as



3.3. The Logistic Model

Consider a linear model

$y_i = \beta_o + \beta_1 x_i + \varepsilon_i$…….……(3.3)

Where $y_i = 0,1 \; and \; i = 1,2, \ldots \ldots n$

Since y_i is 0 or 1, the mean $E(y_i)$ for each x_i becomes the proportion of observations at x_i for which $y_i = 1$.

This can be expressed as

$E(y_i) = P(y_i = 1) = p_i$ and $1 - E(y_i) = P(y_i = 0) = 1 - p_i$……..(3.4)

The distribution in (3.4) is the Bernoulli distribution with mean

$E(y_i) = p_i = \beta_o + \beta_i x_i$ ……..…….(3.5)

And variance

$Var(y_i) = E[y_i - E(y_i)]^2 = p_i(1 - p_i)$…….....(3.6)

$\Rightarrow Var(y_i) = (\beta_o + \beta_i x_i)(1 - \beta_o - \beta_i x_i)$……..…….....(3.7)

This implies that the variance of each y_i depends on x_i which defeat the fundamental assumption of constant variance. The usual least square estimators $\widehat{\beta_0} \; and \; \widehat{\beta_1}$ will not be optimal. To obtain the optimal estimator for β_o and β_1, the generalised least square estimators for $\hat{\beta}_0$ and $\hat{\beta}_1$ are used

3.3.1. Generalized Least Squares

Consider models in which the y variables are correlated or have differing variances so that$cov(y) \neq \sigma^2 I$. In simple linear regression, larger values of x_i may lead to larger

values of $var(y_{i)}$. In either simple or multiple regression, if $y_1, y_2 \ldots\ldots\ldots\ldots y_n$ occur at sequential points in time, they are typically correlated. For cases such as these, in which the assumption that $cov(y) = \sigma^2 I$ is no longer appropriate.

The appropriate model is $y = X\beta + \varepsilon, E(y) = X\beta, cov(y) = \Sigma = \sigma^2 V$ …..…...(3.8)

X is a full rank matrix, V is a known positive definite matrix

The following theorem gives estimators of β *and* σ^2 for the model in (3.8)

Theorem 3.1

Let $y = X\beta + \varepsilon, E(y) = X\beta$ *and* $cov(y) = \Sigma = \sigma^2 V$ where X is a full-rank matrix and V is a known positive definite matrix. For this model, we obtain the following results:

The best linear unbiased estimator for β is

$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y$……..…(3.9)

Proof

Since V is a positive definite, there exist an $n \times n$ non singular matrix P such that $V = PP^{-1}$.

Multiplying $y = X\beta + \varepsilon$ by P^{-1}, we obtain $P^{-1}y = P^{-1}X\beta + P^{-1}\varepsilon$ for which $E(P^{-1}\varepsilon) = P^{-1}E(\varepsilon) = P^{-1}(0) = 0$ and $cov(P^{-1}\varepsilon) = P^{-1}cov(\varepsilon)(P^{-1})'$

$\Rightarrow cov(P^{-1}\varepsilon) = P^{-1}\sigma^2 V(P^{-1})' = P^{-1}\sigma^2 PP^{-1}(P^{-1})' = \sigma^2 P^{-1}PP'(P^{-1})' = \sigma^2 I$

This implies that

$$\hat{\beta} = [X'(P^{-1})'X'P^{-1}X]^{-1}X'(P^{-1})'y$$

$$\hat{\beta} = [X'(P')^{-1}P^{-1}X]^{-1}X'(P')^{-1})P^{-1}y$$

$$[X'(PP')^{-1}X]^{-1}X'(PP')^{-1})y$$

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y$$

$\hat{\beta}$ is the generalised least square estimator for β

Since $E(y_i) = p_i$ is a probability, $0 \leq p_i \leq 1$.

Fitting (3.8) by the generalised least square

$\widehat{p_i} = \widehat{\beta_0} + \widehat{\beta_1}x_i$……..……(3.10)

From (3.10) it means $\widehat{p_i}$ may be less than 0 or greater than 1 for some value x_i

A model for $E(y_i)$ that is bounded between 0 and 1, and approach 0 and 1 asymptotically instead of linearly is suitable. The best and most populous model is the logistic model

$$p_i = E(y_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} = \frac{1}{1+e^{-\beta_0-\beta_1 x_i}} \text{……....(3.11)}$$

In general, the model is expressed as

The probability that the dependent variable $y = 1$ given the independent variables x_i is

$$p(y = 1|x_1, x_2, x_3, \ldots, x_n) = \frac{1}{1+e^{-(\alpha+\sum \beta_1 x_i)}} \text{ ……..……(3.12)}$$

3.3.2. Odds

The odds is the ratio the probability of success and that of failure. If the value is greater than one it means there is a higher probability of success compared to that of failure. A value less than one indicate a higher probability of failure than that of success.

$$The\ Odds\ of\ p(x) = \frac{p(x)}{1-p(x)} \text{.....(3.13)}$$

$$The\ odds\ p(x) = \frac{\frac{1}{1+e^{-(\alpha+\sum \beta_i x_i)}}}{1-\frac{1}{1+e^{-(\alpha+\sum \beta_i x_i)}}} \text{.....(3.14)}$$

$$The\ oddsp(x) = \frac{1}{1+e^{-(\alpha+\sum \beta_i x_i)}} \times \frac{1+e^{-(\alpha+\sum \beta_i x_i)}}{e^{-(\alpha+\sum \beta_i x_i)}} \text{.....(3.15)}$$

Therefore $oddsp(x) = \frac{1}{e^{-(\alpha+\sum \beta_i x_i)}}$ .....(3.16)

$$odds\ P(x) = e^{(\alpha+\sum \beta_i x_i)} \text{.....(3.17)}$$

3.3.4 The Model in Logit Form

This is given by taking the natural logarithm of the quantity $P(x)$ divided by 0ne minus $P(x)$

$$\text{logit}\, p(x) = \ln\left[\frac{p(x)}{1-p(x)}\right] \text{.....(3.18)}$$

Where $p(x) = \frac{1}{1+e^{-(\alpha+\sum \beta_i x_i)}}$

From the above

$$\text{logit}\, p(x) = \ln\left[\frac{p(x)}{1-p(x)}\right] = \ln e^{(\alpha+\sum \beta_i x_i)} \text{.....(3.19)}$$

This will give $\text{logit}\, p(x) = \alpha + \sum \beta_i x_i$……...(3.20)

This implies that logit $p(x) = \ln odds\ P(x)$.....(3.21)

A plot of the logistic distribution for $0 \leq p(x) \leq 1$, indicates that values of $p(x)$ in the range of (0,1) is transformed into the value of the $logit\ p(x)$ in $(-\infty, \infty)$

3.4. Estimating *α and β*

The parameters $\alpha\ and\ \beta$ in (3.3.2) are estimated using the maximum likelihood function. For a random of y_i where i= 1,……,n from the Bernoulli distribution with $p_i(y_i = 1) = p_{i\ and}\ p_i(y_i = 0) = 1 - p_i$ the likelihood function is given by

$$L(\alpha, \beta) = f(y_1, \ldots \ldots . y_n, \alpha, \beta) = \prod f(y_i, \alpha, \beta) \text{……..(3.22)}$$

$$L(\alpha, \beta) = \prod {p_i}^{y_i}(1 - p_i)^{1-y_i} \text{……...(3.23)}$$

Taking the log of both sides

$$logL(\alpha, \beta) = \sum \log(p_i)^{y_i}(1 - p_i)^{1-y_i})$$

$$logL(\alpha, \beta) = \sum \log(p_i)^{y_i} + \sum log(1 - p_i)^{1-y_i}$$

$$logL(\alpha, \beta) = \sum y_i \log p_i + \sum (1 - y_i)\log\,(1 - p_i)$$

$$logL(\alpha,\beta) = \sum y_i \log p_i + \sum \log(1 - p_i) - \sum y_i \log(1 - p_i)$$

$$logL(\alpha,\beta) = \sum y_i \log p_i - \log(1 - p_i) + \sum \log(1 - p_i)$$

$$logL(\alpha,\beta) = \sum y_i \log\left(\frac{p_i}{1 - p_i}\right) + \sum \log(1 - p_i)$$

But from (3.20) $log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 x_i$ and from (3.12)

$p_i = \frac{1}{1+e^{-(\alpha+\beta_1 x_i)}}$ this implies that $1 - p_i = \frac{1}{1+e^{(\alpha+\beta_1 x_i)}}$

And

$$logL(\alpha,\beta) = \sum y_i(\alpha + \beta_1 x_i) + \sum \ln\left(\frac{1}{1 + e^{\alpha+\beta_1 x_i}}\right)$$

$$logL(\alpha,\beta) = \sum y_i(\alpha + \beta_1 x_i) + \sum \ln 1 - \sum ln\left(1 + e^{\alpha+\beta_1 x_i}\right)$$

$$lnL(\alpha,\beta) = \sum y_i(\alpha + \beta_1 x_i) + \sum \ln(1 + e^{\alpha+\beta_1 x_i}) \text{.........}(3.24)$$

Differentiating (3.24) with respect to $\alpha\ and \beta$ and setting the result to zero gives

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \frac{1}{1+e^{\hat{\alpha}+\hat{\beta}_i x_i}} \text{...}(3.25)$$

$$\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} \frac{x_i}{1+e^{\hat{\alpha}+\hat{\beta}_i}} \text{..}(3.26)$$

Equations (3.25) and (3.26) can be solved iteratively for $\hat{\alpha}\ and\ \hat{\beta}$

3.5. Test of correlation between the selected predicted and response variables

Bivariate analysis will be employ to determine the association between the response and the predicted variables, using 5% level of significant. The predicted variables consist of mothers background characteristics which include; age, marital status, region of residence, educational level, employment status, partner's educational level, place of residence and wealth index. The response variable is malaria prevalence. The response variable will be cross tabulated against each of the predicted variables to examine the correlation between them.

The correlation between the response and the predicted variables will be determined by computing the p-value using SPSS. Determining the p-value helps in decision making regarding the relationship between the response and the predicted variables, it also gives us additional insight into the strength of the decision. Since this study will use 5% level of significant, computed p-value will be compare to the value of the significant level (0.05), if the computed p-value is greater than 0.05, then we conclude that there no association between the response and the predicted variables. Again if the p-value is less than 0.05 then we can conclude that there is a relationship between the response and the predicted variables.

3.6. Test of predictive ability of the model

Further analysis will be conducted in this study. Since bivariate analysis cannot completely indicate the strength and direction of association between the response and the predicted variables, logistic regression will be use to determine the variables that add significantly to the improvement in the prediction of the logit. Logistic regression is a very flexible technique because it makes no assumptions about the nature of the relationship between the independent and the dependent variables. The predictor variables do not have to be normally distributed. Although the power of analysis is increased, if the independent variables are normally distributed and do have a linear relationship with the dependent variable.

Once the logistic regression model is fit, the model is then access to see the significance or contribution of each of the variable in the model. The log-likelihood value is also display in the model as well. The value of the log-likelihood helps in determining the of accuracy at which the model predicts the outcome. The higher the log-likelihood value the better fit. The log-likelihood statistics can be used to test hypothesis about the parameters in the model using likelihood ratio test. This test will indicate whether the removal of the insignificant variables from the model can affect the model. The overall value of the chi-square will also give further information about the nature of the model. The lower the chi-square value, the better the model. This information provided by the logistic regression model makes it more appropriate to use, compare to other models.

3.6.1 Deviance

In regression models for binary dependent variables, the comparison of the predicted and observed models is dependent on the log-likelihood function.

The current model is the fitted model which we want to compare with other models.

Deviance is a measure of deviation of the model from realized values. The deviance measure is defined as:

$$y = -2\ln\frac{(likelihood\ of\ the\ curent\ model)}{(likelihood\ of\ the\ structured\ model)} \quad \text{.....3.7.1}$$

When models are compared, we can use deviance as a measure to determine which one to choose.

The model with lower deviance will be chosen.

3.6.2 Pearson Chi-Square Goodness of Fit Statistic

It is a simple non-parametric goodness of fit test which measures how well an assumed model predicts the observed data. The test statistic is:

$$x^2 = \sum_{i=1}^{n}\frac{(Observed\ frequency - fitted\ frequency)^2}{(fitted\ frequency)} \quad \text{.....(3.28)}$$

x^2 is assumed to be chi-square with n − p degrees of freedom

3.6.3 G Likelihood Ratio Chi-Square Statistic

G statistic is a goodness of fit test that depends on log-likelihood function. The purpose of this test is to compare the models with and without independent variables. The test statistic is:

$$G = 2\ln\left(\frac{L_0}{L_1}\right) = 2(\ln L_0 - \ln L_1) \quad \text{.....3.29}$$

Where

L_0 is the likelihood function value of the model without any independent variables and

L_1 is the likelihood function value of the model with independent variables.

G is assumed to be distributed as chi-square with p-1 degrees of freedom.

3.6.4 Pseudo R^2

As in linear regression, pseudo R^2 measures the explained percentage of dependent variables. It also can be called as the determination coefficient. The statistic is:

$$R^2 = \frac{G}{G+n} \quad \text{..... 3.30}$$

Where

G is the value estimated in equation 3.39

Pseudo R^2 ranges between 0 and 1.

When comparing the models, the model with higher pseudo R^2 will be preferred as it is the determination coefficient.

3.6.5. Wald Statistic

To assess the significance of all coefficients we can use Wald statistic as a significance test. It is also known as pseudo t statistic.

The statistic is:

$$W = \frac{\hat{B}_i}{Se(\widehat{B_i})}, i = 1,..p+1 \quad \text{..3.31}$$

Where $\hat{B}_i$ is the maximum likelihood estimate of ith regression coefficient.

$Se(\widehat{B_i})$ is the standard error of ith regression coefficient identified as

$Se(\widehat{B_i}) = \sqrt{cov_{ii}}$

The result of Wald statistic is assumed to be normally distributed. It is asymptotic since the normal distribution provides a valid approximation for large n.

CHAPTER FOUR

DATA ANALYSIS

4.1. Introduction

This chapter presents the data for the empirical study and analysis. After describing the dataset we shall explain the variables use in the model giving descriptive statistics about them and explain how those variables affect the model.

4.2 Data Description and Summary Statistics

This study uses a unique dataset from an international Bank in Ghana. For the purpose of this study the bank will be called Bank A. Bank A specialized in providing Cooperate and Institutional small- and medium-sized loans to retail customers.

The collected data includes several variables such as Age, Sex, marital status, and number of dependence. We also have information on the number of months the borrower has been in his or her current employment and the borrower's residential status.

The data consists of 9939 applications granted loan between January, 2008 and December, 2010. Out of these, 14% defaulted and 86% performed well.

In the empirical analysis we have excluded observation of customers who applied for a loan but were rejected. The true creditworthiness status of the rejected applicants is unknown and their characteristics might differ from those who were granted the loan. The exclusion might cause a potential selection bias but is common in the literature and according to Banasik et al. (2003) has only a minimal effect on results.

The information for the variables is given by the customer at the time of filling loan application. Along with the terms and conditions, the customer is obligated to provide the pay slip to ascertain the applicant's financial status.

Table 4.1 below gives the definition of variables used in the study. Column one with heading "Variable" gives the variable names as they appear in the database. Column two gives the definition Bank A gives to each variable.

Table 4.1. Definition of Variables used in the Study

Variable	Definition
Cus_Code	Customer Code
Dob	Date of Birth
Age	Age at Disbursement
Sex	0 if applicant is Male and 1 if applicant is Female
NetIncome	Applicants net monthly income
Residence	Type of residence the applicant is staying. This is categorise into four 1. If applicant is living in their own house 2. If applicant is living in accommodation provided them by their employers 3. If applicant is living with their parent 4. If applicant is living in rented apartment
Marital Status	1if the applicant is married and 0 for single applicants
Dependents	Number of dependents applicant is having
CurrEmpAge	Number of months the applicant has been in current employment
Disbursedt	Date Loan was disbursed to the applicant
IntRate	Interest rate being charged the customer.
DelinStatus	This shows applicants whom payment is not expected. It helps to identify defaulted customers
Loan Amount	Amount granted the customer
Overdue Days	Number of days the applicant has not made payment that is due. Applicant who have made payment for 90 days and above are said to have defaulted
Tenor	Number of months the loan is running
Default Status	1 if applicant is classified as defaulted and 0 otherwise. Bank A defined default as a situation where has overdue payment for 90days or more. A customer is also said to have defaulted if repayment is not expected.

Table 4.2 shows extract of the data used in the analysis. The various columns are as defined in table 4.1

Table 4.2. Data used in the analysis

Cus_Code	Age	Sex	Net Income	Residence	Marital Status	Dependents	Curr Emp Age	IntRate	Loan Amount	Tenor	Default Status
GHA10002	54	0	506.68	1	M	3	321	24.41	5,500.00	35	0
GHA10004	42	0	473.78	2	M	1	179	25.66	1,600.00	34	1
GHA10047	41	0	1,153.08	2	M	5	121	21.08	7,300.00	54	0
GHA10070	39	0	309.38	4	S	0	123	24.92	3,500.00	48	0
GHA10087	56	0	3,000.00	1	M	0	60	25.16	18,000.00	50	0
GHA10118	46	1	509.03	2	S	6	67	22.44	6,000.00	45	0
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
GHA12967	31	0	587.63	4	S	0	60	20.43	9,800.00	61	0
GHA10227	34	0	534.16	4	M	1	56	23.56	7,000.00	58	0
GHA10233	42	0	613.75	4	M	5	171	24.65	3,000.00	46	1
GHA10240	50	0	428.31	1	M	3	322	23.55	6,500.00	58	0
GHA10251	51	0	744.72	1	M	6	282	26.17	2,500.00	36	0

4.3. Descriptive Statistics of Variables used in the model.

The data consist of ten variables of including both social demographic characteristics financial of the customers. The variables used include Age, Marital Status, Sex, number of months in current employment and Residential Status.

Table 4.3 below gives the descriptive statistic of all applicants (both defaulted and non-defaulted). Column one gives the variables which are defined in table 4.1., column two shows the number of data points used in the analysis for each of the variables. Column three gives the number of data missing in the analysis. Column four to eight give the mean, variance, minimum and maximum values respectively for each of the variables considered.

Table 4.3: Descriptive Statistics for all Applicants

Variables	Valid Cases	Missing	Mean	Std. Deviation	Variance	Minimum	Maximum
Age	9939	0	40.92	9.068	82.225	21	60
Sex	9939	0	0.13	0.338	0.114	0	1
Net Income	9939	0	730.4043	678.0422	459741.2	97.92	14730.32
Residence	9939	0	2.94	1.219	1.486	1	4
Marital Status	9939	0	0.71	0.454	0.206	0	1
Dependents	9939	0	1.62	2.111	4.457	0	66
CurrEmpAge	9939	0	126	108.447	11760.81	0	491
IntRate	9939	0	24.6971	1.87276	3.507	17.74	35.35
Loan Amount	9939	0	6728.523	5623.112	31620000	800	60000
Tenor	9939	0	42.71	8.846	78.252	6	92

From the table 4.3 it is observed that average age of a customer who were granted loan is 41 years. The average income of applicant is GHC 730. The average loan amount granted a customer is GHC 6729. During the period of consideration minimum amount of loan granted a customer was GHS 800 while the maximum was GHC 60000. The average interest charged was 24.7% .Average duration for a loan to mature was 42 months while the maximum number of months for loan duration is 92 months

Table 4.4a gives the descriptive statistics for customers who performed well (non defaulted customers). The columns in table 4.4a are as explained for table 4.3

Table 4.4a: Descriptive Statistics for non-defaulted applicants

	Valid Cases	Missing	Mean	Std. Deviation	Minimum	Maximum
Age	8597	0	41.26	8.969	21	60
Sex	8597	0	0.13	0.34	0	1
Net Income	8597	0	757.8517	686.35678	100.27	9747.34
Residence	8597	0	2.91	1.228	1	4
Marital Status	8597	0	0.72	0.447	0	1
Dependents	8597	0	1.65	2.128	0	66
CurrEmpAge	8597	0	130.3	107.911	0	491
IntRate	8597	0	24.4366	1.40303	20.05	28.78
Loan Amount	8597	0	6948.258	5784.48744	800	60000
Tenor	8597	0	42.71	8.443	12	73

The average of those who did not defaulted is 41 years with average net income of GHC 757.85. Applicants number of months in their current employment on the average is 130

months for those who did not default. With regards to tenor (Loan duration) the average is 42 months.

Table 4.4b below gives the descriptive statistics for customers who performed badly (defaulted customers). The columns in the table are as explained for table 4.3

Table 4.4b: Descriptive Statistics for defaulted applicants

Variable	Valid cases	Missing	Mean	Std. Deviation	Minimum	Maximum
Age	1342	0	38.78	9.41	21.00	59.00
Sex	1342	0	0.12	0.33	0.00	1.00
Net Income	1342	0	554.57	592.96	97.92	14730.32
Residence	1342	0	3.15	1.14	1.00	4.00
Marital Status	1342	0	0.61	0.49	0.00	1.00
Dependents	1342	0	1.40	1.99	0.00	25.00
CurrEmpAge	1342	0	98.48	107.87	2.00	487.00
IntRate	1342	0	26.37	3.19	17.74	35.35
Loan Amount	1342	0	5320.88	4189.27	800.00	35000.00
Tenor	1342	0	42.72	11.09	6.00	92.00

For the applicants who defaulted in loan payment, the average is 38. This is less than the average age of those who performed well in their loan repayment. This could mean younger applicants are more likely to default in their loan repayment. The average loan amount granted to defaulted applicant is GHC5320 as against GHC 6948 for non-defaulted applicants.

Table 4.5 below gives the default rate among the categorical variables used. The variables are Sex, Marital Status, Residence, Loan Amount and income level. Column two gives the how the variables are coded for analysis. Columns three to six give the number not defaulted, number not defaulted, the total for each of the category and the default rate respectively

Table 4.5: Default rate among, Sex, Marital Status Residence, Loan Amount and Income level.

Varibles	Variables	Not Defaulted (0)	Defaulted (1)	Total	% Defaulted
Sex	Male (0)	7450	1179	8629	14%
	Female(1)	1147	163	1310	12%
Marital Status	Married (1)	6224	822	7046	12%
	Single (0)	2373	520	2893	18%
Residence	Owner (1)	1778	205	1983	10%
	Employment (2)	1581	165	1746	9%
	Parent (3)	888	193	1081	18%
	Rent (4)	4350	779	5129	15%
Loan Amount	Less than 2000 (1)	1524	338	1862	18%
	>2000 and < 5000 (2)	2849	510	3359	15%
	>50000 and < 10000 (3)	2686	363	3049	12%
	Greater than 10000 (4)	1538	131	1669	8%
Income Level	Less than or equal to 5000 (1)	3901	835	4736	18%
	> 500 and < 1000 (2)	2811	372	3183	12%
	> =1000 and <1500 (3)	1002	83	1085	8%
	> =1500 and < 2000 (3)	461	35	496	7%
	>=2000 (4)	422	17	439	4%

From table 4.5. it can be seen that default rate high in male applicants as compared to female applicants. Default rate in male applicants is 14% while that of female applicants 12%. On marital status, default rate for married applicant is 12% against 18% in single applicant showing high default rate among the single applicants

Among the residence category, default rate is high for applicants who stay with their parents with a rate of 18%, followed by those living in rented apartment. Those who live in accommodation provided by their employee have the least default rate of 9%.

Default rate decreases as loan amount granted increases. Applicant who were granted loan amount of less than GHC 2000 have higher default rate of 18% while those who were granted amount greater than GHC 10,000 had lower default rate of 8%.

The table shows that default rate decreases as income level increases. From the table those with income level of GHC 500 and below have higher default rate of 18% as against 4% for those with income level above GHC 2000.

4.4. Computation procedure

In analysing the data, SPSS version 16 was used. The data was run on Intel(R) Celeron(R) CPU, 32 BG operating system, 1GB RAM, 2.13GHZ speed, with Window vista laptop computer. The data run successfully on the windows vista.

4.5 Results.

In table 4.6 below, Column two with the heading “B” gives the coefficient of variables in the model. Column three with the heading “S.E” gives the standard error for the coefficient values. The column four with heading “wild” gives the wild test values of the coefficient values. Df is the degree of freedom for the wild test values. The column “sig”, show how significant the variables are to the model. A value less than 0.05 shows the variable is highly significant. Column “Exp(B)” gives the odds of each variable. While column 95% CI for Exp(B) gives the upper and lower confidence interval for the odds.

Table 4.6. Default Probability Model

		B	S.E.	Wald	Df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Age	.005	.005	.849	1	.357	1.005	.994	1.016
	Sex(1)	.019	.098	.038	1	.845	1.019	.841	1.235
	Residence			7.066	3	.070			
	Residence(1)	-.039	.119	.105	1	.746	.962	.761	1.216
	Residence(2)	.181	.130	1.927	1	.165	1.198	.928	1.547
	Residence(3)	.181	.094	3.699	1	.054	1.198	.997	1.441
	MaritalStatus(1)	-.211	.081	6.770	1	.009	.810	.691	.949
	Dependents	.008	.018	.198	1	.657	1.008	.974	1.043
	CurrEmpAge	-.002	.000	18.747	1	.000	.998	.997	.999
	IntRate	.550	.020	779.609	1	.000	1.733	1.668	1.801
	Tenor	.038	.005	61.148	1	.000	1.038	1.029	1.048
	IncomeLevel			40.587	4	.000			
	IncomeLevel(1)	-.377	.098	14.622	1	.000	.686	.566	.832
	IncomeLevel(2)	-.759	.169	20.216	1	.000	.468	.336	.652
	IncomeLevel(3)	-1.061	.238	19.803	1	.000	.346	.217	.552
	IncomeLevel(4)	-1.586	.304	27.293	1	.000	.205	.113	.371
	LoanRange			32.751	3	.000			
	LoanRange(1)	-.543	.096	31.789	1	.000	.581	.481	.702
	LoanRange(2)	-.492	.133	13.629	1	.000	.612	.471	.794
	LoanRange(3)	-.405	.192	4.443	1	.035	.667	.457	.972
	Constant	-16.607	.612	736.925	1	.000	.000		

.

The table show that Marital statues Current Employment Age, Interest Rate, tenor, income level and Loan amount shows highly significant to the model.

Table 4.7 below defines how significant variables are defined in the model. Column two define the significant variable in column one and column three gives the variable as in the model. Column four gives coefficient values in the model.

Table 4.7. Variables Selected for the model

Variable	Definition	Model Variable	Co-efficient
MaritalStatus(1)	1 if applicant is married and 0 otherwise	X_1	-0.211
CurrEmpAge	Number of month the applicant has been in his current employment	X_2	-0.002
IntRate	Interest Rate being charged the customer	X_3	0.55
Tenor	Loan Duration in months	X_4	0.038
IncomeLevel(1)	1 if monthly net income is between 501 and 1000 and 0 otherwise	X_5	-0.377
IncomeLevel(2)	1 if monthly net income is between 1001 and 1500 and 0 otherwise	X_6	-0.759
IncomeLevel(3)	1 if monthly net income is between 1501 and 2000 and 0 otherwise	X_7	-1.061
IncomeLevel(4)	1 if monthly net income is greater than 2000 and 0 otherwise	X_8	-1.586
LoanRange(1)	1 if amount granted the customer is between 2001 and 5000 and 0 otherwise	X_9	-0.543
LoanRange(2)	1 if amount granted the customer is between 5001 and 10000 and 0 otherwise	X_{10}	-0.492
LoanRange(3)	1 if amount granted the customer is greater than 10000	X_{11}	-0.405
Constant	The Regression Constant	α	-16.607

From the table above, the loan default model Loan default model is

$$P(Y = 1|x_i) = \frac{1}{1+e^{-(\alpha+\sum_{i=1}^{11} X_i B_i)}}$$

Where B_is are the co-efficient in column four of table 4.7 above.

The model is presented as

$$P(Y = 1|x_i)$$

$$= \frac{1}{1 + e^{-(16.607 - 0.211X_1 - 0.002X_2 + 0.55X_3 + 0.038X_4 - 0.377X_5 - 0.759X_6 - 1.061X_7 - 1.586X_8 - 0.543X_9 - 0.492X_{10} - 0.405X_{11}}}$$

Table 4.8 shows the chi square test result. Column two gives the chi square value with degree of freedom in column three and column four tells how significant the loan default model is.

Table 4.8. Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	1380.597	12	.000
	Block	1380.597	12	.000
	Model	1380.597	12	.000

Table 4.8 shows that the model is significant with a chi square value of 1380.597 and 12 degree of freedom.

Table 4.9 below shows the outcome of the test Hosmer and Lemeshow test to support the model. Column two gives the chi-square value with column three as the degree of freedom and column four gives the significant value**.**

Table 4.9. Hosmer and Lemeshow Test

Step	Chi-square	Df	Sig.
1	548.037	8	0.092

The Hosmer and Lemeshow test support the model with a significant of 0.092

Table 4.10 shows how the model was able to classify the cases of interest. In this case defaults and not default. Table 4.10a shows the classification without considering the independent variable. Table 4.10b shows how the model the cases of interest when the independent variables are considered.

Table 4.10a Block 0 Classification Table

			Predicted		
			Default Status		Percentage
	Observed		0	1	Correct
Step 0	Default Status	0	8597	0	100.0
		1	1342	0	.0
	Overall Percentage				86.5

Block 0 Classification table gives percentage accuracy of classification 86.5%. This means that without considering the independent variables, the model classified all characteristics with 86.5% accuracy.

Table 4.10b. Block 1 Classification Table

Observed			Predicted		
			Default Status		
			0	1	Percentage Correct
Step 1	Default Status	0	8588	9	99.9
		1	974	368	27.4
	Overall Percentage				90.1

Table 4.10b shows how well the model can predict the default probability. It gives an overall Percentage Accuracy in Classification (PAC) of 90.1. The sensitivity of the model also gives $\left(\frac{368}{9+368}\right) \times 100 = 97.6\%$ which means the model classified 97.6% correctly the characteristics of interest, in this case those who defaulted. The percentage specificity of the model is $\left(\frac{8588}{974+8588}\right) \times 100 = 89.8\%$. This also means that the model classified 89.8% accurately those who did not default.

4.6. Predicting the Probability of default of an applicant:

Consider a customer who applied for a loan with profile as shown in table 4.11 below

Table4.11. Customer Profile

Sex	Male
Net Income	1326
Residential Statues	Rent
Marital Status	Single
Age	32Years
Number of months in Current employment	31
Loan Amount	10000
Interest Rate	25.5
Number of dependents	0
Tenor	36

Current interest rate of the bank is 25.5%, from the above data, the probability of default

This information can be presented as follows for the purpose of the model

Table 4.12 below shows how the customer profile is converted to the variable of interest to be fitted in the default prediction model.

Table 4.12. Model Illustration.

Variable	Definition	Model Variable	Value of X in the model
MaritalStatus(1)	1 if applicant is married and 0 otherwise	X_1	0
CurrEmpAge	Number of month the applicant has been in his current employment	X_2	31
IntRate	Interest Rate being charged the customer	X_3	25.5
Tenor	Loan Duration in months	X_4	36
IncomeLevel(1)	1 if monthly net income is between 501 and 1000 and 0 otherwise	X_5	0
IncomeLevel(2)	1 if monthly net income is between 1001 and 1500 and 0 otherwise	X_6	1
IncomeLevel(3)	1 if monthly net income is between 1501 and 2000 and 0 otherwise	X_7	0
IncomeLevel(4)	1 if monthly net income is greater than 2000 and 0 otherwise	X_8	0
LoanRange(1)	1 if amount granted the customer is between 2001 and 5000 and 0 otherwise	X_9	0
LoanRange(2)	1 if amount granted the customer is between 5001 and 10000 and 0 otherwise	X_{10}	1
LoanRange(3)	1 if amount granted the customer is greater than 10000	X_{11}	0
Constant	The Regression Constant	α	-16.607

From the above table the probability of default for the customer is calculated as follow

$$P(Y=1) = \frac{1}{1+e^{-(16.607-0.211(0)-0.002(31)+0.55(25.5)+0.038(36)-0.377(0)-0.759(1)-1.061(0)-1.586(0)-0.543(0)-0.492(1)-0.405(0))}}$$

$$P(Y=1) = \frac{1}{1+e^{(16.607-0.062+14.025+1.368+0-0.759+0+0+0-0.492+0)}}$$

$$\Rightarrow P(Y=1) = 0.074 \quad \text{.....4.1}$$

This implies the customer has 7.4% probability of default.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

Research objective one is to model loan default as a logistic regression problem. The model is represented as

$$P(Y = 1|x_i) = \frac{1}{1 + e^{-(16.607 - 0.211X_1 - 0.002X_2 + 0.55X_3 + 0.038X_4 - 0.377X_5 - 0.759X_6 - 1.061X_7 - 1.586X_8 - 0.543X_9 - 0.492X_{10} - 0.405X_{11}}}$$

Where $x_1, x_2, \ldots\ldots\ldots\ldots\ldots x_{11}$ are defined as in table 4.8

Research objective two is to predict the probability of default of a customer. for the customer whose profile is presented in table 4.12 the probability of default for customer is predicted as in equation 4.1 and the result is 7.4%.

Nine variables (characteristics) were found to be significant in predicting the default probability, these are Marital Status, Number of months the applicant has been in current employment, interest rate, tenure of loan, income level and loan amount. Variables like Age, Sex, dependents and residence did not show any significance in the model.

From the research, single applicants are 1.24 times more likely to default then those who are married.

In agreement with other literature, lower income earners are more likely to default

compared to higher income earners.

Those who have been in their current employment for longer period are more likely to repay their loan. A unit increase in the number of months in current employment the probability of default reduces by 0.998 times.

5.2 Recommendations

Based on the study, the following recommendations are made:

Banks and financial institutions should use the Internal Rate based in evaluating the credit risk of clients.

Banks should use the Probability of Default (PD) model in access the credit worthiness of clients since it is more objective.

It is recommended that for further research, the borrower's financial commitment to other financial institutions must be taken into consideration when developing PD.

REFRENCES

1. Agarwal, S., Chomsisengphet, S. and Liu, C. (2009). "Consumer Bankruptcy and Default: The Role of Individual Social Capital" Working Paper. Available at SSRN: http://ssrn.com/abstract=1408757.

2. Amarnath, K. N. Statistical Methods in Consumer Credit Scoring.

3. (http://www.hearne.com.au/attachments/Statistical%20methods%20in%20credit%20scoring.pdf)

4. Allen, L., DeLong, G. and Saunders, A. (2004). Issues in the Credit Risk Modelling of Retail Markets". Journal of Banking and Finance, Vol. 27, Issue 4, p. 221-342

5. Altman, E. I. (1968). "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy". *Journal of Finance*, Vol. 23, Issue 4, p. 589-609.

6. Altman, E. I., Eom, Y. H. and Kim, D. W. (2007). "Failure Prediction: Evidence from Korea". *Journal of International Financial Management and Accounting*, Vol. 6, Issue 3, p. 230-249.

7. Altman, E. I., and Saunders, A. (1997). “Credit Risk Measurement: Developments Over the Last 20 Years”. *Journal of Banking and Finance*, Vol. 21, Issue 11-12, p. 1721-1742.

8. Arminger, G., Enache, D. and 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.

9. Autio, M., Wilska, T-A., Kaartinen, R. and Lähteenmaa, J. (2009). “The Use of Small Instant Loans Among Young Adults – a Gateway to a Consumer Insolvency”. *International Journal of Consumer Studies*, Vol. 33, Issue 4, 407-415.

10. Banasik, J., Crook, J. and Thomas, L. (2003). “Sample selection bias in credit scoring models”. *Journal of the Operational Research Society*, Vol. 54, Issue 8, p. 822-832.

11. Bofondi, M. and Lotti, F. (2006). Innovation in the Retail Banking Industry: the Diffusion of Credit Scoring. *Review of Industrial Organization.* Vol. 28, Issue 1, p. 343-358.

12. Boyes, W. J., Hoffman, D. L. and Low, S. A. (2002). An econometric analysis of the bank credit scoring problem. *Journal of Econometrics*, Vol. 40, Issue 1, p. 3-14.

13. Brown, S., Taylor, K. and Price S. W. (2005). Debt and distress: evaluating the psychological cost of credit. *Journal of Economic Psychology*, Vol. 26, Issue 5, p. 642-663.

14. Chen, M. C. and Huang, S. H. (2003). Credit Scoring and Rejected Instances Reassigning Through Evolutionary Computation Techniques. *Expert Systems with Applications*, Vol. 24, Issue 4, p. 433-441.

15. Claessens, S., Krahnen, J. and Lang, W. W. (2005) The Basel II Reform and Retail Credit Markets. *Journal of Financial Services Research*, Vol. 28, Issue 1-3, p. 5-13.

16. Crook, J. N., Hamilton, R. and Thomas, L. C. (1983). A Comparison of a Credit Scoring Model with a Credit Performance Model. *The Service Industries Journal*, Volume 12, Issue 4, p. 558-579.

17. Desai, V. S., Crook, J. N. and Overstreet, G. A. (1996). "A comparison of neural networks and linear scoring models in the credit union environment". *European Journal of Operational Research*, Vol. 95, Issue 1, p. 23-37.

18. Dinh, T. H. T. and Kleimeier, S. (2007). A Credit Scoring Model for Vietnam's Retail Banking Market. *International Review of Financial Analysis*, Vol. 16, Issue 5, p. 571-495.

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

20. Gross, D. and Souleles, N. (2001). "Liquidity Constraints and Interest Rates Matter for Consumer Behaviour: Evidence from Credit Card Data" . *The Quarterly Journal of Economics*, Vol. 117, Issue 1, p. 149-185.

21. 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*. Vol. 160, Issue 3, p. 523-541.

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

23. Jaffee, D. M. and Russell, T. (1976). Imperfect Information, Uncertainty, and Credit Rating. *The Quarterly Journal of Economics*, Vol. 90, Issue 4, p. 651-66.

24. Laitinen, E. and Laitinen, T. (2000). Bankruptcy prediction: Application of the Taylor's expansion in logistic regression. *International Review of Financial Analysis*, Vol. 9, p. 327-349.

25. Laitinen, T. and Kankaanpää, M. (1999). Comparative analysis of failure prediction methods: the Finnish case'. *European Accounting Review*, Vol. 8, Issue 1, p. 67-92.

26. Lawrence, E. C. and Arshadi, N. (1995). A Multinomial Logit Analysis of Problem Loan Resolution Choices in Banking. *Journal of Money, Credit and Banking*, Vol. 27.

27. Lee, T. S., Chiu, C. C., Lu, C. J. and Chen, I. F. (2002). Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications*, Vol. 23, Issue 3, p. 245-254.

28. Lieli, R. P. and White, H. (2008). The Construction of Empirical Credit Scoring Models Based on Maximization Principles. *Journal of Banking and Commerce*, Vol. 27, Issue 4, p. 615-633

29. Luo, J-h. and Lei, H. Y. (2008). "Empirical study of corporate credit default probability based on Logit model. *Journal of Banking and Finance*, Vol. 27, Issue 4, p. 615-633

30. Martin, D. (1977). Early Warning of Bank Failure: A Logit Regression approach. Journal of Banking and Finance, Vol. X, Issue X, p. 249-276.

31. Mester, L. (1997). What's the point of credit scoring? Federal Reserve Bank of Philadelphia Business Review, September/October, p. 3-16.

32. Musto, D. K. and Souleles, N. S. (2006). A Portfolio View of Consumer Credit. *Journal of Monetary Economics*, Vol. 53, Issue 1, p. 59-84.

33. Neophytou, E. and Charitou, A. (2000). Predicting Corporate Failure: Empirical Evidence for the UK. Working paper, University of Southampton, Department of Accounting and Management Science, No. 01-173.

34. Peltoniemi, J. (2004). The Value of Relationship Banking. Empirical Evidence on Small Business Financing in Finnish Credit Markets. MSc thesis, University of Oulu, Oulu.

35. Roszbach, K. (2004). Bank Lending Policy, Credit Scoring and the Survival of Loans. *Review of Economics and Statistics*, Vol. 86, Issue 4, p. 946-958.

36. Steenackers, A. and Goovaerts, M. J. (1989). A Credit Scoring Model for Personal Loans. *Insurance: Mathematics and Economics*, Vol. 8, Issue 1, p. 31-34.

37. Stiglitz, J. E. and Weiss, A. M. (1981). Credit Rationing in Markets with Imperfect Information. *American Economic Review*, Vol. 71, Issue 3, p. 393-410.

38. Straka, J. W. (2000). A Shift in the Mortgage Landscape: The 1990s Move to Automated Credit Evaluations. *Journal of Housing Research*, Vol. 11, Issue 2, p. 207-232.

39. Sullivan, T. A., Thorne, D. and Warren, E. (2001). Young, Old, and In Between: Who Files for Bankruptcy? *Norton Bankruptcy Law Adviser*, p. 1-11.

40. Thomas, L. C. (2000). A Survey of Credit and Behavioural Scoring: Forecasting Financial Risk of Lending to Consumers. *International Journal of Forecasting*, Vol. 16, Issue 2, p. 149-172.

41. Tsai, M. C., Lin, S. P., Cheng, C. C. and Lin, Y. P. (2009). The consumer loan default predicting model – An application of DEA-DA and neural network. *Expert Systems with Applications*, Vol. 36, Issue 9, p- 11682 11690.

42. Van Order, R. and Zorn, P. M. (2000). Income, Location and Default: Some Implications for Community Lending. *Real Estate Economics*, Vol. 28, Issue 3, p. 385-404.

43. Vasanthi, P. and Raja, P. (2006). Risk Management Model: an Empirical Assessment of the Risk of Default. *International Research Journal of Finance and Economics*, Vol. 1, Issue 1.

44. Warren, E. (2002). Financial Collapse and Class Status: Who Goes Bankrupt? *Osgoode Hall Law Review*, Vol. 41, Issue 1, p. 114.

45. Wilson, N., Summers, B. and Hope, R. (2000). Using Payment Behaviour Data for Credit Risk Modelling. *International Journal of the Economics of Business*, Vol. 7, Issue 3, p. 333-346.

46. Yang, Y., Nie, G. and Zhang, L. (2009). Retail Exposures Credit Scoring Models for Chinese Commercial Banks. Computer Science, Vol. 5545, Issue 1, p. 633-642.

47. Zorn, P. and Lea, M. (1989). Mortgage borrower repayment behaviour: a microeconomic analysis with Canadian adjustable rate mortgage data. *Real Estate Economics*, Vol. 17, Issue 1, p. 118-136.

48. Özdemir, Ö. and Boran, L. (2004). An Empirical Investigation on Consumer Credit Default Risk. Turkish Economic Association Working Paper 2004 / 20.

49. Federation of Finnish Financial Services (2010). Kulutusluottoselvitys Tammikuu 2010. (http://www.fkl.fi/www/page/fk_www_3994)

50. The internal rating approach. Technical report, Basel Committee on Banking Supervision, http://www.bis.org, 2001.

51. Aggarawal A., (1990) Categorical Data Analysis, Wiley, New York.

52. Albright H. T. (1994) Construction of a polynomial classifier for consumer loan applications using genetic algorithms, Working Paper, Department of Systems Engineering, University of Virginia.

53. Altman E.I., Marco G., Varetto F., (1994) Corporate distress diagnosis; Comparisons using linear discriminant analysis and neural networks (the Italian experience). J. Banking and Finance 18, 505- 529.

54. Banasik J., Crook J.N., Thomas L.C., (1996) Does scoring a subpopulation make a difference?, Int. Review of Retail, Distribution and Consumer Research 6, 180-195.

55. Banasik J., Crook J.N., Thomas L.C. (1999) Not if but when borrowers default, J. Operational Research Society 50, 1185-1190.

56. Bierman H., Hausman W.H., (1970) The credit granting decision, Management Science 16, 519-532.

57. Black F., Scholes M., (1973), The pricing of options and corporate liabilities, J. of Political Economy 81, 637-654.

58. Boyle M., Crook J.N., Hamilton R., Thomas L.C., (1992) Methods for credit scoring applied to slow payers in Credit scoring and Credit Control ed. L.C.Thomas, J.N.Crook, D.B.Edelman, Oxford University Press, Oxford, pp 75-90.

59. Breiman L., Friedman J.H., Olshen R.A., Stone C.J. (1984) Classification and regression trees, Wadsworth, Belmont, California.

60. Buckley, J, James, I (1979) Linear Regression with Censored Data, Biometrika 66, 429-436

61. Capon N., (1982) Credit scoring systems: a critical analysis, J. Marketing, 46, 82-91.

62. Carter C., Catlett J. (1987) Assessing credit card applications using machine learning, IEEE Expert 2, 71-79.

63. Chandler G.G, Ewert D.C., (1976) Discrimination on basis of sex and the Equal Credit Opportunity Act, Credit Research Centre, Purdue University, Indiana

64. Chatterjee S., Barcun S., (1970) A nonparametric approach to credit screening, J. American Statistical Assoc. 65, 150-154.

65. Cheng B., Titterington D.M., (1994), Neural Networks: A review from a Statistical Perspective, Statistical Science 9, 2-30.

66. Choi S.C. (1986), Statistical Methods of Discrimination and Classification, Pergamon Press, New York.

67. Churchill G.A., Nevin J.R., Watson R.R., (1977) The role of credit scoring in the loan decision, Credit World, March, 6-10,

68. Coffman J.Y., (1986) The proper role of tree analysis in forecasting the risk behaviour of borrowers, MDS Reports, Management Decision Systems, Atlanta, 3,4,7 and 9.

69. Corcoran A.W., (1978) The use of exponentially smoothed transition matrices to improve forecasting of cash flows from accounts receivable, Management Science 24, 732-739.

70. Crook J.N, (1999), The demand for household debt in the US: evidence from the Survey of Consumer Finance, to appear in Applied Financial Economics

71. Crook J.N., Hamilton R., Thomas L.C., (1992), The degradation of the scorecard over the business cycle, IMA J. of Mathematics applied in Business and Industry 4, 111-123.

72. Cyert R.M., Davidson H.J., Thompson G.L., (1962) Estimation of allowance for doubtful accounts by Markov chains, Management Science 8, 287-303

73. Davis R.H., Edelman D.B., Gammerman A.J., (1992) Machine-learning algorithms for credit-card applications, IMA J. Mathematics applied in Business and Industry 4, 43-52.

74. Desai V.S., Crook J.N., Overstreet G.A., (1996) A comparison of neural networks and linear scoring models in the credit environment, European J. Operational Res.95, 24-37.

75. Desai V.S., Convay D.G., Crook J.N., Overstreet G.A., (1997) Credit scoring models in the credit union environment using neural networks and genetic algorithms, IMA J. Mathematics applied in Business and Industry 8, 323-346.

76. Dirickx Y.M.I., Wakeman L., (1976) An extension of the Bierman-Hausman Model for credit granting, Management Science 22, 1229-1237.

77. Durand D., (1941) Risk elements in consumer instalment financing, National Bureau of Economic Research, New York.

78. Edelman D.B., (1997), Credit scoring for lending to small businesses, Proceedings of Credit Scoring and Credit Control V, Credit Research Centre, University of Edinburgh

79. Eisenbeis R.A., (1977) Pitfalls in the application of discriminant analysis in business, finance and economics, J. of Finance 32, 875-900.

80. Eisenbeis R.A., (1978) Problems in applying discriminant analysis in credit scoring models, J. Banking and Finance 2, 205-219.

81. Equal Credit Opportunity Act, (1975) U.S.C., Title 15, Sec 1691 et seq.

82. Equal Credit Opportunity Act Amendments of 1976, (1976) Report of the Committee on Banking

83. Housing and Urban Affairs, 94th Congress, Washington D.C., U.S. Government Printing Office.

84. Fishelson-Holstine H., (1998), Case studies in credit risk model development, in Credit Risk Modeling ed by E. Mays, pp 169-180, Glenlake Publishing, Chicago.

85. Fogarty T.C., Ireson N.S., (1993) Evolving Bayesian classifiers for credit control - a comparison with other machine learning methods, IMA J. Mathematics applied in Business and Industry 5, 63-76.

86. Freed N., Glover F., (1981a) A linear programming approach to the discriminant problem, Decision Sciences 12, 68-74.

87. Freed N., Glover F., (1981b) Simple but powerful goal programming formulations for the discriminant problem, European J. Operational Research 7, 44-60

88. Frydman H., Kallberg J.G., Kao D-L, (1985), Testing the adequacy of Markov chains and Mover-Stayer models as representations of credit behaviour, Operations Research 33, 1203-1214.

89. Fung R., Lucas A., Oliver R., Shikaloff N., (1997), Bayesian Networks applied to credit scoring, Proceedings of Credit Scoring and Credit Control V, Credit Research Centre, University of Edinburgh

90. Glen J.J., (1997), Integer programming Models for normalisation and variable selection in mathematical programming models for discriminant analysis, Proceedings of Credit Scoring and Credit Control V, Credit Research Centre, University of Edinburgh

91. Grablowsky B.J., Talley W.K. (1981) Probit and discriminant functions for classifying credit applicants; a comparison, J. Economics and Business, 33, 254-261.

92. Hand D.J., (1981) Discrimination and Classification, Wiley, Chichester.

93. Hand D.J., McConway K.J., Stanghellini E., (1997) Graphical models of applications for credit. IMA J. Mathematics applied in Business and Industry 8, 143-155.

94. Hand D.J., Henley W.E., (1993) Can reject inference ever work?, IMA J. Mathematics applied in Business and Industry 5, 45-55.

95. Hand D.J., Henley W.E., (1997) Statistical classification methods in consumer credit, J. Royal Stat. Soc., Series A, 160, 523-541.

96. Hand D.J., Jacka S.D., (1998) Statistics in Finance, Arnold, London.

97. Hand D.J., Oliver J.J., Lunn A.D., (1996), Discriminant analysis when the classes arise from a continuum, Pattern Recognition, 31, 641-650.

98. Hardy W.E., Adrian J.L., (1985), A linear programming alternative to discriminant analysis in credit scoring, Abribus 1, 285-292.

99. Henley W.E. (1995) Statistical aspects of credit scoring, Ph.D. thesis, Open University.

100. Henley W.E., Hand D.J., (1996) A k-NN classifier for assessing consumer credit risk, The Statistician 65, 77-95.

101. Hopper M.A., Lewis E.M., (1992), Behaviour Scoring and Adaptive Control Systems, In Credit scoring and Credit Control ed by L.C.Thomas, J.N.Crook, D.B.Edelman pp 257-276, Oxford University Press, Oxford.

102. Hsia D.C., (1978) Credit scoring and the Equal Credit Opportunity Act, The Hastings Law Journal 30, 371-448.

103. Ignizio J.P., Soltys J.R., (1996) An ontogenic neural network bankruptcy classification tool, IMA J. Mathematics applied in Business and Industry, 7, 313-326.

104. Joachimsthaler E.A., Stam A., (1990) Mathematical programming approaches for the classification problem in two-group discriminant analysis, Multivariate Behavioural Research 25, 427-454.

105. Joanes D.N., (1993), Reject inference applied to logistic regression for credit scoring, IMA J. of Mathematics applied in Business and Industry 5, 35-43

106. Johnson R.W., (1992) Legal, social and economic issues implementing scoring in the US, in Credit Scoring and Credit Control ed L.C.Thomas, J.N.Crook, D.B.Edelman, Oxford University Press, Oxford, pp 19-32.

107. Jost A., (1998), Data Mining, In Credit Risk Modeling ed by E. Mays pp 129-154, Glenlake Publishing, Chicago.

108. Kolesar P., Showers J.L., (1985) A robust credit screening model using categorical data, Management Science 31, 123-133.

109. Krzanowski W.J., (1975) Discrimination and classification using both binary and continuous variables, J. American Stat. Assoc. 70, 782-790.

110. Lachenbruch P.A., (1975) Discriminant analysis, Hafner Press, New York.

111. Lai, T L, Ying, Z.L.,A Missing information principle and M-estimators in regression analysis with censored and truncated data, Ann. Stats, 22, 1222-255 (1994)

112. Lando D., (1994), Three Essays on contingent claims pricing, Ph.D. thesis, Cornell University. Ithaca.

113. Leonard K.J., (1993) Empirical Bayes analysis of the commercial loan evaluation process, Statistics and Probability Letters 18, 289-296.

114. Leonard K.J., (1993) Detecting credit card fraud using expert systems, Computers and Industrial engineering 25, 103-106.

115. Lewis E.M., (1992) An introduction to credit scoring, Athena Press, San Rafael, California.

116. Li H.G., Hand D.J., (1997), Direct versus indirect credit scoring classification, Proceedings of Credit Scoring and Credit Control V, Credit Research Centre, University of Edinburgh.

117. Lovie A.D., Lovie P., (1986) The flat maximum effect and linear scoring models for prediction, J. Forecasting 5, 159-186.

118. Makowski P., (1985) Credit scoring branches out, The Credit world, 75, 30-3

119. Mangasarian O.L., (1965) Linear and nonlinear separation of patterns by linear programming, Operations Research 13, 444-452.

120. Mangasarian O.L., (1993) Mathematical Programming in Neural Networks, ORSA J. on Computing 5, 349-360

121. Martell T.F., Fitts R.L., (1981) A quadratic discriminant analysis of bank credit card user characteristics, J. of Economics and Business 33, 153-159

122. Mays E. (1998), Credit Risk Modeling, Glenlake Publishing, Chicago.

123. Mehta D., (1968) The formulation of credit policy models, Management Science 15, 30-50.

124. Merton R.C., (1974) On the pricing of corporate debt: the risk structure of interest rates, Journal of Finance 29, 449-470.

125. Myers J.H., Forgy E.W., (1963), The development of numerical credit evaluation systems, J. American Stats. Assoc. 58 (September), 799-806.

126. Narain B., (1992), Survival Analysis and the credit granting decision, in Credit Scoring and Credit Control ed by L.C.Thomas, J.N.Crook, D.B.Edelman, pp 109-122, Oxford University Press, Oxford.

127. Nath R., Jackson W.M., Jones T.W., (1992) A comparison of the classical and the linear programming approaches to the classification problem in discriminant analysis, J. Statistical Computation and Simulation 41, 73-93.

128. Oliver R.M.,(1993), Effects of Calibration and Discrimination on Profitability Scoring, Proceedings of Credit Scoring and Credit Control III, Credit Research Centre, University of Edinburgh

129. Orgler Y.E., (1971) Evaluation of bank consumer loans with credit scoring models, J. Bank Research, Spring, 31-37

130. Overstreet G.A., Bradley E.L., Kemp R.S., (1992) The flat maximum effect and generic linear scoring models: a test. IMA J. Mathematics applied in Business and Industry 4, 97-110.

131. Platts G. Howe I., (1997), A single European scorecard, Proceedings of Credit scoring and Credit Control V, Credit Research Centre, University of Edinburgh.

132. Quinlan J.R., (1993) C4.5: Programs for Machine Learning, Morgan Kaufman, San Mateo, California Reichert A.K., Cho C-C, Wagner G.M., (1983) An examination of the conceptual issues involved in developing credit scoring models, J. Business and Economic Statistics 1, 101-114.

133. Rosenberg E., Gleit A., (1994) Quantitative methods in credit management : a survey, Operations Research 42, 589-613.

134. Safavian S.R., Landgrebe D., (1991) A survey of decision tree classifier methodology, IEEE Trans. On Systems, Man and Cybernetics, 21, 660-674

135. Sewart P., Whittaker J.,(1998) Fitting graphical models to credit scoring data, IMA J. Mathematics in Business and Industry, 9, 241-266

136. Showers J.L., Chakrin L.M. (1981) Reducing revenue from residential telephone customers, Interfaces 11, 21-31.

137. Srinivasan V., Kim Y.H., (1987a) The Bierman-Hausman credit granting model: a note, Management Science 33, 1361-1362.

138. Srinivasan V., Kim Y.H. (1987b) Credit granting: a comparative analysis of classification procedures, J. of Finance 42, 665-683.

139. Tam K.Y., Kiang M.Y., (1992) Managerial applications of neural networks: the case of bank failure predictions, Management Science 38, 926-947.

140. Tessmer A.C., (1997) What to learn from near misses: on inductive learning approach to credit risk assessment, Decision Sciences 28, 105-120.

141. Thomas L.C., (1992) Financial risk management models, In Risk analysis, assessment and management ed. J. Ansell, F. Wharton, Wiley, Chichester.

142. Thomas L.C. (1994) Applications and solution algorithms for dynamic programming, Bulletin of the I.M.A. 30, 116-122.

143. Thomas L.C. (1998) Methodologies for classifying applicants for credit, in Statistics in Finance, ed. By D. J. Hand, S.D.Jacka, pp 83-103, Arnold, London

144. Thomas L.C., Allen D., Morkel- Kingsbury N., (1998) A hidden Markov Chain model of credit risk spreads, Working Paper, Department of Finance and Business Economics, Edith Cowan University.

145. Thomas L.C., Crook J.N., Edelman D.B., (1992) Credit Scoring and Credit Control, Oxford University Press, Oxford

146. Titterington D.M., (1992), Discriminant Analysis and related topics, in Credit Scoring and Credit Control ed L.C.Thomas, J.N.Crook, D.B.Edelman, pp 53-73, Oxford University Press, Oxford.

147. Vlachonikolis I.G., (1986) On the estimation of the expected probability of misclassification in discriminant analysis with mixed binary and continuous variables, Comp. and Maths with Applications 12A, 187-195.

148. Van Kuelen J.A.M., Spronk J., Corcoran A.W. (1981) Note on the Cyert-Davidson-Thompson Doubtful Accounts Model, Management Science 27, 108-112.

149. Wiginton J.C. (1980) A note on the comparison of logit and discriminant models of consumer credit behaviour, J. Financial and Quantitative Analysis 15, 757-770.

150. Yobas M.B., Crook J.N., Ross P. (1997) Credit scoring using neural and evolutionary techniques, Working Paper 97/2, Credit Research Centre, University of Edinburgh

151. Ziari H.A., Leatham D.J., Ellinger P.N., (1997), Development of statistical discriminant mathematical programming model via re-sampling estimation techniques, Amer. J. Agricultural Economics 79, 1352- 1362.

152. Zocco D.P., (1985) A framework for expert systems in bank loan management , J. Commercial Bank Lending 67, 47-54.