# BEHAVIORAL CREDIT SCORING MODEL FOR CREDIT CARDHOLDERS

Omanyo John Oriema

Reg No: I56/64862/2010

UNIVERSITY OF NAIROBI

College of Biological and Physical Sciences

School of Mathematics

In Partial Fulfillment of the Requirement for the award of the degree

of

Master of Science in Social Statistics

# Declaration

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

Omanyo, John Oriema

I56/64862/2010

Signature Date

........................... 10/07/2012 ...........................

**Declaration by supervisor**

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

Dr. Mwaniki Joseph Ivivi

Signature Date

........................... 10/07/2012 ...........................

**Table of Contents**

# List of Tables and figures

## Acknowledgement

I am very grateful to my employer Kenya Commercial Bank, Lending Risk Unit for providing data for analysis for Masters of Science degree (Project). I would like to thank (Philip Gichangi Edward Osewe and John Njoroge) of Lending Risk team for their data mining and programming skills that have significantly contributed to the success of my MSc. project. May I thank the teaching and non-teaching staff in the School of Mathematics, University of Nairobi for their support. I would like 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.

I appreciate the support of my family and above all, I thank the Almighty God for good health throughout the study period.

## Dedication

I dedicate this degree to my family for their outstanding support, prayers and encouragement during my study period.

## Abstract

Credit risk being the most common problem facing credit card issuers has attracted a lot of interest in research. Default risk manifestations are more severe in card holders who have missed payments above 60 days and progression to default in the next one year is more rapid than in cardholders who missed payments less than 60 days. Default probability also hastens when average amount past due is more than KES 15000. The study has also found out that cardholders who utilize there card limits below 50% are less risk to default in the next twelve months.

We performed a retrospective cohort study at a large data set obtained from the local bank, and use Logit Model to identify predictors of default and risk factors among cardholders followed for a period of eighteen months, and in particular to determine the impact of financial and non-financial on the outcomes among cardholders. Results showed that, 72% of the cardholders who had average amount past due KES 15000defaulted within a period of twelve months.

## Chapter1

### 1.1 Background

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 honor the obligation. Access to credit 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.

According to Anderson (2007) 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). Available data is integrated into a single value that implies some quality, usually related to desirability or suitability. Scores are usually presented as numbers that represent a single quality, based on this concept he defines credit scoring as the use of statistical models to transform relevant data into numerical measures that guide credit decisions.

Credit scoring is therefore a technique mainly used in 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. In brief, 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.

Fundamentally, 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. Particular common aims can be categorized as follows: either as part of a Decision process, or Probability estimation.

## 1.2 Decision process

Application scoring: use scores to decide who to accept for a loan or other financial product and who to reject.

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

Predict probability of default, Measurement of expected profitability or return as well as capital requirement calculations as per required by regulatory authorities.

## 1.4 Statement of the problem

In recent years Kenyan banks have been faced with influx in number of credit applications and as result they have been faced with credit risk, the risk arising from the obligor (debtor) failure to meet the agreement to repay the debt upon advancement of the credit. The approach applied by credit analysts to monitor on whom to consider and good or a bad customer been subjective on the personal opinion of the analyst. This therefore has left banks with a serious challenge on the objective method on which to monitor the behavior of their customers and be able to identify those customers who are more likely to default in the next one year. In the long run the bank faces a high risk of reviewing and approving addition credit to customers who are very likely to default.

The risk of default is major risk that card issuers and banks are facing and this study focuses on how an objective approach with fewer inconsistencies can be applied on regular basis to discriminate low risk customers from high risk customers.

Predictive variables to be considered include, but are not limited to, prior credit performance, current level of indebtedness, card utilization, employee status, and number of days account has been delinquent.

**1.5 Main objective**

To determine the customers' behavioral characteristics that predicts the probability of a customer defaulting on his financial obligation with the bank.

**1.6 Specific objectives**

i. To determine the financial behavioral factors associated with a bad customer (defaulter).
ii. To determine a combination of non-financial behavioral factors that can be used to predict the probability of default.

## Chapter 2

### Literature Review

The traditional methods of deciding whether to grantor 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.

The objective of credit scoring models is to assign loan customers to either good credit or bad credit (Lee et al, 2002), or predict the bad creditors (Lim & Sohn, 2007). 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 (Lim & Sohn, 2007; Lee et al, 2002). 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; 11 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 paper 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 paper that such variables are implicitly deemed influential.

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 possess 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)

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.

Through the review of studies we can conclude that the evolution of credit scoring can be categorized into broad statistical techniques as follows;

### 2.1 Linear Regression

Linear regression has been used in credit scoring applications, as the two class problem can be represented using a dummy variable.

(Orgler, 1970) used regression analysis for commercial loans; this model was limited to the evaluation of existing loans and could be used for loan review and examination purposes. Later on, (Orgler, 1971) used a regression approach for evaluating outstanding consumer loans. He came to the conclusion that information not included on the application form had greater predictive ability than information included on the original application form, in assessing future loan quality. The use of regression analysis extended such applications to include further aspects (see, Henley, 1995; Hand & Henley, 1997; Hand & Jacka, 1998) as quoted by (Thomas, 2000).

### 2.2 Discriminant analysis

This is a statistical technique that is used to determine group membership, in this case good and bad groups. (Fisher, 1936) used discriminant analysis to differentiate groups in a population through measurable attributes, when the common characteristics of the members of the group are unobservable. (Durand, 1941), recognized that the same approach could be used to distinguish between good and bad loan. Much of the approach has been used in bankruptcy prediction field, notably (Beaver, 1966) empirical study. The author analyzed thirty financial ratios among failed and survived firms. Employing univariate analysis, three financial ratios i.e., total debt / total assets, net income/total assets and cash flow/total debt were found significant in determining financial distress of a company. (Altman, 1968) Study extended the work of (Beaver, 1966) by employing multivariate discriminant analysis to the prediction of corporate bankruptcy with what he called the "z-score". Using accounting data of 66 healthy and bankrupt companies, Altman calculated the financial ratios used by accountants and analysts to assess the solvency of business firms; he ended with up discriminant function to distinguish healthy companies from those with high probability of bankruptcy. He ended up with 5 variables suggesting a cutting point of z-score greater than 2.99 falls into —non-bankrupt category while firms having a z-score below 1.81 are all bankrupt. In his work the z-score is defined as:

$$z - score = 0.1717X_1 + .0847X_2 + 3.107X_3 + .420X_4 + .998X_5$$

These variables best discriminated between healthy and bankrupt companies were:

$$X_1 = \frac{Working\ Capital}{Total\ assets}$$

$$X_2 = \frac{Retained\ Earnings}{Total\ assets}$$

$$X_3 = \frac{EBIT}{Total\ assets}$$

$$X_4 = \frac{Market\ Value\ of\ shares}{Total\ assets}$$

$$X_5 = \frac{Working\ Capital}{Total\ assets}$$

Thus by obtaining the "z-score" for a particular company one can classify the company into healthy or bankrupt.

(Bandyopadhyay, 2006), using logistic and z-score approaches developed a model with high classification power of 91% to predict default for Indian firms. The new z-score model developed in his paper depicted not only a high classification power on the estimated sample, but also exhibited a high predictive power in terms of its ability to detect bad firms in the holdout sample. In the logit analysis, the empirical results reveal that inclusion of financial and non-financial parameters would be useful in more accurately describing default risk. (Bandyopadhyay, 2006), uses two approaches Multiple Discriminate Analysis (MAD) for developing z-score models for predicting corporate bond default in India and Logistic regression model to directly estimate the probability of default.

The known drawback of using discriminant analysis in credit scoring is the assumptions associated with it. Major one being high misclassification errors when predicting rare group, so equal sample for each group is usually used (Anderson, 2007). While linear Discriminant Analysis was the original, logistic regression is now preferred because of fewer assumptions.

### 2.3 Probit Regression

Probit model has been used in credit scoring applications for many years. The idea of probit was published in 1934 by Chester Bliss in an article in science on how to treat data such as the percentage of pest killed by pesticide. The method was carried forward in toxicological applications. (Grablowsky & Talley, 1981), noted that probit analysis was first pioneered for the analysis of "toxicology problems" by (Finney, 1952) who used it to "determine the relationship between the probability that an insect will be killed and the strength of the dose of poison administrated". However, early in the 1930s the term "Probit" was developed and stood for probability unit (Pindyck & Rubinfeld, 1997; Maddala, 2001). The probit model applies the inverse cumulative distribution function or quantile function associated with the standard normal distribution. (Grablowsky & Talley, 1981), stated that, under probit analysis, normal distributions of the "threshold values" are assumed, while multivariate normal distributions and equal variances are assumed under discriminant analysis; and using a likelihood ratio test, estimates of coefficients under a probit function can be tested individually for significance because of their uniqueness. But, this is not the case for discriminant coefficients, which cannot be individually tested, whilst this is possible in a regression as well as under a probit function, but the latter is much more difficult than that for a linear, logistic or Poisson regression model. Finally, they note that multicollinearity can cause, under probit analysis, incorrect signs for coefficients, although the probability values from the likelihood ratio tests are not affected. Otherwise, this problem is not an issue under discriminant analysis.

The application of probit analysis in credit scoring has also been investigated and compared with other statistical scoring models (Abdou, 2009c;Guillen & Artis, 1992; Banasik et al, 2003; Greene, 1998); also classification results were very close to other techniques (Green, 1998), and better than techniques, such as discriminant analysis, linear regression and the Poisson model (Guillen & Artis, 1992). Furthermore, probit analysis is used as a successful alternative to logistic regression.

## 2.4 Logistic regression

Logistic regression model differs from linear regression model is that the outcome variable in logistic regression is binary (0 or 1). (Wiginton, 1980) gave one of the first published accounts of logistic regression applied to credit scoring in comparison with discriminant analysis. He concluded that logistic regression gave superior classification. (Srinivasan & Kim, 1987), included logistic regression in comparative study with other methods, but in this case dealing with corporate credit granting problem. (Leonard, 1993), also applied logistic regression to a commercial loan evaluation process exploring several models, including a model using random effects for bank branches.

The simple logistic regression model can easily be extended to two or more independent variables. The additional of more variables, the harder it is to get multiple observations at all levels of all variables. For this reason the use of maximum likelihood estimation method has enable logistic regression to handled more than one independent (Freund & William, 1998). On theoretical grounds it might be supposed that logistic regression is a more proper statistical instrument than linear regression, given that the two classes "good" credit and "bad" credit have been described (Hand & Henley, 1997).

The application of logistic regression has been extensive in credit scoring applications (see for example: (Crook, et al, 2007), (Abdou, et al, 2008) and (Desai, et al, 1996). In building the scoring models, statistical techniques such as discriminant analysis, regression analysis, probit analysis and logistic regression, have been evaluated (Sarlija et al, 2004; Banasik et al, 2001; Greene, 1998; Leonard, 1992; Steenackers & Goovaerts, 1989; Boyes et al, 1989; Orgler, 1971). Other methods are: mathematical programming, nonparametric smoothing methods, Markov chain models, expert systems, neural networks, genetic algorithms and others (Hand & Henley, 1997). Also, case studies have been the subject of investigation in the credit scoring literature (see, for example: Lee & Chen, 2005; Lee et al, 2002; Banasik et al, 2001; Leonard, 1995; Myers & Forgy, 1963).

## Methodology

### 3.1 Study area

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

### 3.2 Study population

The populations consist of credit cardholders from the local bank who have not defaulted at the end of learning period which in this case is 6 months.

### 3.3 Study procedure

Behavioral scoring is associated with the account management of the existing consumers which are refreshed at regular intervals usually monthly for assessment of consumer credit risk (Anderson 2007). A sample of customers is chosen so that data available on their performance on either side of an arbitrarily chosen observation point in this case March 2010. The period prior to the observation time is called the performance or observation period and is usually 6 to 12 months in length (Thomas et al., 2001). 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.

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 etal, 2001). Basel II defines a bad customer to be someone who is 90 days past due or in excess of the agreed limit at this point. One of the disadvantages of behavioral scoring is the need for two years 20 worth of history to build a scorecard.

Consequently the population that the scorecard is then applied to may be quite different from that it was built on. One way used to reduce this is to take a shorter observation period and/or performance period of six months.

### 3.4 Data Source

The data used for this study are obtained from internal local bank's database which is dynamic in nature, the data frequently changes in the banking systems: these are credit card status, number of cash withdrawal, value of cash withdrawn and other past bad indicators.

### 3.5 Data Cleaning

Missing data was addressed through ignoring the variables with high percentage of NULLS or dealing with NULLs as separate attribute in categorical variables. When Inclusion of characteristics with missing values in the scorecard is done, then "missing" can then be treated as a separate attribute, grouped, and used in regression as an input. The scorecard would 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. (Siddiqi, 2006) acknowledges the importance of missing data in the credit scorecards, he noted that missing values may be part of a trend, linked to other characteristics, or indicative of bad performance. 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. If characteristics or records with missing values are excluded, none of these insights can be made. Therefore he recommendes that missing data be included in the analysis, and be assigned points in the final scorecard. His approach recognizes that missing data holds some information value, and that there is business benefit in including such data in your analysis.

## 3.6 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.

The learning period data were subjected to overtime transformation as follows:

**Average(x):** Average (amount past due in six months)

**Sum(x):** sum (no of rejected card payments, in six months)

**Ratio(x:y):** ratio (repayment amount to card limit)

## 3.7 Study Design

The study begins by selecting individuals from 18 months panel data such an individual meets the following criteria:

i. That an individual *i* is a good customer (has not defaulted) in the first six month period, which we consider the learning period and has not become a bad customer at the observation point (at the end of six months). Default refers to the 90 days past due without repayment by the individual $i$.

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.

Therefore, we define $\boldsymbol{T}$ to be a random variable for an individual's survival time and denote $\boldsymbol{t}$ to be any specific value of interest for random variable $\boldsymbol{T}$.

We denote $\pi$ to define a (0, 1) random variable indicating either an individual has become bad within the 12 months period or not (censored). Such that:

$$\pi = \begin{cases} 1, & \textit{if default/failure} \\ 0, & \textit{if non default/censored} \end{cases}$$

## 3.8 Data layout

The learning data is such individuals and their characteristics are observed and variables of interest taken at regular time period such that an individual $i$ is observed six times as shown in the Table 3.4a data layout.

| Individual | period | $\pi$ | $X_1$ | $X_2$ | $X_3$ | ... | $X_p$ |
|---|---|---|---|---|---|---|---|
| 1 | $t_1$ | $\pi_1$ | $X_{11}$ | $X_{12}$ | $X_{13}$ | ... | $X_{1p}$ |
| 2 | $t_2$ | $\pi_2$ | $X_{21}$ | $X_{22}$ | $X_{23}$ | ... | $X_{2p}$ |
| . | | | | | | | . |
| . | | | | | | | . |
| . | | | | | | | . |
| $i$ | $t_i$ | $\pi_i$ | $X_{i1}$ | $X_{i2}$ | $X_{i3}$ | ... | $X_{ip}$ |
| . | | | | | | | . |
| . | | | | | | | . |
| . | | | | | | | . |
| $n$ | $t_n$ | $\pi_n$ | $X_{n1}$ | $X_{n2}$ | $X_{n3}$ | ... | $X_{np}$ |

**Table 3 - 1: data layout**

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

### 3.9 The scorecard

A scorecard links the borrower characteristics to a credit score using a statistical model. In this regard a credit score is a linear combination of weighted variables values. Thus for $p$ borrower characteristics we have

$$s = \omega_0 + \sum_{i=1}^{p} \omega_i x_i \tag{3.1}$$

Where:

$s =$ credit score

$x_1, \dots, x_p$ are the borrower characteristics

$\omega_1, \dots, \omega_p$ are the weights on each characteristics $\omega_0$ is a constant term

Expressed in vector notation:

$$s = \omega_0 + \boldsymbol{\omega} \cdot \mathbf{X} \tag{3.2}$$

Where $\mathbf{X} = (x_1, \dots, x_p)$ and $\boldsymbol{\omega} = (\omega_1, \dots, \omega_p)$ are vectors of borrower characteristics and weights respectively. The vector of weights $\boldsymbol{\omega}$ form a scorecard which is used to score a customer. The credit score is linked to the risk of default. We use binary outcome $\pi \in \{0, 1\}$.

### 3.9.1 Credit score and Probability of default

The risk possessed by a customer is quantified by assigning a probability of default to the credit score $s$ using a link function:

$$P(Y = 0/s) = 1 - f(s) \tag{3.3}$$

Such that increasing scores reflect deterioration in creditworthiness, thus the link function should increase with increase in the score.

### 3.9.2 Log-odds function

The log-odds link function is used to link the scores,

$$s = log\left(\frac{P(y = 1 / s}{P(y = 0 / s)}\right) \tag{3.4}$$

This gives us the link function

$$P(y = 1/s) = f_0(s) = \frac{1}{1 + e^{-s}} \qquad (3.5)$$

The log-odds link function two properties:

- Greater resolution at extreme probabilities
- present in logistic regression

### 3.10 Logistic regression

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.

Logistic regression model differs from the linear regression model in that the outcome variable in logistic regression is *binary* or *dichotomous.* This distinction is reflected on their assumptions. In credit scoring problems we are normally interested on how several customers behavioral characteristics are related to the default or non-default of the customer. The outcome variable is default which is coded with a value of zero to indicate no default or 1 to indicate that the customer has defaulted.

In all regression problems the key quantity is the mean value of the outcome variable, given the value of the independent variable. This quantity is called the *conditional mean* and expressed as $E(Y/\mathbf{x})$, where $Y$ denotes the outcome variable (default or non-default) and $\boldsymbol{x}$ denotes a value of the independent variables (covariates).

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:

(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.

In this study we consider a collection of $\boldsymbol{p}$ behavioral covariates denoted by the vector $\boldsymbol{x}^T = (x_1, x_2, \dots, x_p)$. Then our conditional probability that the customer has defaulted be denoted by

$P(Y = 1/\boldsymbol{x}) = \pi(\boldsymbol{x})$. Then the logit of the multiple logistic regression model is expressed as

$$g(x) = log\left(\frac{P(y = 1 / s}{P(y = 0 / s)}\right) = s = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad (3.6)$$

Where $\beta_j$ $where\ j = 1,2, \dots p,$ are coefficients of the p covariates of the customers.obtained from the learning data set.

and the logistic regression model is now expressed as:

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}} \qquad (3.7)$$

### 3.10.1 Fitting multiple logistic regression model

Considering $n$ independent customers, $(x_i, y_i)$, $i = 1,2,\ldots,n$. Fitting the model requires that we obtain estimates of the vector $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \ldots, \beta_p)$ using maximum likelihood estimation method. From $P(y_i = 0) = 1 - \pi(x)$ and $P(y_i = 1) = \pi(x)$, the likelihood function becomes:

$$L(\boldsymbol{\beta}) = f(y_1, y_2, \ldots, y_n; \boldsymbol{\beta}) = \prod_{i=1}^{n} f_i(y_i; \boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x)^{y_i} (1 - \pi(x))^{1-y_i} \qquad (3.8)$$

Taking the logarithm of both sides of (3.8), we obtain the log likelihood function

$$\ln L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i \ln \pi(x) + (1 - y_i) \ln(1 - \pi(x))\} \qquad (3.9)$$

The model which best fits the data has values of $\boldsymbol{\beta}$ that maximize the likelihood equation (3.9). This is achieved by differentiation of this equation with respect to $\beta_0, \beta_1, \ldots, \beta_p$. The method yields $p + 1$ likelihood equations

The likelihood equation that result are expressed as

$$\sum_{i=1}^{n} [y_i - \pi(x)] = 0$$

and

$$\sum_{i=1}^{n} x_{ij}[y_i - \pi(x)] = 0, \qquad for\ j = 1,2, \ldots, p$$

Therefore, $\hat{\boldsymbol{\beta}}$ denotes the solution to these equations, and the fitted values for the multiple logistic regression model are $\hat{\pi}(x_i)$

The form of logistic regression is defined as:

$$p(x) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p}} \tag{3.10}$$

The logit function (3.10) can be transformed into:

$$g(x) = \ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \tag{3.11}$$

Logistic regression use Maximum Likelihood Estimation (MLE) to estimate the values of the unknown parameters which maximizes the probability of obtaining the observed set of data. The maximum likelihood estimators of these parameters are chosen to be those values that maximize the function.

## 3.11 Variable selection

### 3.11.1 Initial Characteristic Analysis

This involves the assessment of characteristics (variables) individually as a predictor to the performance of default. This is done to screen out illogical or non-predictive characteristics. As a result univariate analysis approach is performed on each variable. On categorical data with few integer values, a contingency table of outcome (d=0, 1) versus the values of $k$ levels of independent variable is applied. The likelihood ratio chi-squ re test with $k-1$ degrees of freedom is exactly equal to the value of the likelihood ratio test for the significance of the coefficients for the $k-1$ 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.

On 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. Although a scorecard can be produced using continuous (ungrouped) variables, however, grouping provides a number of advantages:

- It provides an easier way to deal with outliers in interval variables and rare cases
- 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.
- It allows for nonlinear dependencies to be modeled by linear models
- It allows unprecedented control over the development process by shaping the groups; one shapes the final composition of the scorecard.
- 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.

Finally when the strongest characteristics are grouped and ranked, variable selection is done. At the end of initial characteristic analysis, we have a set of strong, grouped characteristics, preferably representing independent information types, for use in the regression step.

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.

### 3.11.2 Weight of Evidence (WOE)

The WOE, as mentioned previously, measures the strength of each attribute, or grouped attributes, in separating good and bad accounts. It is a measure of the difference between the proportion of goods and bads in each attribute (i.e., the odds of a person with that attribute being good or bad). The WOE is based on the log of odds calculation:

$$Distr\ Good/Distr\ Bad \qquad (3.12)$$

This is a measure of odds of being good for a particular attribute in a selected characteristic.

A more user-friendly way to calculate WOE, and one that is used in this study is:

$$\left[\ln\left(\frac{Distr\ Good}{Distr\ Bad}\right)\right] * 100 \qquad (3.13)$$

The negative value from equation (3.13) implies that the particular attribute is isolating a higher proportion of bads than goods.

### 3.11.3 Information Value (IV)

Information Value, or total strength of the characteristic, comes from information theory, and is measured using the formula:

$$\sum_{i=1}^{n}(Distr\ Good_i - Distr\ Bad_i) * \ln\left(\frac{Distr\ Good_i}{Distr\ Bad_i}\right) \qquad (3.14)$$

(Siddiqi, 2006)Based on this methodology, one rule of thumb regarding IV is:

- Less than 0.02 : unpredictive
- 0.02 to 0.1: weak
- 0.1 to 0.3: medium
- 0.3 +: strong

The application of IV measure is widely used in the industry, and different practitioners have different rules of thumb regarding what constitutes weak or strong characteristics. In this regard, this study will consider all these characteristics which have met this threshold or not and find out the validity of the rule of thumb on IV suggested by (Siddiqi, 2006). (Anderson R. , 2007) noted that weak characteristics may: (i) provide value in combination with others; or (ii) have individual attributes that could provide value as dummy variables. They should thus not be discarded indiscriminately. Further, he quoted (Mays, 2004) who mentioned that even if these characteristics are not considered for the model, they should still be retained for future analysis.

In cases where the scorecard is being developed using non-grouped characteristics, statistics to evaluate predictive strength include R-square and Chi-square. These methods use goodness-of-fit criteria to evaluate characteristics. The R-squared technique uses a stepwise selection method that rejects characteristics that do not meet incremental R-square increase cutoffs. A typical cutoff for stepwise R-squared is 0.005. Chi-square operates in a similar fashion, with a minimum typical cutoff value of 0.5. The cutoffs can be increased if too many characteristics are retained in the model.

As with the technique using grouped variables, the objective here is to select characteristics for logistic regression (or another modeling step). Once these variables are identified we began with a model containing all of the selected variables in input variables are replaced by WOE.

### 3.11.4 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.

### 3.11.5 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 4

## Exploratory Data analysis

The panel data consists of customers followed from October 2010 to March 2012, with the following variables being measured; number of transactions, Number of cards, Amount in arrears, outstanding balance, employment status, Overdue cycles, number of cash withdrawals among others are selected and analyzed over six months period. The figure 4-1 shows the breakdown of periods for behavioral modeling.

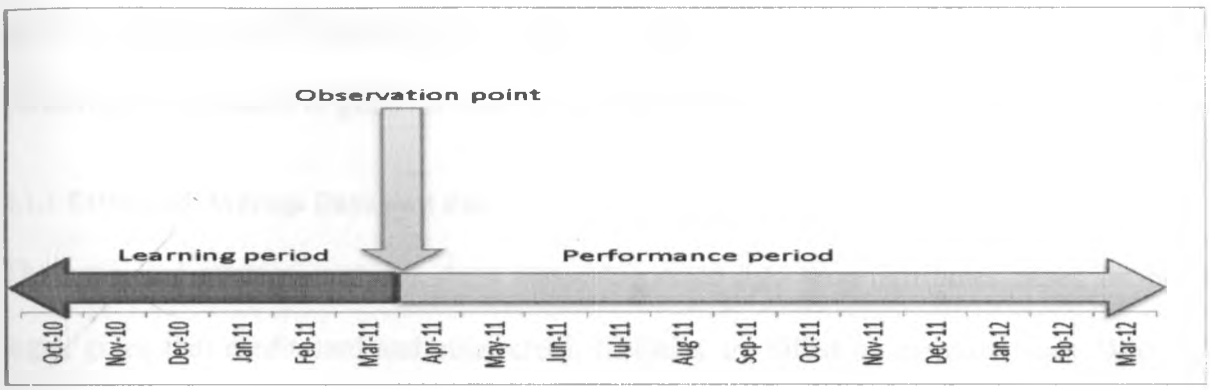


Figure 4 - 1 : period dates for behavioral modeling

Exploratory data analysis of data provides the purpose of formulating hypotheses worth testing, complementing the tools of conventional statistics for testing hypotheses. It is the approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to

- maximize insight into a data set
- uncover underlying structure
- extract important variables

- detect outliers and anomalies
- test underlying assumptions
- develop parsimonious models and

## 4.1 Financial Characteristics

EDA is carried on financial characteristics to achieve the bulleted desired results, this involves obtaining the counts of goods and bads from each characteristic, deriving both their good and bad distributions, the bad rate as well as WOE and IV. The total number of cardholders was 8,283 with 7,076 being good and 1,207 being bad. These numbers of goods and bads are same in all the characteristics. The distribution good and distribution bad columns give the column wise percentage distribution of good and bad cases respectively.

### 4.1.1 Estimated Average Days past due

The estimated average past due days was categorized into three attributes based on business logic, given that credit card and other credit facilities are billed on monthly basis. With this in mind 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 4-1; shows that 99% of good cases and 75% of bad cases fall into (0-29) days past due group. The bad rate is highest among the cardholders whose days past due range is (59-89). Both the bad rate and WOE are sufficiently different from one group to the next. These attributes therefore identify and separate good cases from bad cases and as a result this characteristic is a candidate for the modeling of the scorecard. The positive sign on the WOE for attribute (0-29) implies that the case with this attribute likely to be a good cardholder. The IV for this characteristic is sufficiently strong.

| Lower Limit | Higher Limit | Count | GOOD | BAD | Distribution Good | Distribution Bad | Bad Rate | WOE | IV |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 29 | 7,892 | 6,981 | 911 | 99% | 75% | 12% | 26.78 | 0.06 |
| 29 | 59 | 307 | 82 | 225 | 1% | 19% | 73% | (277.80) | 0.49 |
| 59 | 89 | 84 | 13 | 71 | 0% | 6% | 85% | (346.63) | 0.20 |
| | | 8,283 | 7,076 | 1,207 | 100% | 100% | 15% | | 0.75 |

Table 4 - 1: Estimated average days past due in the last 6 months by grouping

### 4.1.2 Number of cash withdrawals

The credit cards normally charge various types interest depending on the nature of transactions/card usage. If the cardholder choses to make cash withdrawal then he faces higher interest rate compared to someone who uses it for purchases of goods and payments of services. It is on this background that credit card issuer may find necessary to determine how this behavior relates to default. Table 4-2; shows analysis on how number cash withdrawals are related to good and bad customers. It is evident that the WOE of is sufficiently different from those who have not used and those that have used their credit cards for cash withdrawal. These two attributes therefore separates the good and bad customers. IV for this characteristic is low and suggests a weak relationship.

| Lower Limit | Higher Limit | Count | GOOD | BAD | Distribution Good | Distribution Bad | Bad Rate | WOE | IV |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 5,708 | 4,934 | 774 | 70% | 64% | 14% | 8.38 | 0.00 |
| 1 | 99999 | 2,575 | 2,142 | 433 | 30% | 36% | 17% | (16.98) | 0.01 |
| | | 8,283 | 7,076 | 1,207 | 100% | 100% | 15% | | 0.01 |

Table 4 - 2: Number of Cash withdrawals over the last 6 months

### 4.1.3 Overdrawn Amount

Overdrawn amount is the utilization of the card above the set limit by the card holder. In banking, consumption of more than what is allocated is considered a sign of bad behavior and the logical trend expected is that, the more one overdraws above their limit, the higher one is expected to default. This has been verified by this study as clearly indicated in Table 4-3. Bad rate and WOE are different from each attribute and meets the logical trend requirement. It also concur with the business logic that card holder who has not exceeded their limits have higher odds of being good while those who have exceeded their limit by (15,000 to 25,000) have lower odds of being bad compared to have (25,000 & above). This characteristic therefore becomes a candidate for modeling stage and is also supported by strong value of IV.

| Lower Limit | Higher Limit | Count | GOOD | BAD | Distribution Good | Distribution Bad | Bad Rate | WOE | IV |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 6347 | 5695 | 652 | 80% | 54% | 10% | 39.87 | 0.11 |
| 0 | 5,000 | 1702 | 1298 | 404 | 18% | 33% | 24% | (60.14) | 0.09 |
| 5,000 | 15,000 | 187 | 73 | 114 | 1% | 9% | 61% | (221.43) | 0.19 |
| 15,000 | 25,000 | 37 | 9 | 28 | 0% | 2% | 76% | (290.36) | 0.06 |
| 25,000 | & Above | 10 | 1 | 9 | 0% | 1% | 90% | (396.58) | 0.03 |
| | | 8283 | 7076 | 1207 | 100% | 100% | 15% | | 0.48 |

Table 4 - 3: Overdrawn amount on the credit card

### 4.1.4 Total Number of transactions

The number of transaction refers to the number of times that card has been used in the last six months in this study. In business sense it is expected that those card holders who do not conduct any transactions with their cards are not likely to default on their payments. Under normal circumstances, it is expected that the higher the number of transactions the higher the chances of default. However, Table 4-4 indicates otherwise. It indicates that those cases that have number of transactions (0 – 20), have higher odds of defaulting compared to those who have transactions above 20. 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 they company meets the charges on their behalf. The IV for this attribute is weak and thus variable should be excluded from the modeling stage.

| Lower Limit | Higher Limit | Count | GOOD | BAD | Distribution Good | Distribution Bad | Bad Rate | WOE | IV |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 20 | 6,711 | 5,664 | 1,047 | 80% | 87% | 16% | (8.04) | 0.01 |
| 21 | 40 | 1,067 | 941 | 126 | 13% | 10% | 12% | 24.21 | 0.01 |
| 41 | & above | 505 | 471 | 34 | 7% | 3% | 7% | 85.99 | 0.03 |
| | | 8,283 | 7,076 | 1,207 | 100% | 100% | 15% | | 0.05 |

Table 4 - 4: Number of transactions in the last 6 months.

### 4.1.5 Total value of transactions

The total value of transaction of the last six months shows some illogical pattern, Table 4-5 gives WOE of this characteristic with the IV of 0.02. The WOE trend indicates that the cases whose value of transactions are (0 -10,000) have higher odds of being good cases while cases whose value are contained in (10,000 to 60,000) are likely to be defaulters. However, those cases whose values of transactions are above 60,000 are most likely to be good. This can be either be false or true depending on the credit analysts opinion. To clearly determine the nature of the relationship, this characteristic would be included in the model for the scorecard and to find out if this behavior has an impact in the scorecard.

| Lower Limit | Higher Limit | Count | GOOD | BAD | Distribution Good | Distribution Bad | Bad Rate | WOE | IV |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 10000 | 5,311 | 4,551 | 760 | 64% | 63% | 14% | 2.12 | 0.00 |
| 10000 | 30000 | 1,722 | 1,451 | 271 | 21% | 22% | 16% | (9.07) | 0.00 |
| 30000 | 40000 | 293 | 235 | 58 | 3% | 5% | 20% | (36.94) | 0.01 |
| 40000 | 60000 | 365 | 307 | 58 | 4% | 5% | 16% | (10.22) | 0.00 |
| 60000 | &above | 592 | 532 | 60 | 8% | 5% | 10% | 41.37 | 0.01 |
| | | 8,283 | 7,076 | 1,207 | 100% | 100% | 76% | (12.73) | 0.02 |

Table 4 - 5: The total value of transaction in the last six months

### 4.1.6 Average amount past due

The average amount that the customer has missed repayment in the last six months has a very high correlation to default. Table 4-6 shows the relationship between the average amount past due and the performance of the cardholder. The customers who have average past due balance (-15,000 and below) have higher odds of defaulting, this attribute is predictive. The cases with past due amount above (-5,000) have higher odds of being good customers. The bad rates and WOE are significantly different in each attribute. The characteristic has a very strong IV value of 0.6 and therefore this characteristic is a candidate for modeling.

| Lower Limit | Higher Limit | Count | GOOD | BAD | Distribution Good | Distribution Bad | Bad Rate | WOE | IV |
|---|---|---|---|---|---|---|---|---|---|
| Below | (15,000) | 123 | 35 | 88 | 0% | 7% | 72% | (269.06) | 0.18 |
| (15,000) | (5,000) | 391 | 161 | 230 | 2% | 19% | 59% | (212.52) | 0.36 |
| (5,000) | &above | 7,769 | 6,880 | 889 | 97% | 74% | 11% | 27.77 | 0.07 |
| | | 8,283 | 7,076 | 1,207 | 100% | 100% | 15% | | 0.60 |

Table 4 - 6: Average amount past due in the last 6 months

### 4.1.7 Outstanding balance

The current outstanding balances in the credit cards determine the performance of the cardholders. Table 4-7 describes the relationship on how different attributes in this characteristic relates to the default of a case. Attributes of balance (below 40,000) have higher odds of being bad customer than good. However, on the other hand cases with outstanding balance (above 5,000) have higher odds of being good customers. This is further supported by the significant differences in both bad rates and WOE for each attribute. The IV above 0.1 indicates that these characteristic has a moderate relationship with default, and should be included in the modeling stage.

| Lower Limit | Higher Limit | Count | GOOD | BAD | Distribution Good | Distribution Bad | Bad Rate | WOE | IV |
|---|---|---|---|---|---|---|---|---|---|
| (40,000) | &below | 1,337 | 1,058 | 279 | 15% | 23% | 21% | (43.56) | 0.04 |
| (40,000) | (20,000) | 1,640 | 1,334 | 306 | 19% | 25% | 19% | (29.62) | 0.02 |
| (20,000) | (15,000) | 824 | 692 | 132 | 10% | 11% | 16% | (11.18) | 0.00 |
| (15,000) | (5,000) | 1,924 | 1,680 | 244 | 24% | 20% | 13% | 16.08 | 0.01 |
| (5,000) | &above | 2,558 | 2,312 | 246 | 33% | 20% | 10% | 47.20 | 0.06 |
| | | 8,283 | 7,076 | 1,207 | 100% | 100% | 15% | | 0.12 |

Table 4 - 7: Outstanding balance at the observation point

### 4.1.8 Card utilization

Card utilization is the ratio of outstanding balance to the credit card limit. The attributes assigned to this ratio has shown bad rate and WOE are sufficiently different from one group to another. These attributes can therefore identify and separate customers into the good and bad group. Table 4-8 shows how cases that have utilization above 100% have higher odds of becoming bad compared to other categories of this characteristic. In this regard this characteristic is selected for modeling of the scorecard.

| Lower Limit | Higher Limit | Count | GOOD | BAD | Distribution Good | Distribution Bad | Bad Rate | WOE | IV |
|---|---|---|---|---|---|---|---|---|---|
| 100% | &above | 499 | 297 | 202 | 4% | 17% | 40% | (138.31) | 0.17 |
| 100% | 50% | 2,719 | 2,260 | 459 | 32% | 38% | 17% | (17.45) | 0.01 |
| 50% | below | 5,065 | 4,519 | 546 | 64% | 45% | 11% | 34.49 | 0.06 |
| | | 8,283 | 7,076 | 1,207 | 100% | 100% | 15% | | 0.25 |

Table 4 - 8: Card utilization at the observation point

## 4.2 Non-Financial characteristics

Non-financial characteristics refer to those characteristics that are already categorical and they do not change overtime. Most of these characteristics are derived from the variables at the observation point.

### 4.2.1 Employee category

The characteristic of employee category has distinctive and sufficiently different bad rate and WOE. The measures indicate that non-bank employees are highly to default compared to their counterpart employed in the bank. The IV for this characteristic is sufficiently strong enough and therefore a candidate to development of the credit risk scorecard. Table 4.2-1 illustrates how these two attributes are related to the credit card performance.

| Bank Employee | Count | GOOD | BAD | Distribution Good | Distribution Bad | Bad Rate | WOE | IV |
|---|---|---|---|---|---|---|---|---|
| No | 5,163 | 4,137 | 1,026 | 58% | 85% | 20% | (37.43) | 0.10 |
| Yes | 3,120 | 2,939 | 181 | 42% | 15% | 6% | 101.88 | 0.27 |
| | 8,283 | 7,076 | 1,207 | 100% | 100% | 15% | | 0.37 |

Table 4.2 - 1: Bank Employee category

### 4.2.2 Brand Name

The credit card performance has weak relationship with the credit card brand name; this is illustrated by the IV value in Table 4.2-2. Regardless of the weak relation the WOE indicates the direction and strength of the attributes in this characteristic. It can be deduced that Visa Gold cardholders have higher odds of default compared to the other two brand names. This variable to our opinion should be a candidate for modeling.

| Brand Name | Count | Good | Bad | Distribution Good | Distribution Bad | Bad Rate | WOE | IV |
|---|---|---|---|---|---|---|---|---|
| Others | 1,026 | 915 | 111 | 13% | 9% | 11% | 34.08 | 0.01 |
| Internl. Classic | 6,236 | 5,342 | 894 | 75% | 74% | 14% | 1.91 | 0.00 |
| Visa Gold | 1,021 | 819 | 202 | 12% | 17% | 20% | (36.88) | 0.02 |
| | 8,283 | 7,076 | 1,207 | 100% | 100% | 15% | | 0.03 |

Table 4.2 - 2: Credit card brand name

### 4.2.3 Card General status

General status of the card does not have any relationship with the performance of the credit card. Table 4.2-3 illustrates the nature of relation, even though the bad rate and WOE indicate sufficiency in the two attributes, the IV is close to zero. In this regard the variable general status is not selected for modeling purposes.

| Gen-Status | Count | Good | Bad | Distribution Good | Distribution Bad | Bad Rate | WOE | IV |
|---|---|---|---|---|---|---|---|---|
| BLCK | 36 | 27 | 9 | 0% | 1% | 25% | (67.00) | 0.00 |
| NORM/NOAU | 8,247 | 7,049 | 1,198 | 100% | 99% | 15% | 0.37 | 0.00 |
| Grand Total | 8,283 | 7,076 | 1,207 | 100% | 100% | 15% | | 0.00 |

Table 4.2 - 3: General status of the card

### 4.2.4 Authorization Status

The authorization status of the credit card has negligible value of IV. The bad rate and WOE has relatively predictive. But due to IV value the variable is not predictive if it were included in the modeling. Therefore it is dropped based on result indicated in Table 4.2-4.

| Aut-Status | Count | Good | Bad | Distribution Good | Distribution Bad | Bad Rate | WOE | IV |
|---|---|---|---|---|---|---|---|---|
| NOAU | 39 | 31 | 8 | 0% | 1% | 21% | (41.40) | 0.00 |
| NORM | 8,244 | 7,045 | 1,199 | 100% | 99% | 15% | 0.23 | 0.00 |
| Total | 8,283 | 7,076 | 1,207 | 100% | 100% | 15% | | 0.00 |

Table 4.2 - 4: Authorization status of the card

# Chapter 5

## Credit Scoring Model

Once a list of characteristics for inclusion in the scorecard is obtained, these characteristics are then regressed again as a group, to obtain final regression parameters.

The process involves combining statistical modeling technique with business considerations in "designing" a scorecard that is strong and stable. In this study the characteristics from financial and non-financial are combined to represent different independent information types that together form a risk profile.

At this stage, logistic regression is performed with the strongest set of characteristics chosen from the initial characteristics analysis, and that all weak criteria 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. Most importantly, it is a useful business tool that can be used by Risk Managers and other decision makers to create risk-adjusted strategies for monitoring their card holders.

### 5.1 Logit Model

The logit model confirmed only three financial and one non-financial characteristic to be predictive classifying a customer on default or non-default. The four characteristics are therefore used for developing the scorecard. Current outstanding balance on the card account, value of transaction, overdrawn amount on credit cards and credit card brand type are not predictive enough in estimating the cardholder default in the next one year. Table 5-1 illustrates the characteristics that meet business logic and statistically significant.

| Coefficients: | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -2.02363 | 0.06022 | -33.602 | < 2e-16 |
| PD Days 0-29 | 0 | | | |
| 30-59 | 2.22307 | 0.1812 | 12.269 | < 2e-16 |
| 60-89 | 2.64993 | 0.35783 | 7.406 | 1.31E-13 |
| Amount Past Due <5000* | 0 | | | |
| 5000-15000 | 1.05963 | 0.15945 | 6.646 | 3.02E-11 |
| Above 15000 | 1.1306 | 0.29707 | 3.806 | 0.000141 |
| Card Utilization <50% | 0 | | | |
| 50% -100% | 0.6039 | 0.09394 | 6.429 | 1.29E-10 |
| >=100% | 1.65353 | 0.16586 | 9.969 | < 2e-16 |
| Employee Category No | 0 | | | |
| Yes | -1.64742 | 0.12066 | -13.653 | < 2e-16 |

Table 5 - 1: Logistic regression model output

*excluded attribute

The model attributes have small standard errors compared to the respective coefficients estimates. P-values are smaller than 5%, indicating that these characteristics are predictive enough and conclude that they are statistically significant. These results can assist banks to predict those customers who are likely to default in the next one year. From the twelve characteristics only four have statistical significance in explaining default in the next twelve months.

## 5.2 Model Assessment

### 5.2.1 Deviance and Pearson Chi-Square

The residual deviance for the model is 3767.5 on 5790 degrees of freedom and null deviance of 4751.1 on 5797 degrees of freedom. The change in deviance between the null and residual is 1083.6 on 7 degrees of freedom. With 7 degrees of freedom, a $\chi^2 = 1083.6$ has a P-value very close to zero. Such a small P-value indicates that such a large change in deviance is not attributable to chance alone. That is Days Past Due, Amount Past Due, Utilization and Employment status of a cardholder are significant explanatory variables in the prediction of whether cardholder will default in the next one year or not. Table 5-2 gives the model assement statistics.

| |
|---|
| (Dispersion parameter for binomial family taken to be 1) |
| Null deviance: 4751.1 on 5797 degrees of freedom |
| Residual deviance: 3767.5 on 5790 degrees of freedom |
| AIC: 3783.5 |
| Number of Fisher Scoring iterations: 6 |

Table 5 - 2: Model Fit statistics

### 5.2.2 The Receiver Operating Characteristic (ROC) Curve and K-S statistic

The ROC curve is the plot of $X = \Pr[S_{FP} \leq S_{Cut-off}]$ against $Y = \Pr[S_{TP} \leq S_{Cut-off}]$ as the cut-off is varied, where X= sensitivity, the true positive rate, or hit rate and $\mathrm{Y} = 1 - \text{specificity}$, which is the false positive rate, or false alarm rate. The results for the model with four characteristics are shown in figure 5-1.

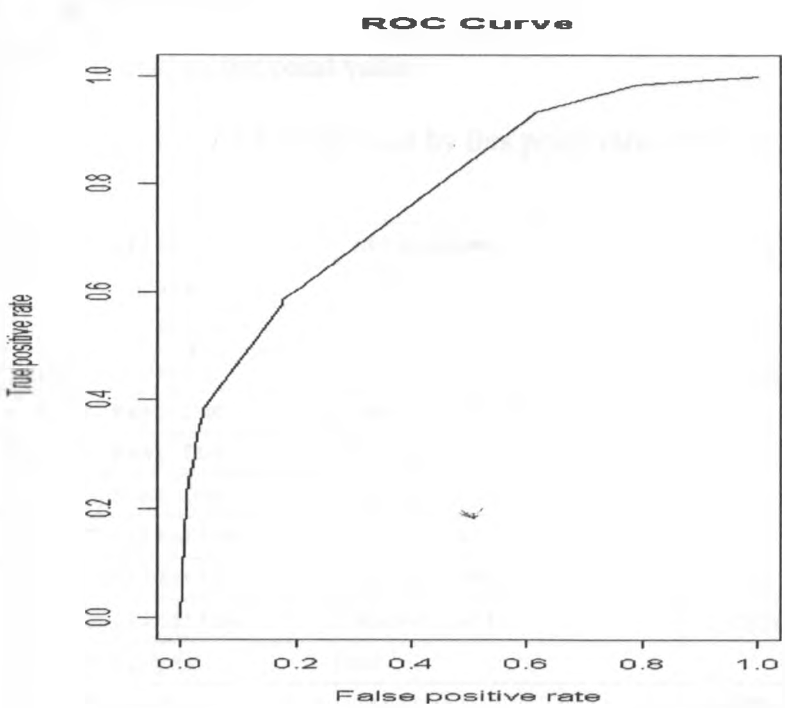


Figure 5 - 1: ROC curve of the model of the four characteristics

The Receiver Operating Characteristic (ROC) curve for the model indicates that inclusion of the four variables in the model yields a better model compared to a random model whose line would be $45^0$ from the origin.

#### 5.2.3 Kolmogorov-Smirnov (K-S)

This measures the maximum (deviation) between the cumulative distributions of bads and goods. The K-S statistic gives a value of 0.413, this could be used as a cutoff point at which to reject a cardholder whose PD is above this point.

### 5.3 The Scorecard

The scorecard is developed using the Table 5.1 three columns. The following procedure was followed to develop the scorecard:

i. Observing the category with the highest value in each characteristic and summing over these values.
ii. Calculate the point value
iii. Multiply each coefficient by this point value from step (ii) to obtain the score.

| Characteristic | Attributes | Estimate | Score |
|---|---|---|---|
| Past Due Days | 0-29 | 0 | 0 |
| Past Due Days | 30-59 | 2.43402 | 170 |
| Past Due Days | 60-89 | 2.69408 | 189 |
| Amount Past Due | Less than 5000 | 0 | 0 |
| Amount Past Due | Between 5000 &15000 | 1.08345 | 76 |
| Amount Past Due | Above 15000 | 1.2897 | 90 |
| Card Utilization | less than 50% | 0 | 0 |
| Card Utilization | Between 50% & 100% | 0.56371 | 39 |
| Card Utilization | Above 100% | 1.49746 | 105 |
| Bank Employee | No | 0 | 0 |
| Bank Employee | Yes | 1.66267 | 116 |

Table 5 - 3: The scorecard

Table 5 - 3: The scorecard

| | |
|---|---|
| sum | 7.14391 |
| The highest score | 500 |
| point value | 69.99 |

## Chapter 5

### Conclusions

This study has found out that average past due days on card, average amount past due in the last six months and card utilization are the key three financial characteristics that can be used to determine the probability of default in the next one year of the credit card holder.

Besides, the status of whether a cardholder is a bank employee or not has proved to be one of the non-financial characteristic that can be applied in monitoring the performance of the cardholder.

### Recommendations

The project has only identified four key characteristics that can be used to estimate the probability of default of customer in the next twelve months. However, there are other characteristics that need to be studied to find out how they affect the default. These include different category of credit card types, total number of transactions and the duration the customer has been using the card- banking relationship.

## Reference


Abdou, H. (2009). Genetic programming for credit scoring: The case of Egyptian public sector banks. *Expert Systems with Applications*, 36.

Abdou, H., & Pointon, J. (2011). Credit Scoring Techniques and Evaluation Criteria. *Intelligence Systems in Accounting, Finance & Management*, 59-88.

Abdou, H., El-Masry, A., & Pointon, J. (2007). On The Applicability of Credit Scoring Models in Egyptian Banks. *Bank and Bank Systems / Vol. 2,Issue 1*, 4-20.

Abdou, H; Pointon, J; El Masry, A. (2008). Neural nets versus conventional techniques in credit scoring in Egyptian banking. *Expert Systems with Applications*, 35 (3):1275-1292.

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and Prediction of Corporate Bankruptcy. *The Journal of Finance, Vol. 23, No.4*, 589-609.

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.

Banasik, J., & Crook, J. (2007). Reject inference, augmentation, and sample selection. *European Journal of Operational Research*, 183, pp 1582-1594.

Bandyopadhyay, A. (2006). Predicting probability of default of Indian corporate bonds: logistic and Z-score model approaches. *Journal of Risk Finance, The Vol 7 Iss: 3*, 255-272.

Beaver, W. (1966). Financial Ratios as Predictors of failure. *Journal of Accounting Research*, 4, 71-111.

Belloti, T., & Crook, J. (2009). Credit scoring with macroeconomic variables using survival analysis. *The Journal of the Operation Research Society*, Vol 60, No.12, pp 1699-1707.

Crook, J., Edelman, D., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operation Research*, 183(3):1447-1465.

Desai, V S; Crook, J N; Overstreet, G A;. (1996). A comparison of Neural networks and Linear Scoring Models in the Credit Union Environment. *European Journal of Operation Research*, 95 (1): 24-37.

Durand, D. (1941). *Risk Elements in Consumer Instalment Financing.* New York: 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.

Finney, P. (1952). *Probit analysis.* Cambridge, M.A: Cambridge University Press.

Fisher, R. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics 7 (2)*, 179-188.

Freund, R., & William, W. J. (1998). *Regression Analysis: Statistical modelling of response variable.* San Diego: Academic press.

Grablowsky, B., & Talley, W. (1981). Probit and discriminant functions for classifying credit applicants: a comparison. *Journal of Economic and Business, 33(3)*, 254-261.

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

Guillen, M., & M. Artis. (1992). Count Data Models for a Credit Scoring System. *The European Conference Series in Quantitative Economics and Econometrics on Econometrics of Duration, Count and Transition Models.* Paris.

H, M., & Forgy, E. (1963). "The development of numerical credit evaluation systems". *Journal of the American Statistical Association.*

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.

Hsieh, N. C. (2004). An integrated data mining and behavioral scoring model for analyzing bank customers. *Expert Systems with Applications 27*, 623-633.

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.

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

Leonard, K. J. (1993). Who's Keeping Score?: A Credit Scoring Model for Commercial Loan Applications. *Business Credit*, Vol. 95, No. 10 pp 8.

Lim, M., & Sohn, S. (2007). Cluster-based dynamic scoring model. *Expert Systems with Applications, 32*, 427-431.

Orgler, Y. E. (1970). A credit scoring model for commercial loans. *Journal of money, Credit and Banking*, Vol. 2 (November), 435-445.

Orgler, Y. E. (1971). Evaluation of Bank Consumer Loans with Credit Scoring Models. *Journal of Bank Research, 2*, 31-37.

Pindyck, R. S., & Rubinfeld, D. L. (1997). *Econometric Models and Economic Forecasts.* Irwin: McGraw-Hill.

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

Siddiqi, N. (2006). *Credit Risk Scorecards: Developing and Implementing Intelligence Credit Scoring.* New Jersey: John Wiley & Sons, Inc.

Srinivasan, V., & Kim, Y. H. (1987). Credit Granting: A Comparative Analysis of Classification Procedures. *Journal of Finance*, 42(3), 665-681.

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

Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International journal of forecasting 16*, 149-172.

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

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.