University of Nairobi

School of Mathematics

College of Biological and Physical Sciences

Application of Linear Logistic and Discriminant Analysis on Forecasting Creditworthiness of Individual Borrowers

This research project is submitted to the University of Nairobi in partial fulfilment of the requirement for the degree of Masters of Science in Social Statistics

By

Maranga Bokea Samuel

July 2013

Declaration

This dissertation is my original work and has not been presented for a degree in any other university.

Maranga Bokea Samuel

I56/71364/2011

Signature Date

…..…… **……..**

Declaration by Supervisor

This dissertation has been submitted for examination with my approval as supervisor

Dr. Mwaniki Joseph Ivivi

Signature Date

…….…… **..……..**

Dedication

I would like to dedicate this research project to my dear wife Dinnah Kwamboka, my mother Mary Maranga, my daughter Hazellindsay and my son Fonzell.

Acknowledgement

I would like to acknowledge the contribution of several personalities whose input helped in one way or another to the successful completion of the research project. First and foremost the Almighty God for this far I have come is by His Grace.

I appreciate the contribution and encouragement from my dear loving wife Dinnah Kwamboka and my two children Hazellindsay Kemunto & Fonzell Bokea who stood encouraged me and for their understanding when they knew that I would not make time for them in pursuit of completion of the Project. My mother for teaching me the benefits of education and taking me to school despite the hard times.

I am very grateful to my employer Kenya Commercial Bank, Lending Risk Unit for providing data for analysis for Masters of Science degree (Project). May I thank the teaching and non-teaching staff in the School of Mathematics, University of Nairobi for their support.

I would like to recognize the contribution of all MSc. Lecturers (Prof. M. Manene, Prof. J.A.M Otieno, Dr. C.I. Kipchirchir and Mr. Nderitu). May I single out my supervisor (Dr. Mwaniki J. Ivivi) for his guidance right from the start of my project to its completion.

Many thanks to colleagues and friends with whom I interacted with in the course of my studies. It could not have been possible without all of you.

Abstract

This research study summarizes the evaluation of credit risk using credit scoring method. Credit scoring is a technique that helps banks decides whether to grant credit to applicants who apply to them or not. The main objective of the research was to evaluate credit risk in commercial banks using credit scoring models by ranking them based on their behavioural financial and non- financial characteristics to honour their debt obligation in future. We applied both logistic regression and discriminant analysis to identify predictors of default and risk factors among cardholders followed for a period of eighteen months. A credit scoring model was developed which can be used by commercial banks to determine the creditworthiness of individual borrowers requesting for credit cards. The results showed out that females constituted 64.3% of the population and they were the most disciplined. Type I and type II errors had been calculated for all the credit scoring models used. The results shows that the proposed model - Linear Logistic Model has more accuracy rate with less misclassification cost errors as compared Discriminant Analysis. Also, several suggestions for further research were presented.

Tables of contents

Declaration i

Dedication ii

Acknowledgement iii

Abstract iv

Chapter One 1

1.1 Introduction 1

1.2 Decision process 2

1.3 Probability estimation 3

1.4 Problem Statement 3

1.5 Overall Objective 4

1.6 Specific objectives 4

1.7 Significance of the study 4

Chapter Two 5

Literature Review 5

Chapter Three 10

Methodology and Research Design 10

3.0 Introduction 10

3.1 Model Specification 10

3.1.1 Logistic Regression 10

3.1.1.1 Model Assumptions 12

3.1.1.2 Interpreting log odds and the odds ratio 12

3.1.1.3 Model fit and the likelihood function 13

3.1.2 Discriminant Analysis 14

3.1.2.1Model Assumptions: 15

3.1.2.2Test of significance 16

3.2 Study area 16

3.3 Study population 16

3.4 Study procedure 16

3.5 Data Source 17
3.6 Data Cleaning 18
3.7 Data Handling 18
3.8 Data Design 19
3.9 Data Layout 19
3.9 Description of the Sample 21
3.10 Variable selection 21
3.11 Logical Trend 23
3.12 Business logic 23
Chapter Four 25
4.1 Data Analysis Tools 25
4.2 Developing Credit Scoring Model 25
4.3 Description of the Sample 27
4.4 Non-financial Factors 27
4.5 Financial Factors 33
Chapter Five 38
5.0 Credit Scoring Models 38
5.1 Logistic Model and Model assessment 38
5.2 Discriminant Analysis and model assessment 41
5.3 Comparing Credit Scoring Models for Individuals 47
5.4 Conclusion 48
5.5 Recommendations 49
References 50
Appendix 53

Chapter One

1.1 Introduction

Peoples' ability to 'buy-now, and pay-later' have so far been driven by the growing/demands of economy. Our modern world today depends upon credit. In the past in about 2000 years ago, credit was not considered as such important but later a privilege. In today's industrialized societies it is considered as a right.

The word 'credit' comes from the old Latin word 'credo', which means, 'trust in', or 'rely on'. If you lend something to somebody, then you have to have trust in him or her to honour the obligation. Many people today view access to credit as a right, but it comes with its own obligations. Borrowers must pay the price of creating the impression of trust; repaying according to the agreed terms; and paying a risk premium for the possibility they might not repay. This gives rise to concepts like: creditworthiness—borrowers' willingness and ability to repay; and credit risk—the potential financial impact of any real or perceived change in borrowers' creditworthiness.

Scoring refers to the use of a numerical tool to rank order cases (people, companies, fruit, countries) according to some real or perceived quality (performance, desirability, saleability, risk) in order to discriminate between them, and ensure objective and consistent decisions (select, discard, export, sell).

According to *Anderson (2007)*, providing credit is a risky business though, as borrowers differ in their ability and willingness to pay. At the extreme, lenders may lose the full amount, and perhaps even get sucked in for more. In other instances, they may lose only a part, or just incur extra costs to get the money back. It is a gamble, and lenders are always looking for means of improving their odds.

Credit scoring was first used in the 1960s, to determine whether people applying for credit would repay the debt, honour the obligation, and—in general—act in a manner deemed acceptable by the treasury's gatekeeper. At that time, it was associated exclusively with 'accept/reject' decisions generated by the new-business application process (application

scoring), and many people still use the term in that limited sense. In the twenty-first century, however, the label is used more broadly to describe any use of statistical models to extend and manage credit generally. This includes the measurement of risk, response, revenue, and retention, whether for marketing, new-business processing, account management, collections and recoveries, or elsewhere (the credit risk management cycle, or CRMC).

Credit scoring is therefore a technique mainly used to assist credit-grantors in making lending decisions. Its aim is to construct a classification rule that distinguishes between "good" and "bad" credit risks according to some specified definition. The rule is developed on a sample of the past applicants, whose performance is known. As such a scoring model evaluates an applicant's creditworthiness by bundling key attributes of the applicant and aspects of the transaction into a score and determines, alone or in conjunction with an evaluation of additional information, whether an applicant is deemed creditworthy.

To develop a model, the modeler selects a sample of consumer accounts (either internally or externally) and analyzes it statistically to identify predictive variables (independent variables) that relate to creditworthiness. The model outcome (dependent variable) is the presumed effect of, or response to, a change in the independent variables.

The aim of credit scoring is to provide banks with intelligence about the borrower (or applicant) that allows them to assess risk and potential reward and this can be categorized either as part of a decision process, or probability estimation.

1.2 Decision process

Common Terms

Application scoring: takes information from the applicants' application and uses to determine the score for purposes of evaluating applicants for acceptance or denial.
Behavioral scoring: Use scores to determine how well-behaved existing borrowers are and therefore to anticipate any problems in the future.
Fraud detection: Use scores to detect unusual credit use which may be the result of fraud.
Cross-selling: Decide who to target for additional financial products.

1.3 Probability estimation

Credit scoring through modelling will evaluate the risk of default by examining the different borrower's characteristics attributing to different weight to explanatory variables on risk of default to determine the probability of default (PD) for each borrower.

1.4 Problem Statement

Increasing amount of non-performing facilities in the credit portfolio is unwelcoming to banks in achieving their objectives. These facilities are directly related to the financial performance of a bank. An increase in the NPA of a bank suggests that there is a high probability of a large number of credit defaults. This in turn affects the net-worth of the bank and also erodes the value of the bank's asset. Historical evidence suggest that most bank failures are directly associated with poor management of credit risk.

The approach applied by lenders to monitor on whom to consider as good or a bad customer has been skewed on the individual estimation of the analyst and the risk appetite to risk. This has left most of the banks with a serious challenge on the objective method on which to monitor the behaviour of their customers and be able to identify those customers who are more likely to default in the next one year. Banks in the long run end up facing a high risk of reviewing and approving additional credit to customers who are very likely to default.

Risk of default has been a challenge that card issuers and banks are facing. The purpose of this study is introduce and show an objective approach that can be applied on a regular basis in evaluating and discriminating low risk customers from high risk customers

Predictive variables to be considered include, but are not limited to, days past due, type of account, conduct of the existing account, segmentation of card, card limit, professionalism [employee status], age and industry.

1.5 Overall Objective

To build a model that ranks customers based on behavioural financial and nonfinancial characteristics to honour their debt obligation in future.

1.6 Specific objectives

a) To determine Financial and non-financial behavioral predictive factors linked/attached to loan defaulters
b) To establish the relationship between the non-financial risk components and financial predictive factors.
c) To compare the proposed credit scoring model with the pre-existing statistical credit scoring models.

1.7 Significance of the study

Banks and other financial institutions are often faced with risks that are mostly of financial nature. These institutions must balance risks as well as returns. For a bank to have a large consumer base, it must offer products that are reasonable enough. As a business, the main objective of commercial banks is to maximize profits for its stakeholders. This study will examine the current credit assessment practices on behavioural risk modelling and how they affect the level of nonperforming loans and in so doing guide management on improvements it should undertake to minimize losses arising from loan defaults.

This study presents a methodology that can serve both purposes—validating credit-scoring models used for customer decree and validating the estimation of the risk components. Application and behaviour scores may be used as input for pooling retail portfolios as well as for estimating the risk components.

The study will also benefit the researchers who may be interested in this topic and may find the results opening up new avenues for further research in a similar area. It will provide reference material to future researchers on banking and customer satisfaction. It will also indicate other areas of possible research like ways of improving customer satisfaction in the banking industry.

Chapter Two

Literature Review

Lending money is risky, but at the same time profitable. Interest and fees on loans are source of profits for the banks. Banks do not want to grant credit to those borrowers who are not able to repay the loan. Over time, some of the loans can become bad even if the banks do not want to have bad loans.

The traditional methods of deciding whether to grant or extend credit to a particular individual use human judgment of the risk of default based on experience of previous decisions. Due to increased demand for credit combined with increased creditors competition and advanced computing technology have opened the application of statistical models in credit decisions. Behavioral/performance scoring is the monitoring and predicting the repayment behavior of a consumer to whom credit has already been granted.

Thomas et al (2002) described "Credit Scoring is the set of decision models and their underlying techniques that aid lenders in the granting of consumer credit. These techniques decide who will get credit, how much credit they should get, and what operational strategies will enhance the profitability of the borrower to the lenders."

According to new Basel II Capital Accord, default is defined as 90 days delinquent this is defined by Siddiqui (2006). Kanwar (2005) defined credit risk as risk arises when the borrower either is unwilling to repay the loan or he is not able to repay the loan granted which results in economic loss to the bank.

Credit scoring has used the data on consumer behavior for the first time so it can be declared as the grandfather of data mining. Firstly, a lender should take two decisions in the credit approval process; one is whether to give loan to a fresh borrower; the technique that used to make this judgment is credit scoring and, other, whether to increase the credit limits of the existing debtors; the techniques that assist the second decision are called behavioral scoring. According to Thomas et al (2002) Lenders in developed countries analyze the creditworthiness of borrowers based on their credit histories taken from credit bureau and

also check borrower's salary and experience before loan approval see (Schreiner, 2000), for example.

According to Thomas, Edelman and Crook (2002) lending institutions started adopting the credit scoring models in evaluating personal loans, after few years for the evaluation of mortgage and small business loans in 1980, after analyzing the effectiveness and accuracy of credit scoring models in the evaluation of credit cards.

The objective of credit scoring models is to assign loan customers to either good credit or bad credit, or predict the bad creditors. Therefore, scoring problems are related to classification analysis (Anderson, 2003). Probably the earliest use of statistical scoring to distinguish between "good" and "bad" applicants was by (Durand, 1941), who analyzed data from financial services, such as commercial and industrial banks, and finance and personal finance companies. Statistical models called scorecards or classifiers, use predictor variables from application forms and other sources to yield estimates of probability of defaulting.

The categorization of good and bad credit is of fundamental importance, and is indeed the objective of a credit scoring model .The need of an appropriate classification technique is thus evident. But what determines the categorization of a new applicant? From the review of literature, characteristics such as gender, age, marital status, dependents, having a telephone, educational level, occupation, time at present address and having a credit card are widely used in building scoring models (Hand et al. 2005; Lee and Chen 2005; Sarlija et al., 2004; Banasik et al. 2003; Chen & Huang, 2003; Lee et al., 2002; Orgler 1971; Steenackers and Goovarts 1989). Time at present job, loan amount, loan duration, house owner, monthly income, bank accounts, having a car, mortgage, purpose of loan, guarantees and others have been also used in building the scoring models (Lee and Chen, 2005; Greene 1998; Sarlija et al., 2004; Orgler 1971; Steenackers and Goovarts 1989). In some cases the list of variables has been extended to include spouse personal information, such as age, salary, bank account and others (Orgler, 1971). Of course, more variables are less frequently used in building scoring models, such as television area code, weeks since the last county court judgment, worst account status, time in employments, time with bank and others (Bellotti and Crook, 2009; Banasik and Crook, 2007; Andreeva, 2006; Banasik et al. 2003).

Insights can be gained from parallel research, pertaining to small business and corporate loans, by identifying other variables, such as main activity of the business, age of business,

business location, credit amount, and different financial ratios, for example, profitability, liquidity, bank loans and leverage have been used in scoring applications (Emel et al. 2003; Bensic et al, 2005; Zekic-Susac et al. 2004; Min and Lee, 2008; Min and Jeong, 2009; Lensberg et al. 2006; Cramer, 2004; Liang 2003).

In some cases the final selection of the characteristics was based on the statistical analysis used, i.e. stepwise logistic regression, regression or neural network (Lee and Chen, 2005; Nakamura, 2005; Kay & Titterington, 1999; Lenard, et al., 1995; Steenackers and Goovarts 1989; Orgler 1971). However, to the best of our knowledge, none of the research reviewed in this study has clearly established a theoretical reason why such variables have been chosen. In addition, in most cases, authors have stated that a particular set of data was provided by a particular institution. Therefore, the selection of the variables used in building scoring models depends on the data providers and the data availability as stated by those authors. It is the view in this study that such variables are implicitly deemed influential.

Both the lenders and the borrowers could bear the costs of loan delinquencies. The creditor will not get the interest payments and also the loan given. The debtor will come in the list of defaulters so his character will be affected as well as he cannot further take loans from the same creditor and also could not invest that loan taken, (Baku & Smith, 1998).

Lieli and White (2010) analyzed that credit is granted to applicants after assessing their creditworthiness, when an applicant meeting the cut off score the he/she will be a accepted and considered as good applicant and increase their credit limits while all those applicants having credit score with total scores lower than cut off score is rejected.

Classification models for credit scoring are used to categorize new applicants as either accepted or rejected with respect to these characteristics. These need to be contextualized to the particular environment, as new variables are appropriately included (see, for example, the inclusion of corporate guarantees and loans from other banks within the Egyptian environment in the investigation by (Abdou and Pointon, 2009). The classification techniques themselves can also be categorized into conventional methods and advanced statistical techniques. The former include, for example, weight of evidence, multiple linear regression, discriminant analysis, probit analysis and logistic regression. The latter comprise various approaches and methods, such as, fuzzy algorithms, genetic algorithms, expert systems, and neural networks (Hand & Henley, 1997). On the one hand, the use of only two groups of customer credit, either “good” or “bad” is still one of the most important approaches to credit

scoring applications (Kim & Sohn, 2004; Lee et al, 2002; Banasik et al, 2001; Boyes et al, 1989; Orgler, 1971). On the other hand, the use of three groups of consumer credit may become one of the approaches for classification purposes in credit scoring models. Some have used "good" or "bad" or "refused" (Steenackers & Goovaerts, 1989), whilst others have used "good" or "poor" or "bad" (Sarlija et al, 2004). (Lim & Sohn, 2007) argue that the way existing models are used is quite worrying, especially at the time when the middle of the repayment term occurs, when it is important to be able to re-evaluate the creditability of borrowers with high default risks for the remaining term.

Although most literature presents probability of default based on application attributes of the applicants. It has been examined that after acceptance of an applicant, their future behavior possesses potential indication of their future repayment ability for granted credit. Indeed it has been cited that behavior of the customer are key indicators to default (Anderson, 2007)

According to Chijoriga (2011), Credit scoring models can be qualitative as well as quantitative in nature. Qualitative technique is judgmental and subjective; the disadvantage of qualitative method is that there is no objective base for deciding the default risk of an applicant. While, quantitative technique is a systematic method to categorize into performing or non- performing loans and it has removed the shortcomings of qualitative technique and proved to be more reliable & accurate model.

The quantitative approach has been applied by large number of studies utilizing various statistical techniques based on credit applicants‟ information that are obtained from lending institutions. The key objective of these studies is to reveal the distinctive indicators among the defaulters and non-defaulters.

According to Basel II rules, banks should have a sound internal rating system to assess the credit risk of debtors through which bank loan officers can effectively and accurately quantify risk and define credit limits accordingly (Hasan & Zazzara, 2006). Lopez and Saidenberg (2000) defined that according to Basel Capital Accord; banks must keep 8% capital against the risk-weighted assets.

Barefoot (1996) described several key benefits of credit scoring: credit scoring lowers the cost of lending as it has reduced the part of human in evaluating a loan application. Credit scoring models has increased the accuracy of predicting the actual credit risk of debtor. According to Ponicki (1996), for banks credit scoring provided a standard technique of loan

evaluation across the entire bank, efficient way of executing the transactions and also enhances the collection of loan. Credit scoring models provide benefits to customers by offering simple application process, results of credit approval in a timely manner, access to credit when they need it.

Lending institutions adopt seventy percent of credit scoring models to evaluate microcredit and 97% to assess the credit card requests.(Mester, 1997)

According to Schreiner (2002), statistical scoring cannot replace the loan officers because ultimately it is the duty of the credit analysts to make the credit decision and these scoring techniques can act as a help guide. Statistical scoring reminds the credit manager the elements of risks that they have ignored.

Chapter Three

Methodology and Research Design

3.0 Introduction

There are three main approaches for credit scoring (Thomas, 2000):

a) Judgmental,
b) Statistical and
c) Non-statistical, non-judgmental.

This paper focuses on the statistical approach, which is based on historical data and includes methodologies as discriminant analysis (DA) and logistic regression (LR)

Discriminant analysis is a computationally efficient procedure, but is hampered by the assumption of normally distributed data. As the models presented in this study include multiple dummy variables, the normality assumption is violated and therefore we opted to use both DA and LR to compare their effectiveness and which one be adopted.

3.1 Model Specification

Model-building techniques used in statistics are aimed at finding the best fitting and reasonable model to describe the relationship between an outcome (dependent or response) variable and a set of independent (predictor or explanatory) variables. These independent variables are often called covariates. The traditional method used is often linear regression model where the outcome variable is assumed to be continuous.

3.1.1 Logistic Regression

While logistic regression gives each predictor a coefficient ‘b’ which measures its independent contribution to variations in the dependent variable, the dependent variable can only take on one of the two values: 0 or 1.

Predicted values are interpreted as probabilities and are now not just two conditions with a value of either 0 or 1 but continuous data that can take any value from 0 to 1. like in the case

of linear regressions, predicted values needs to be transformed so that the outcome is not a prediction of a Y value, as in linear regression, but a probability of belonging to one of two conditions of Y, which can take on any value between 0 and 1 rather than just 0 and 1.

A log transformation – is needed to normalize the distribution. This log transformation of the p values to a log distribution enables us to create a link with the normal regression equation. The log distribution (or logistic transformation of p) is also called the logit of p or logit(p). *Logit*(*p*) is the log (to base *e*) of the *odds ratio or likelihood ratio* that the dependent variable is 1. In symbols it is defined as:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) \quad (1.0)$$

Whereas *p* can only range from 0 to 1, *logit*(*p*) scale ranges from negative infinity to positive infinity and is symmetrical around the *logit* of .5 (which is zero).

Equation (1.1) below shows the relationship between the usual regression equation (a+bx+ …), which is a straight line formula, and the logistic regression equation.

The form of the logistic regression equation is:

$$\text{logit}\,[p(x)] = \log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_{1i} + \ldots\ldots\ldots\ldots + \beta_k x_{ki} + \varepsilon_i \quad (1.1)$$

Instead of using a *least-squared deviations* criterion for the best fit, it uses a *maximum likelihood* method, which maximizes the probability of getting the observed results given the fitted regression coefficients. *p* can be calculated with the following formula (formula 1.2) which is simply another rearrangement of formula 1.1:

$$p = \frac{e^{a+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}} \quad (1.2)$$

Where:
p = the probability that a case is in a particular category,

a = the constant of the equation and,

b = the coefficient of the predictor variables.

Logistic regression – involves fitting an equation of the form to the data:

$$\text{logit}\,[p(x)] = \log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_{1i} + \ldots\ldots\ldots\ldots + \beta_k x_{ki} + \varepsilon_i,$$

$$i = 1,2,3 \ldots, N$$

The variable that the regression seeks to explain is coded p = 1 or p = 0. The independent variables that can affect the dependent variable are noted with X.

$$p = \begin{cases} 0 \text{ if average delay per installment} < 90 \text{ days} \\ 1 \text{ if average delay per installment} > 90 \text{ days} \end{cases}$$

3.1.1.1 Model Assumptions

Many distributions functions have been proposed in modeling binary outcome data for example Linear Discriminant Analysis (LDA), today logistic regression is the most preferred because of the following:-

(i) there are fewer assumption violations, especially as it does not demand normally distributed independent variables;

(ii) it works better where group sizes are very unequal;

(iii) Mathematically the resulting models are easier to interpret due to its mathematical simplicity.

3.1.1.2 Interpreting log odds and the odds ratio

The *Logits (log odds)* are the b coefficients (the slope values) of the regression equation.

The slope can be interpreted as the change in the average value of Y, from one unit of change in X.

Logistic regression calculates changes in the log odds of the dependent, not changes in the dependent value as OLS regression does. For a dichotomous variable the odds of membership of the target group are equal to the probability of membership in the target group divided by the probability of membership in the other group. Odds value can range from 0 to infinity and tell you how much more likely it is that an observation is a member of the target group rather than a member of the other group. If the probability of membership in the target group is .50, the odds are 1 to 1 (.50/.50), as in coin tossing when both outcomes are equally likely.

Odds ratio (OR), estimates the change in the odds of membership in the target group for a one unit increase in the predictor. It is calculated by using the regression coefficient of the predictor as the exponent or exp. Example, if we are predicting accountancy success by a maths competency predictor that b = 2.69. Thus the odds ratio is exp(2.69) or 14.73. Therefore the odds of passing are 14.73 times greater for a student, for example, who had a pre-test score of 5, than for a student whose pre-test score was 4.

SPSS actually calculates this value of the *ln*(odds ratio) for us and presents it as **EXP(B**) in the results printout in the **'Variables in the Equation'** table.

3.1.1.3 Model fit and the likelihood function

Just as in linear regression, we are trying to find a best fitting line of sorts but, because the values of Y can only range between 0 and 1, we cannot use the least squares approach. The Maximum Likelihood (or ML) is used instead to find the function that will maximize our ability to predict the probability of Y based on what we know about X.

Likelihood just means probability. It always means probability *under a specified hypothesis*.
In logistic regression, two hypotheses are of interest:

a) the null hypothesis, which is when all the coefficients in the regression equation take the value zero, and
b) the alternate hypothesis that the model with predictors currently under consideration is accurate and differs significantly from the null of zero, i.e. gives significantly better than the chance or random prediction level of the null hypothesis.

Log likelihood is the basis for tests of a logistic model. *The likelihood ratio test* is based on –2LL ratio. It is a test of the significance of the difference between the likelihood ratio (–2LL) for the researcher's model with predictors (called model chi square) minus the likelihood ratio for baseline model with only a constant in it.

Significance at the .05 level or lower means the researcher's model with the predictors is significantly different from the one with the constant only (all 'b' coefficients being zero). It measures the improvement in fit that the explanatory variables make compared to the null model. Chi square is used to assess significance of this ratio.
When probability fails to reach the 5% significance level, we retain the null hypothesis that knowing the independent variables (predictors) has no increased effects (i.e. make no difference) in predicting the dependent.

3.1.2 Discriminant Analysis

Discriminant Function Analysis (DA) undertakes the same task as multiple linear regression by predicting an outcome.

DA is used when:

The dependent is categorical with the predictor IV's at interval level such as age, income, attitudes, perceptions, and years of education, although dummy variables can be used as predictors as in multiple regression. Logistic regression predictor IV's can be of any level of measurement.

There are more than two DV categories, unlike logistic regression, which is limited to a dichotomous dependent variable. DA involves the determination of a linear equation like regression that will predict which group the case belongs to. The form of the equation or function is:

$$D = a + b_1x_1 + b_2x_2 \ldots + b_nx_n \quad (1.3)$$

Where D = discriminate function

b = the discriminant coefficient or weight for that variable

X = respondent's score for that variable

a = a constant

n = the number of predictor variables

We use maximum likelihood technique to assign a case to a group from a specified cut-off score.

a) If group size is equal, the cut-off is mean score.
b) If group size is not equal, the cut-off is calculated from weighted means.

The aim of the statistical analysis in DA is to combine (weight) the variable scores in some way so that a single new composite variable, the discriminant score, is produced.

This function is similar to a regression equation or function. These b's maximize the distance between the means of the criterion (dependent) variable. Good predictors tend to have large weights. Discriminant analysis – creates an equation which will minimize the possibility of misclassifying cases into their respective groups or categories.

Purpose of Discriminant analysis is:

a) To maximally separate the groups.
b) to determine the most parsimonious way to separate groups
c) to discard variables which are little related to group distinctions

3.1.2.1Model Assumptions:

- Cases should be independent.
- Predictor variables should have a multivariate normal distribution, and within-group variance-covariance matrices should be equal across groups.
- Group membership is assumed to be mutually exclusive

3.1.2.2Test of significance

- For two groups, the null hypothesis is that the means of the two groups on the discriminant function-the centroids, are equal.
- Centroids are the mean discriminant score for each group.
- Wilk's lambda is used to test for significant differences between groups.
- Wilk's lambda is between 0 and 1. It tells us the variance of dependent variable that is not explained by the discriminant function.
- Wilk's lambda is also used to test for significant differences between the groups on the individual predictor variables.
- It tells which variables contribute a significant amount of prediction to help separate the groups.

3.2 Study area

This study was carried out on credit cardholders for a local bank over a period of eighteen months.

3.3 Study population

The study population was restricted to existing customers who have applied for the grant of credit and were accepted by the bank and they have not defaulted at the end of learning period which in this case is 6 months.

3.4 Study procedure

Based on a customers' credit history, a score is calculated to predict the likelihood of a customer defaulting on a new account. Behaviour scoring models use credit and account performance data to determine whether to increase credit lines, re-price accounts etc.

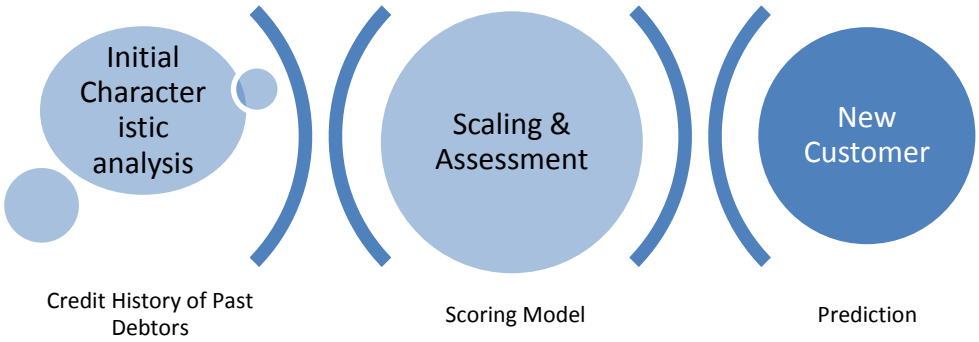
According to new Basel II Capital Accord, default is defined as 90 days delinquent this is defined by Siddiqui (2006). Credit risk is defined as risk that arises when the borrower either is unwilling to repay the loan or he is not able to repay the loan granted which results in economic loss to the bank.

The entire data that was chosen covered a period of 18 months which was categorised as:

a) Performance or observation period and is usually 6-12 months in length. Typical performance data would be average, maximum and minimum levels of balance, credit turnover, and debit turnover. Some of the characteristics are indicators of delinquent behavior; overdrawn amount, value of cash withdrawals, number of missed payments, times in over credit limit, number of cash withdrawals among others
b) The period after the observation point is the outcome period, which is usually taken as 12 months, and the customer, is classified as a good or a bad depending on their status at the end of this outcome period (Thomas et al, 2001).

Scorecard development steps

Figure 1: Credit Scoring Process



3.5 Data Source

The individual data was collected from a well reputed internal local commercial bank, Kenya Commercial Bank as a case study.

3.6 Data Cleaning

Predictive models are heavily reliant upon the data used for their development, and if the data is substandard, it affects the quality of the final result. The major sources of problems are missing data, misrepresentation, and miscapture. While some statistical techniques such as decision trees are neutral to missing values, logistic regression requires complete datasets with no missing data (i.e., complete case analysis). Ways to deal with missing values in our case the NULLS are as follows:-

a) Exclude all data with missing values—this is complete case analysis, and in most financial industry cases, will likely result in very little data to work with.
b) Include characteristics with missing values in the scorecard. The "missing" can then be treated as a separate attribute, grouped, and used in regression as an input. The scorecard can then be allowed to assign weights to this attribute. In some cases this assigned weight may be close to the "neutral" or mean value, but in cases where the weight is closer to another attribute, it may shed light on the exact nature of the missing values.

Missing values are not usually random. For example, those who are new at their work may be more likely to leave the "Years at Employment" field blank on an application form. Missing values may be part of a trend, linked to other characteristics, or indicative of bad performance (Siddiqi 2006). In addition, having assigned points for missing value in the scorecard will facilitate the scoring of applicants who leave fields blank in the applications form in the future.

3.7 Data Handling

Candidate variable construction was undertaken in which variable categorization was done and transformation carried out on the selected variables. Categorization of attributes was performed based on three criteria for binding attributes:

- Attributes with small number of observations were combined together
- Attributes with same default rate
- Based on business logic

Numeric variables were also transformed into categories by creating bins with different default rates and combined adjacent groups with similar default rates.

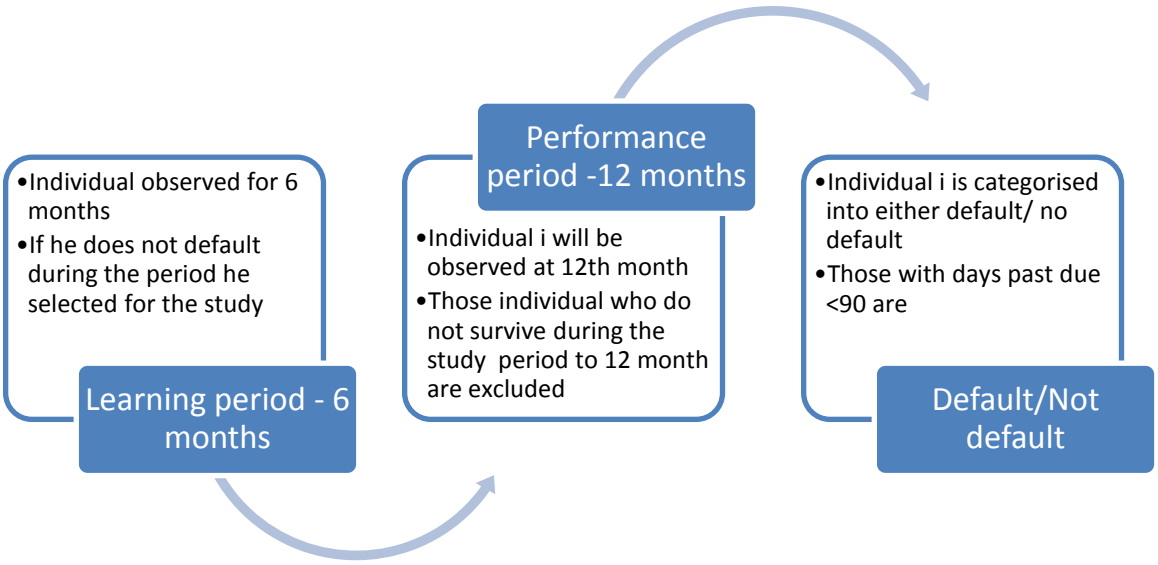
3.8 Data Design

The selection of the variables is from evaluation of the socio- economic and demographic environment of the customers. It was also based on an assessment of the information available. A key determinant of variable selection was the information provided by the loan officers as well as a thorough review of KCB loan application. Individuals who were selected met the following criteria:

i. That an individual ***i*** is a good customer (has not defaulted) in the first six month period (Learning Period) and has not become a bad customer at the observation point (at the end of six months).

ii. The selected individuals are then observed for the next twelve months to identify if they become bad or not and the number of days they take to become a bad customer. The twelve month period is the performance period.

3.9 Data Layout



Study data will be in such away that individual I will be observed for 6 months during the learning period. Individuals who have not defaulted i.e having estimated past due days less than 90 are selected for the study.

It is assumed that individual i will survive during the performance period of 12 months. Those individuals who do not survive during the period are excluded from the study. Table 1 below shows the data layout

Individual	**period**	$\boldsymbol{\pi}$	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	**…**	$\mathbf{X_p}$
1	t_1	π_1	X_{11}	X_{12}	X_{13}	…	X_{1p}
2	t_2	π_2	X_{21}	X_{22}	X_{23}	…	X_{2p}
.							.
.							.
.				.			
i	t_i	π_i	X_{i1}	X_{i2}	X_{i3}	…	X_{ip}
.							.
.							.
.							.
n	t_n	π_n	X_{n1}	X_{n2}	X_{n3}	…	X_{np}

These data consists of covariates $\mathbf{X_{ij's}}$ which are derived from the learning period of six months and the random variables **T** and $\boldsymbol{\pi}$ are obtained from the performance outcome.

Dependent variable

The dependent variable in this model is late repayment, labeled in the data as "defaultstastus"

(Default status). It is a dummy variable in which one is equivalent to if the repayment was late and zero if the card payments where on time.

Independent variables include

Estimated days past due; Amount past due, gender and credit limit amongst others

Method of data selection

They are usually chosen using stratified-random sampling that is separate sampling of several predefined groups.

3.9 Description of the Sample

The sample size is made up of 11391 borrowers for whom the credit cards were disbursed and or reimbursed during the period from 01st July 2011 to the 31st December 2012. In the sample we have 8340 good borrowers and 705 bad borrowers who were selected using Basel II accord. The sampling was based on the 20959 customers who applied for a credit card during that period.

3.10 Variable selection

Our goal is to select those variables that result in a 'best' model within the scientific context of the problem. Selection begins with careful univariable analysis for each variable. For nominal, ordinal and continuous variables with few integer values, contingency table of outcome (y=0,1) versus the n levels of the independent variable is done.

The likelihood ratio chi-square test with n-1 degrees of freedom is exactly equal to the value of the likelihood ratio test for the significance of the coefficients for the n-1 design variables in univariable logistic regression model that contains that single independent variable.
In the case of continuous variable, the univariate analysis involves fitting the univariable logistic regression model to obtain the estimate of the coefficient, standard error, the likelihood ratio test for the significance of the coefficient and the univariable Wald statistic.

Upon completion of univariable analyses, we selected variables for the multivariable analysis, based on the univariate test for any variable that had a p-value <0.25 to be included for the multivariable model along with all variables of known credit risk importance.
The strongest characteristics are then grouped. This applies to attributes in both continuous and discrete characteristics, and is done for an obvious reason.

The grouping is done because it is required to produce the scorecard. Scorecard can be produced using continuous (ungrouped) variables; however, grouping provides a number of advantages:

a) It provides an easier way to deal with outliers in interval variables and rare cases

b) Grouping simplifies the understanding of relationships; as a result gain more knowledge of the portfolio. A chart displaying the relationship between attributes of a characteristic and performance is a much more powerful tool than a simple variable strength statistic. It allows users to explain the nature of this relationship, in addition to the strength of the relationship.

c) It allows for nonlinear dependencies to be modeled by linear models

d) It allows unprecedented control over the development process by shaping the groups; one shapes the final composition of the scorecard.

e) The process of grouping characteristics allows the user to develop insights into the behavior of risk predictors and increases knowledge of the portfolio, which can help in developing better strategies for portfolio management.

Variable selection is done and the strongest characteristics are grouped and ranked. The strength of a characteristic is gauged using four main criteria:-

- Predictive power of each attribute. The weight of evidence (WOE) measure is used for this purpose.
- The range and trend of weight of evidence across grouped attributes within a characteristic.
- Predictive power of the characteristic. The Information Value (IV) measure is used for this.
- Operational and business considerations (e.g., using some logic in grouping postal codes, or grouping debt service ratio to coincide with corporate policy limits).

The first step into performing initial characteristic analysis is to perform initial grouping of variables, and rank order them by IV, this can be done by using a number of binning techniques. In this study, we started by binning variables into a large number of equal groups and calculation of WOE and IV for attributes and characteristics were done. The spreadsheet software was then used to fine-tune the groupings for the stronger characteristics based on principles outlined in the next section. Similarly for categorical characteristics, the WOE for each unique attribute and the IV of each characteristic were calculated. Sometime were then spent fine-tuning the grouping for those characteristics that surpass a minimum acceptable strength.

3.11 Logical Trend

The statistical strength, derived in terms of WOE and IV, is not the only factor in choosing a characteristic for further analysis, or designating it as a strong predictor. In grouped scorecards, the attribute strengths must also be in a logical order, and make operational sense. In other words groupings in this characteristic must have linear relationship with WOE; that is, they should denote a linear and logical relationship between the attributes in a characteristic and proportion of bads. This should conform to business experience in the credit. Establishing such logical (not necessarily linear) relationships through grouping is the purpose of the initial characteristic analysis exercise. The process of arriving at a logical trend is one of trial and error, in which one balances the creation of logical trends while maintaining a sufficient IV value.

3.12 Business logic

Other than statistical measures and logical trends, business logic contributes a very important component in developing credit risk scorecards. Characteristics included in the model must have business sense for the scorecard to be predictive and meet business requirements. Most of the business logics are embedded in the internal lending institution’s policies and manuals that guide the day to day operations of lending. The business rules may define what portfolios to be treated in a special way or not, as well as define characteristics that are known to affect the performance of default.

Upon undertaking the above steps, a multivariable logistic regression model was fitted; the importance of each variable included in the model was verified by an examination of the Wald statistic for each variable. Variables that did not contribute to model based on these criteria were eliminated. The new model was compared to the old, larger model using the likelihood ratio test. Further the estimated coefficients for the remaining variables were compared to those from the full model

Chapter Four

4.1 Data Analysis Tools

Financial tools that were used to calculate the creditworthiness of individuals which includes Descriptive Statistics (Frequency Distribution &Cross Tabulation), the Discriminant Analysis (DA), Logistic Regression analysis on SPSS 17.0 and Ms Excel for data maintenance.

4.2 Developing Credit Scoring Model

4.2.1 Extraction of factors: Principal Component Analysis Method

The main objective of the research is to apply both LA and DA to design & develop new and potentially more effective credit scoring model for Individuals. The 1st step was finding the different components affecting the creditworthiness of applicants.

While selecting the variables, we did principal component analysis to determine the factors to consider during model development: Out of the subjected 17 factors, 10 variables were selected for further univariate analysis. Most of these factors are socio- demographic variables.

There is a statistical significance for the factors under study ($p < 0.05$), at KMO 64.8%.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.648
Bartlett's Test of Sphericity	Approx. Chi-Square	38778.143
	df	136
	Sig.	.000

The table below represents the percentage of variability attributed to the model amongst the factors that were being investigated. Amount past due accounted for 90.9% of the variance of the extracted factors, average estimated past due days rated 86.7% while overdrawn amount rated at 83.6%. Other factors which rated above the threshold variation of 60% were average cash withdrawals (74.7%) gender (69.9%), card limit (69.6%), outstanding balance (69.1%), residential status (68.4%), brand name (66.4%), and marital status (64.1%). However KCBEmployee, level of education, dependants, average number of transactions, number of cards, and occupation were rated below the threshold variation thus disqualified to be included in the proposed model.

Table 1.1 Significance of the factors

Factors	Extraction
MaritalStatus	.641
KCBEmployee	.548
Gender	.699
Rstatus – Residential Status	.684
LEducation – Level of Education	.256
BrandName	.664
Occupation	.456
Avgoutstbal – Average Outstanding Balance	.691
Dependants	.374
Avgestdays - average Estimated past due Days	.867
Avglimit - Average Credit Limit	.696
Avgoverdrawn – Average overdrawn amount	.836
Avgamtpst – average amount past due	.909
Avgnotransc – average number of transactions	.587
Avgcashwtd – Average cash withdrawals	.747
NumberOfCards	.457
Defaultstatus – Default status	.377

The coefficients on individual variables may be insignificant [$p>0.05$] when the regression as a whole is significant. Intuitively, this is because highly correlated independent variables are explaining the same part of the variation in the dependent variable, so their explanatory power and the significance of their coefficients is "divided up" between them.

4.3 Description of the Sample

The sample size is made up of 11391 borrowers for whom the credit cards were disbursed and or reimbursed during the period from 01st July 2011 to the 31st December 2012. In the sample we have 8340 good borrowers constituting 92.2% and 705 bad borrowers at 7.8% who were selected using Basel II accord.

Card product is a lifestyle product, it targets individuals with certain characteristics and it requires discipline.

Table 1.2 Description of the sample under study

Default Status		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0- Non Defaulters	8340	92.2	92.2	92.2
	1-Defaulters	705	7.8	7.8	100.0
	Total	9045	100.0	100.0	

4.4 Non-financial Factors

4.4.1 Gender

Out of the entire population 5819 were females comprising of 64.3% of the population and 3226 males who constituted 35.7% of the total population

Table 1.3 below clearly demonstrates that there were 4.7 % defaulter female borrowers as compared to 13.3% of defaulter male borrowers, so females have less probability of default as compared to males. There were 86.7% of non-defaulter male borrowers as compared to 95.3% of non-defaulters female borrowers, so it is concluded that females were more creditworthy, they have less probability of default because they were more disciplined when it comes to the use of the credit card as compared to male borrowers.

Table 1.3: Gender and default status

			Gender		Total
			Female	Male	
defaultstatus	0	Count	5543	2797	8340
		% within Gender	95.3%	86.7%	92.2%
	1	Count	276	429	705
		% within Gender	4.7%	13.3%	7.8%
Total		Count	5819	3226	9045
		% within Gender	100.0%	100.0%	100.0%

We further tested the significance of the variable using logistic regression as show on table below. The variable contributes up to 52% of the predictive accuracy in the model. Therefore it's included in the model development.

		B	S.E.	Wald	df	Sig.
Step 1[a]	Gender(1)	-1.125	.081	194.963	1	.000
	Constant	-1.875	.052	1307.424	1	.000

4.4.2 Residential Status

Table 1.4 below shows that all those individuals who have their own house [Owner with mortgage] have high creditworthiness and less probability of default. Those without mortgage were rated 11.7% which was close to those who stays at rental houses at 11.3%. The business logic has it that individuals from both residential statuses don't have any similarities at all, they behave differently.

Table 1.4 Residential status and default status

			Residential Status					Total
			With Mortgage	Without Mortgage	Rental	Employer Owned	Living with Parents	
defaultstatus	0	Count	4537	953	2743	53	54	8340
		% within Rstatus	95.6%	88.3%	88.7%	73.6%	93.1%	92.2%
	1	Count	208	126	348	19	4	705
		% within Rstatus	4.4%	11.7%	11.3%	26.4%	6.9%	7.8%
Total		Count	4745	1079	3091	72	58	9045
		% within Rstatus	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

From the business logic it is assumed that individuals who have their own homes without mortgage are less likely to default than rest. The table above shows a contrary feedback since they behave the same way as those individuals who are renting. Hence the factor is eliminated from modeling as it's not significant.

		B	S.E.	Wald	df	Sig.
Step 1[a]	Rstatus			168.476	4	.000
	Rstatus(1)	-.480	.523	.842	1	.359
	Rstatus(2)	.579	.527	1.210	1	.271
	Rstatus(3)	.538	.521	1.065	1	.302
	Rstatus(4)	1.577	.583	7.313	1	.007
	Constant	-2.603	.518	25.227	1	.000

4.4.3 Brand Name

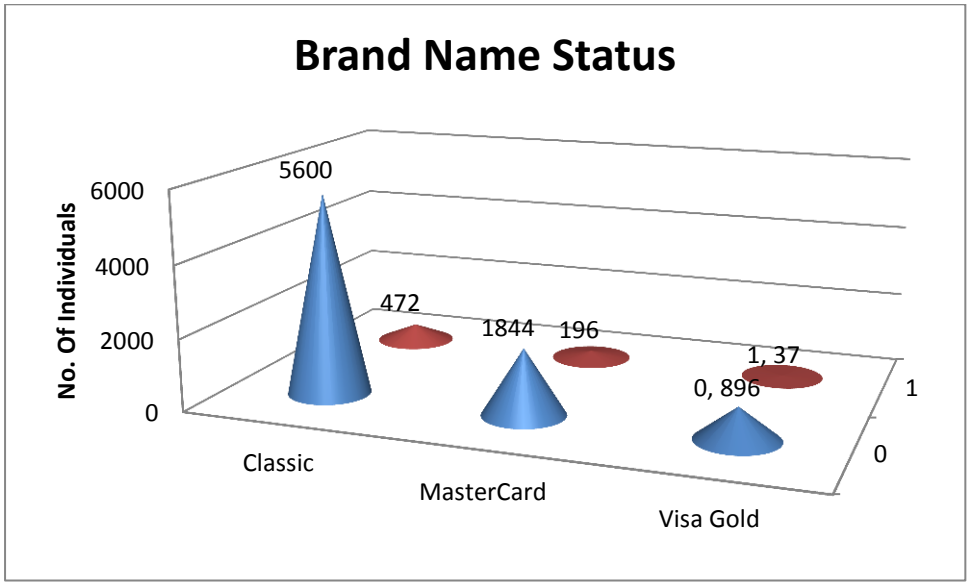
Brand name is an identifier of the type of card you have. Card products are classified to various categories that is: - International Classic, Visa Gold, Local Classic, MasterCard Corporate and MasterCard Co-Branded. In our study we have reclassified them into three categories:- Master Card, Visa Gold and International classic.

Master cards are named in respect to a given institutions or an event like Tuskys, serena etc. The other two have same characteristics although for gold card have a large limit starting from a minimum of KES 200,000/=.

Table 1.5 Crosstabulation of default status and BrandName

			BrandName			Total
			Classic	Mastercard	Visa Gold	
defaultstatus	0	Count	5600	1844	896	8340
		% within defaultstatus	67.1%	22.1%	10.7%	100.0%
		% within BrandName	92.2%	90.4%	96.0%	92.2%
	1	Count	472	196	37	705
		% within defaultstatus	67.0%	27.8%	5.2%	100.0%
		% within BrandName	7.8%	9.6%	4.0%	7.8%
Total		Count	6072	2040	933	9045
		% within defaultstatus	67.1%	22.6%	10.3%	100.0%
		% within BrandName	100.0%	100.0%	100.0%	100.0%

The above table shows that individuals with visa Gold cards are less likely to default as compared with individuals who hold mastercard and classic card who rated at 9.6% and 7.8% respectively.



By doing further univariate analysis using logistic regression, the variable is significant and it will be included in the construction of the proposed model. This will be supported by the business logic in the sense that high net-worth customers are less likely to default. This variable will go hand in hand with the card limit, in the sense that if a borrower comes for a higher limit then the brand will be relative. Since the variable is collinear to limit, its not included in the model.

4.4.4 Marital Status

Individuals customer statuses are classified to various categories:- married, single, divorced, widowed and separated. In study we have reclassified the mentioned categories into two that is married and single. This is because the other categories apart from the married group exhibit the same characteristics. In this study they both tended to have same behaviour, which is why they were grouped as single.

		defaultstatus			
MaritalStatus	Gender	0	1	Grand Total	% Default
Married	Female	4973	248	5221	5%
	Male	1118	134	1252	11%
Single	Female	570	28	598	5%
	Male	1679	295	1974	15%
Grand Total		8340	705	9045	8%

Table 1.6 Marital Status, gender against default status

The table above conforms that female on both status behave the same way as well men. Whichever the status whether married or not, women are les riskier [5%] as compared to men. Of the sample applicants, it can be shown that both single and married tend to have the same characteristics.

However, In general married individuals comprise of 5.9% of defaulters as compared to 12.6% of the individuals who are not married. It can therefore be concluded that married individuals are more responsible than unmarried individuals.

Business logic supports that married individual are more responsible than the single individuals. Therefore with the trend above it's noted that the two groups behave in the same manner. This suggest that majority of the respondents didn't give their correct identification status. Hence not significant and cannot be considered as a predictive factor for model development

4.4.5 Other non-financial factors

Of the sampled applicants, 7.6% of those individuals with professional education background were less likely to default than those individuals with elementary education who constituted 17.1% of the population. It is clear that majority of the individuals who take credit card are learned above the elementary education. So it is concluded that as the education level increases the creditworthiness also increases and probability of default decreases and vice versa [table 1.7]

Table 1.7 Default Status * Level of Education Crosstabulation

			LEducation		Total
			Elementary	professional	
defaultstatus	0	Count	142	8198	8340
		% within LEducation	83.0%	92.4%	92.2%
	1	Count	29	676	705
		% within LEducation	17.0%	7.6%	7.8%
Total		Count	171	8874	9045
		% within LEducation	100.0%	100.0%	100.0%

Table 1.8 defaultstatus * Occupation Crosstabulation

			Occupation				Total
			Employed	Businessman	H/wife	Pensioner	
defaultstatus	0	Count	8068	257	2	13	8340
		% within Occupation	92.7%	79.6%	100.0%	100.0%	92.2%
	1	Count	639	66	0	0	705
		% within Occupation	7.3%	20.4%	0.0%	0.0%	7.8%
Total		Count	8707	323	2	13	9045
		% within Occupation	100.0%	100.0%	100.0%	100.0%	100.0%

7.3% of Individuals who are salaried are less likely to default as compared 20.4% of the businessmen. H/wife and pensioners (unemployed group) do not default. Most of the unemployed individuals have supplementary cards from the family members who supports them financially and they are the ones who pays any amount past due.

Model Entropy

	Model Entropy
MaritalStatus	.387
Gender	.379
Rstatus	.382

Smaller model entropy indicates higher predictive accuracy of the binned variable on guide variable defaultstatus.

4.5 Financial Factors

Financial characteristics are concerned with the understanding of the personal resources available by examining the net worth and household cashflow. These characteristics are expressed through the attachment of an individual with a financial institution i.e. by the operation of the account within a certain period of time. They are referred financial behavioral characteristics- the historical behavior.

In our study we examined a group of individual cardholders who are bad and good

Model Entropy

	Model Entropy
avgoutstbal	.388
avgestdays	.338
avglimit	.386
avgoverdrawn	.365
avgamtpst	.344

Smaller model entropy indicates higher predictive accuracy of the binned variable on guide variable defaultstatus.

4.5.1 Overdrawn Amount

The extent to which the card is overdrawn i.e. the difference between the card account balance and the credit limit where the card balance is greater than the credit limit. That is the amount utilized by the individual above the limit set by the card holder. If an individual over exceeds the limit allocated, the likelihood of default is highly expected and this forms a significant parameter.

The overdrawn variable was binned based on the default status and categorized into three categories as shown below:

Table 1.9 Categorical Variables Codings

Bin	End Point		Number of Cases by Level of defaultstatus			
	Lower	Upper	0	1	Total	% Default
1	[a]	337	7038	378	7416	5.1%
2	337	3841	1045	204	1249	16.3%
3	3841	[a]	257	123	380	32.4%
Total			8340	705	9045	

Each bin is computed as Lower <= avgoverdrawn < Upper.

a. Unbounded

Therefore this variable is somehow directly related to the number of the transaction an individual carries out. Business logic has it that, it doesn't matter to what extent an individual can overdraw an account, as long us he/she is able to regularize the account. Hence, not considered fit for model predictive.

4.5.2 Outstanding Balance

This is the Total amount owed by the customer; this explains how an individual is performing under the credit card. The variable was binned and attributes of two categories were considered as show on table 1.11 below. Majority of the respondents fall below (2,374) and they have higher percentage of bad accounts at 9.2% as compare to the individual above (2,374) at 3.8%.

The variable is significant although business logic has it that the outstanding balance does not have so much implication as long as the individual borrower honors and regularises his account. Hence not included in our proposed predictive model

Table 1.11Categorical Variables Codings

Bin	End Point		Number of Cases by Level of defaultstatus			
	Lower	Upper	0	1	Total	% Default
1	[a]	-2374	6094	617	6711	9.2%
2	-2374	[a]	2246	88	2334	3.8%
Total			8340	705	9045	

Each bin is computed as Lower <= avgoutstbal < Upper.

a. Unbounded

4.5.3 Amount past due

The total overdue amount on the card account; this is the average amount the customer has missed to repay in the last six months. Tables 1.14 below show that individual with amount below (1,214) are more likely to default as compared to individuals whose amount overdue is above (377). It clear that the more the amount past due is high the more the applicant is likely to default, Making the attribute to be more predictive. The variable is significant and forms part of the modeling parameters

Table 1.14 Categorical Variables Codings

Bin	End Point		Number of Cases by Level of defaultstatus			
	Lower	Upper	0	1	Total	% Default
1	[a]	-1214	1077	361	1438	25.1%
2	-1214	-377	609	96	705	13.6%
3	-377	[a]	6654	248	6902	3.6%
Total			8340	705	9045	

Each bin is computed as Lower <= avgamtpst < Upper.

a. Unbounded

4.5.4 Number of transactions

These are the total number of transactions undertaken by the customer during the past six months. The grouping was done using business logic. It is assumed that the lesser the transactions during the month the likelihood of default on their payments is minimal. However table 1.15 below indicates on contrary. It indicates that those with less than 2 transactions have higher odds of defaulting (8.9%) as compared to those who have more than 2 transactions.

The reason could be that these customers are large corporates who occasionally record high number of transaction and are less likely to default. For this it may be difficult to ascertain the behavior of these card holders since their company meets the charges on their behalf. Therefore it's excluded from the modeling variables

Table 1.15 Categorical Variables Codings

Bin	End Point		Number of Cases by Level of defaultstatus		
	Lower	Upper	0	1	Total
1	[a]	2	5394	530	5924
2	2	[a]	2946	175	3121
Total			8340	705	9045

Each bin is computed as Lower <= avgnotransc < Upper.

a. Unbounded

4.5.4 Estimated past due days

The estimated average past due days was categorized into two attributes based on business logic, given that credit card and other credit facilities are billed on monthly basis. A cardholder who has not yet defaulted can be delinquent once or twice, that is he might have missed at most two repayments during the learning period. Table 1.16 below shows that 6% of bad cases fall within 30 days and below. The bad rate is highest among the cardholders whose days past due 30-90 days. Business logic is justified in the sense that all items below 30 days are assumed to be no default. The IV for this characteristic is sufficiently strong.

Table 1.16 Average estimated past due days

Bin	End Point		Number of Cases by Level of defaultstatus			
	Lower	Upper	0	1	Total	% Default
1	[a]	11	7487	342	7829	4.4%
2	11	25	439	105	544	19.3%
3	25	[a]	414	258	672	38.4%
Total			8340	705	9045	

Each bin is computed as Lower <= avgestdays < Upper.

a. Unbounded

4.5.6 Credit Limit

This is the total credit limit available on the customer's card(s). Card limit was categorized into two groups those with less than 20,833 limit and above 20,833. Table 1.17 below shows that the variable forms a significant variable as it agrees with the business logic that the higher the limit the lesser the default. The reason being that the more limit you have the more responsible you are and the high chances are that the individual is financially stable. Therefore the variable forms part of the model.

Table 1.17 Categorical Variables Codings

Bin	End Point		Number of Cases by Level of defaultstatus		
	Lower	Upper	0	1	Total
1	[a]	20833	2762	376	3138
2	20833	[a]	5578	329	5907
Total			8340	705	9045

Each bin is computed as Lower <= avglimit < Upper.

a. Unbounded

Chapter Five

5.0 Credit Scoring Models

For the purpose of determining creditworthiness of individuals we have used several credit scoring techniques such as credit scoring model for individuals, logistic regression (LR) and discriminant analysis (DA). We have used the LR and DA to compare the accuracy of the developed credit scoring model. We have discussed the results of each credit scoring model and also compared their results.

LR and DA were performed with the strongest set of characteristics chosen from the initial characteristics analysis, weak characteristics have been eliminated. All tests for significance are followed in selecting the final composition of the scorecard. The scorecard produced has measurable strength and impact. That can be used by Risk Managers and other decision makers like credit analysts to make a decision on how to control and monitor card holders.

5.1 Logistic Model and Model assessment

As we can see from table 1.1 above all factors are significant except for a few of which they did not meet the threshold of 60% significance. The logit model confirmed only four characteristic to be predictive classifying a customer on default or non-default. They include Gender, estimated past due days, amount past due, and credit limit.

Interpretation:

Logistic regression compares the model with the model including all the predictors to determine whether the latter model is more appropriate. Table 2.0 below tells us that the model is explained at 92.4% accuracy at a cut off value of 0.5

Table 2.0 Classification Table[a]

Observed			Predicted		
			defaultstatus		Percentage Correct
			0	1	
Step 1	defaultstatus	0	8276	64	99.2
		1	622	83	11.8
	Overall Percentage				92.4

a. The cut value is .500

The model appears good although we have to but we need to evaluate model fit and significance.

Model Chi- square: the overall significance is tested using chi-square which is derived from the likelihood that the model that has been fitted is accurate. There are two hypotheses to test in relation to the overall fit of the model;

i) H_0 - The model is a good fitting model.

ii) H_1 - The model is not a good fitting model (i.e. the predictors have a significant effect).

In our case chi- square has 6 degrees of freedom, a value of 1042.154 and a $p<0.000$ (table 2.1). Thus, the indication is that the model has a poor fit, with the model containing only the constant indicating that the predictors do have a significant effect.

Table 2.1 Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	1042.154	6	.000
	Block	1042.154	6	.000
	Model	1042.154	6	.000

We reject the null hypothesis as the variable does not make a significant contribution at $p<0.05$

Our ***H-L statistic*** has a significance of .709 which means that it is not statistically significant and therefore our model is quite a good fit (Table 2.2)

Table 2.2 Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	2.940	5	.709

We let the variables in the equation be:-

$x1 =$ $\text{avgestdays}_{\text{bin}}$
$x2 =$ $\text{avgestdays}_{\text{bin(1)}}$
$x3 =$ $\text{avgestdays}_{\text{bin(2)}}$
$x4 =$ $\text{avglimit}_{\text{bin(1)}}$
$x5 =$ $\text{avgamtpst}_{\text{bin}}$
$x6 =$ $\text{avgamtpst}_{\text{bin(1)}}$
$x7 =$ $\text{avgamtpst}_{\text{bin(2)}}$
$x8 =$ $\text{Gender}(1)$

Then our logistic regression model will be of the form

Variables	Limits	Estimates	S.E.	Wald(zscore)	Sig.	Exp(B)
x1	avgestdays < 11	0	0	84.612	0.000	0
x2	11 <= avgestdays < 25	-1.444	0.164	77.438	0.000	0.236
x3	25 <= avgestdays	-0.891	0.145	37.812	0.000	0.41
x4		0.85	0.09	89.564	0.000	2.34
x5	avgamtpst < -1214	0	0	99.754	0.000	0
x6	-1214 <= avgamtpst < -377	1.441	0.162	79.338	0.000	4.224
x7	-377 <= avgamtpst	1.136	0.142	64.2	0.000	3.115
x8		-0.989	0.088	125.585	0.000	0.372
Constant		-1.72	0.191	81.023	0.000	0.179

a. Variable(s) entered on step 1: avgestdays_bin, avglimit_bin, avgamtpst_bin, Gender.

We can interpret EXP(B) in terms of the change in odds. If the value exceeds 1 then the odds of an outcome occurring increase; if the figure is less than 1, any increase in the predictor leads to a drop in the odds of the outcome occurring. For example, the EXP(B) value associated with Credit card limit is 2.34. Hence when credit card limit is raised by one unit the odds ratio is 2.34 times as large and therefore limits are 2 more times likely to belong to the *non-defaulter* group.

Summary

A logistic regression analysis was conducted to predict probability of default of 8045 Individual (card borrower) using estimated past due days, credit limit, amount past due and gender as predictors. A test of the full model against a constant only model was statistically significant, indicating that the predictors as a set reliably distinguished between defaulters and non-defaulters of the credit card facility (chi square = 1042.154, p <.000 with df =6). H-L statistic of 0.709 indicated a moderately strong relationship between prediction and grouping. Prediction success overall was 92.4% (99.2% for non-defaulters and 11.8% for defaultert. The Wald criterion demonstrated that all predictors made a significant contribution to prediction (p =.000). Those factors which were not significant predictors were eliminated. EXP(B) value indicates that amount past due and Credit card limit when raised by one unit the odds ratio is 4 and 2 times as large and therefore individuals are 4 and 2 more times likely to default; respectively.

5.2 Discriminant Analysis and model assessment

Interpretation

In discriminant analysis we are trying to predict a group membership, so firstly we examined whether there are any significant differences between groups on each of the independent variables using group means and ANOVA results data. Tests of Equality of Group Means table below provide this information. In the ANOVA table 2.3 below, the smaller the Wilks's lambda, the more important the independent variable to the discriminant function. In our case the average estimated days is the most important variable for it produces a very high F value.

Table 2.3 Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
Gender	.977	216.367	1	9043	.000
Binned input variable avgestdays based on guide variable defaultstatus	.878	1253.132	1	9043	.000
Binned input variable avglimit based on guide variable defaultstatus	.987	118.767	1	9043	.000
Binned input variable avgamtpst based on guide variable defaultstatus	.911	879.641	1	9043	.000

Log determinants and Box's M tables

Box's M tests the null hypothesis that the covariance matrices do not differ between groups formed by the dependent. The researcher wants this test not to be significant so that the null hypothesis that the groups do not differ can be retained. For this assumption to hold, the log determinants should be equal. When tested by Box's M, we are looking for a non-significant M to show similarity and lack of significant differences. In this case the log determinants appear similar and Box's M is 891.344 with F = 88.952 which is significant at $p < .000$ (Tables 2.4 and 2.5).

Table 2.4 Log Determinants

defaultstatus	Rank	Log Determinant
0	4	-5.913
1	4	-4.262
Pooled within-groups	4	-5.686

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

Table 2.5Test Results

Box's M		891.344
F	Approx.	88.952
	df1	10
	df2	6452903.408
	Sig.	.000

Tests null hypothesis of equal population covariance matrices.

Table of Eigen values

This provides information on each of the discriminate functions (equations) produced. The maximum number of discriminant functions produced is the number of groups minus 1. We are only using two groups here, namely 'defaulters' and 'non-defaulters', so only one function is displayed. The canonical correlation is the multiple-correlation between the predictors and the discriminant function. This is the measure of association between the discriminant function and the dependent variable. The square of canonical correlation coefficient is the percentage of variance explained in the dependent variable.

In our case (Table 2.6) a canonical correlation of .391 suggests the model explains 15.29% of the variation in the grouping variable, i.e. whether a respondent a defaulter or not. The larger the eigenvalue, the more of the variance in the dependent variable is explained by that function.

Table 2.6 Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.180[a]	100.0	100.0	.391

a. First 1 canonical discriminant functions were used in the analysis.

Table 2.7 Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.847	1497.357	4	.000

Wilks' lambda is a measure of how well each function separates cases into groups. Smaller values of Wilks' lambda indicate greater discriminatory ability of the function.

This table (Table 2.7) indicates a highly significant function ($p < .000$) and provides the proportion of total variability not explained, i.e. it is the converse of the squared canonical correlation. So we have 84.7% unexplained.

The standardized canonical discriminant function coefficients table

Table 2.8 provides an index of the importance of each predictor. The sign indicates the direction of the relationship. Estimated past due days was the strongest predictor while gender was next in importance as a predictor. These two variables with large coefficients stand out as those that strongly predict allocation to the probability of default or not default group. Credit limit and amount past due were less successful as predictors.

Table 2.8 Standardized Canonical Discriminant Function Coefficients

	Function
	1
Gender	.327
Binned input variable avgestdays based on guide variable defaultstatus	.673
Binned input variable avglimit based on guide variable defaultstatus	-.276
Binned input variable avgamtpst based on guide variable defaultstatus	-.293

The structure matrix table

Table 2.9 provides another way of indicating the relative importance of the predictors and it can be seen below that the same pattern holds. We may consider the structure matrix correlations because they are considered more accurate than the Standardized Canonical Discriminant Function Coefficients. These Pearson coefficients are structure coefficients or discriminant loadings. They serve like factor loadings in factor analysis. Generally, just like factor loadings, 0.30 is seen as the cut-off between important and less important variables.

Table 2.9 Structure Matrix

	Function
	1
Binned input variable avgestdays based on guide variable defaultstatus	.877
Binned input variable avgamtpst based on guide variable defaultstatus	-.735
Gender	.364
Binned input variable avglimit based on guide variable defaultstatus	-.270

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions. Variables ordered by absolute size of correlation within function.

The canonical discriminant function coefficient table

These unstandardized coefficients *(b)* are used to create the discriminant function (equation). It operates just like a regression equation. In this case we have (Table 2.10):

$$D = 0.691x1 + 1.283x3 - 0.584x4 - 0.412x4 - 5526.734$$

Where x1= gender

X2= Estimated past due days

X3= Credit limit

X4= Amount past due

Table 2.10 Canonical Discriminant Function Coefficients

	Function
	1
Gender	.691
Binned input variable avgestdays based on guide variable defaultstatus	1.283
Binned input variable avglimit based on guide variable defaultstatus	-.584
Binned input variable avgamtpst based on guide variable defaultstatus	-.412
(Constant)	-5526.734

Unstandardized coefficients

Classification table

The classification results (Table 2.12) reveal that 86.3% of respondents were classified correctly into 'defaulters' or 'non-defaulters' groups. This overall predictive accuracy of the discriminant function is called the 'hit ratio'. Non-defaulters were classified with slightly better accuracy (88.8%) than defaulters (56.7%).

Table 2.12 Classification Results[a,c]

		defaultstatus	Predicted Group Membership		Total
			Non-defaulters	Defaulters	
Original	Count	0	7404	936	8340
		1	305	400	705
	%	0	88.8	11.2	100.0
		1	43.3	56.7	100.0
Cross-validated[b]	Count	0	7404	936	8340
		1	305	400	705
	%	0	88.8	11.2	100.0
		1	43.3	56.7	100.0

a. 86.3% of original grouped cases correctly classified.

b. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

c. 86.3% of cross-validated grouped cases correctly classified.

Summary

A discriminant analysis was conducted to predict whether the borrower was a defaulter or not. Predictor variables were gender, estimated days past due, amount past due and credit limit. Significant mean differences were observed for all the predictors on the DV. While the log determinants were quite similar, Box's M indicated that the assumption of equality of covariance matrices was violated. However, given the large sample, this problem is not regarded as serious. The discriminate function revealed a significant association between groups and all predictors, accounting for 15.29% of between group variability, although closer analysis of the structure matrix revealed only two significant predictors, namely gender (.0.364) and Estimated past due days (0.877). The cross validated classification showed that overall 86.3% were correctly classified.

5.3 Comparing Credit Scoring Models for Individuals

The results from the classification table 2.0 of LR shows that there are 83 applicants predicted to be bad or defaulters, comprising of 0.92% of the total population and there are 8276 applicants (91.50%) out of 9045 applicants who are above the cut point 0.5 and acceptable for the grant of loan, hence they are good applicants. The correct classification rate was 92.4% of LR having cut value equal to 0.5, as the P-value of LR shown to be lower than 0.01 so it resulted that default predictors are significantly related at the 95% confidence level.

There are two types of error which must be mentioned are Type I and Type II error. Type I error is predicting a bad credit application as a good credit application while Type II error is predicting a good credit application as a bad credit application. According to our results, there is 88.2% Type I error and the Type II error is 0.8%.

The result from the classification table 2.12 of DA shows that there are originally 400 applicants (56.7%) predicted to be bad and 7404 applicants (88.8%) as good applicants. It can be observed that Type I error rate is 43.3% and Type II error is 11.2%. There is 86.3% of accuracy that the original group cases correctly classified having cut value equal to 0.500, as the P-value of DA shown to be lower than 0.01, so it resulted that default predictors are significantly related at the 95% confidence level. The overall model is also significant as the p-value is less than 0.01.

Credit Scoring Model	Credit Scoring results		
	Bad-Bad (0-0)	Good-Good (1-1)	Accuracy rate*
LR	0.92%	91.50%	92.40%
DA	56.70%	88.20%	86.30%

* cut off point is 0.5

(LR) has the accuracy rate of 92.4%, with 0.92% accurately classified the bad applicants and 91.5% accurately predicted the good customers. The discriminant analysis credit scoring model has the accuracy rate of 86.3%, with 56.7% accurately classified the bad applicants and 88.2% accurately predicted the good applicants. Hence it is concluded from the credit scoring results that the proposed LR have the highest accuracy rate and also the most effective model as compared to discriminant analysis (DA).

Comparing Errors

Credit Scoring Model	Error Results	
	Type I	Type II
LR	88.2%	0.8%
DA	43.3%	11.2%

Discriminant analysis has the highest Type I as well as Type II error as compared to LR. The Misclassification cost of DA would be higher as compared to other two credit scoring models.

The credit scoring model which has the highest accuracy rate and lowest error rates are considered to be the most effective, accurate, efficient and useful model.

The results from credit scoring model for individuals proved that the gender and estimated past due days were the strong predictor of credit risk. We can estimate that female applicants are considered by banks to be less risky and more disciplined as compared to male counterpart.

5.4 Conclusion

This research study shows an evaluation of creditworthiness of individuals having credit card facility to improve the credit approval process and to decrease the non-performing loans in the commercial banks of Kenya.

In this research study we have taken a sample set of 9045 individual borrowers who have taken credit card from Kenya commercial bank of Kenya, out of which 8340 applicants who have clear history having no default ever, there were 705 who applicants have over 90 days default.

Logistic Regression and discriminant Analysis were applied to support the results of developed credit scoring model. The accuracy rate of Credit Scoring Model for Individuals on logistic regression (LR) was 92.4% and the discriminant analysis credit scoring model for individuals had the accuracy rate of 86.3%. It shows that proposed LR CSMI have the highest accuracy rate and also the most effective model as compared to discriminant analysis.

5.5 Recommendations

I would recommend commercial banks to apply the proposed credit scoring model as a part of their evaluation process. By adopting this model banks can reduce their non-performing loans. Non-financial factors have been included amongst other financial factors that banks consider but in a systematic way.

Future research studies are recommended to use the advanced credit scoring techniques like genetic algorithms, discriminant analysis and neural networks. For the generalization and accuracy of the results generated by the credit scoring models, it is recommended to have a large data of individual borrowers.

New variables can also be added to help in predicting the probability of default of individuals and corporations. It is highly advisable to collect the data of both accepted and rejected applicants by banks, so that more versatile results could be obtained.

References

Anderson, R. (2007). Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation. New York: Oxford University Press.

Anderson, T. W. (2003). An Introduction to Multivariate Statistical Analysis. New York: Wiley Interscience.

Agresti, A. 1996. *An Introduction to Categorical Data Analysis*. John Wiley and Sons.

Asia Samreen and Farheen Batul Zaidi (2012). Design and Development of Credit Scoring Model for the Commercial banks of Pakistan: Forecasting Creditworthiness of Individual Borrowers (September, 2012). *International Journal of Business and Social Science* Vol. 3 No. 17; September 2012.

Amelie Jouault and Allen M. Featherstone (2011). Determining the Probability of Default of Agricultural Loans in a French Bank. *Journal of Applied Finance & Banking,* vol.1, no. 1, 2011, 1-30

Baku, E., & Smith, M. (1998). Loan Delinquency in Community Lending Organizations: Case Studies of NeighborWorks Organizations. Housing Policy Debate, 9 (1), 151-175.

Banasik, J., & Crook, J. (2007). Reject inference, augmentation, and sample selection. *European Journal of Operational Research*, 183, pp 1582-1594. NeighborWorks Organizations. Housing Policy Debate, 9 (1), 151-175.

Barefoot, A. (1996, June). Credit Scoring at a Crossroads. *ABA Banking Journal*, 26.

Chen, M., & Huang, S. (2003). Credit Scoring and Rejected Instances Reassigning Through Evolutionary Computation Techniques. Expert Systems with Applications, 24, 433-441.

Chijoriga, M. M. (2011). Application of multiple discriminant analysis (MDA) as a credit scoring and risk assessment model . *International Journal of Emerging Markets,* 6 (2), 132-147.

Emel, A. B., Oral, M., Reisman, A., & Yolalan, R. (2003). A credit scoring approach for the commercial banking sector. Socio-Economic Planning Sciences, 37, 103–123.

Durand, D. (1941). Risk Elements in Consumer Finstalment Financing. National Bureau of Economic Research.

Emel, A. B., Oral, M., Reisman, A., & Yolalan, R. (2003). A Credit scoring approach for commercial banking sector. *Soc Econ Plan Sci 37*, 103-123.

Green, W. (1998). Sample Selection in Credit-Scoring Models. *Japan and the World Economy*, 10, pp 299-316.

Grigoris Karakoulas (2004). Empirical Validation of Retail Credit-Scoring Models. *The RMA Journal* September 2004

Hasan, I., & Zazzara, C. (2006). Pricing Risky Bank Loans in the New Basel 2 *Environment. Journal of Banking Regulation,* 7 (3-4), 243-267.

Hand, D., & Henley, W. (1997). Statistical Classification Methods in Consumer Credit Scoring: A Review. *Journal of Royal Statistical Society: Series A (Statistics in Society) 160 (3)*, 523-541.

Hosmer, D., & Lemeshow, S. (1989). *Applied Logistic Regression.* New York: John Wiley & Sons, Inc.

Hamid Sepehrdoust and Adel Berjisian (2012). Credit risk management of commercial banks in Iran; Using logistic model (28 January, 2013). *African Journal of Business Management Vol. 7(4), pp. 265-272,* 28 January, 2013

Kanwar, A. A. (2005). Risk Management for Banks. *Journal of Market Forces*, 1 (1), 1-7.

Lieli, R. P., & White, H. (2010). The Construction of Empirical Credit Scoring Models Based on Maximization Principles. *Journal of Econometrics*, 157 (1), 110-119.

Lopez, J. A., & Saidenberg, M. R. (2000). Evaluating credit risk models. *Journal of Banking & Finance*, 24, 151- 165.

Lee, T., Chui, C. Lu, & I. Chen. (2002). Credit Scoring Using the Hybrid Neural Discriminant Technique. *Expert Systems with Applications, 23*, 245-254.

Lee, T., & I.Chen. (2005). A Two-Stage Hybrid Credit Scoring Model Using Artificial Neural Networks and Multivariate Adaptive Regression Splines. *Expert Systems with Applications, 28*, 743-752.

Mugenda, O. M. and Mugenda, A.G. (1999) Research Methods: Quantitative and Qualitative Approaches. Nairobi: Acts Press.

Mester, L. (1997, September/October). *What's the Point of Credit Scoring? Federal Reserve Bank of Philadelphia Business Review* , 3-16.

Natasa Sarlija, Mirta Bensic and Marijana Zekic-Susac (2006). Modeling customer revolving credit scoring using logistic regression, survival analysis and neural networks. *Proceedings of the 7th WSEAS International Conference on Neural Networks, Cavtat, Croatia,* June 12-14, 2006 (pp164-169)

Orgler, Y. E. (1971). Evaluation of Bank Consumer Loans with Credit Scoring Models. *Journal of Bank Research* , 29, 31-37.

Ponicki, C. (1996). Case Study: Improving the Efficiency of Small-Business Lending at First National Bank of Chicago . *Commercial Lending Review*, 11 (2), 51-60.

Ousseni Kinda and Audrey Achonu (2012). Building A Credit Scoring Model For The Savings And Credit Mutual Of The Potou Zone (MECZOP)/Senegal. Consilience: *The Journal of Sustainable Development* Vol. 7, Iss. 1 (2012), Pp. 17–33

Sullivan, A. (1981). Consumer Finance, in Altman, E.I. Financial Handbook. New York: John Wiley & Sons.

Schreiner, M. (2000). Credit Scoring for Microfinance: Can It Work? *Journal of Microfinance Risk Management*, 2 (2), 105-118.

Schreiner, M. (2002). Scoring: *The Next Breakthrough in Microcredit? Occasional Paper No*. 7 , 6-7.

Siddiqui, N. (2006). Credit Scorecards. John Wiley&Sons Inc.

Steenackers, A., & Goovaerts, M. J. (1989). A Credit Scoring Model for Personal Loans. Insurance: Mathematics and Economics, 8, 31-34.

Sarlija, N., M. Bensic, & Z. Bohacek. (2004). Multinomial Model in Consumer Credit Scoring. *10th International Conference on Operation Research.* Trogir: Croatia.

Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002). Credit Scoring and its Applications, Society for Industrial Mathematics

Zekic-Susac, M., Sarlija, N., & Bensic, M. (2004). Small Business Credit Scoring: A Comparison of Logistic Regression, Neural Network, and Decision Tree Models. *26th Int. Conf. Information Technology Interfaces ITI.* Cavtat: Croatia.

Appendix

Discriminant Syntax

```
GET
  FILE='D:\FinaldataSAM.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.
DISCRIMINANT
  /GROUPS=defaultstatus(0 1)
  /VARIABLES=avgestdays_bin avglimit_bin avgamtpst_bin Gender
  /ANALYSIS ALL
  /SAVE=CLASS PROBS
  /PRIORS EQUAL
  /STATISTICS=MEAN STDDEV UNIVF BOXM COEFF RAW CORR TABLE CROSSVALID
  /PLOT=SEPARATE
  /CLASSIFY=NONMISSING POOLED.
```

Logistic Regression syntax

```
LOGISTIC REGRESSION VARIABLES defaultstatus
  /METHOD=ENTER avgestdays_bin avglimit_bin avgamtpst_bin Gender
  /CONTRAST (avgestdays_bin)=Indicator
  /CONTRAST (avglimit_bin)=Indicator
  /CONTRAST (avgamtpst_bin)=Indicator
  /CONTRAST (Gender)=Indicator
  /SAVE=PRED LRESID
  /PRINT=GOODFIT CI(95)
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```