University of Nairobi

School of Mathematics

Modelling Credit Risk for Personal Loans Using Product-Limit Estimator

By

Okumu Argan Wekesa

146/60006/07

A Research Project Submitted In Partial Fulfillment for the Award of Post Graduate Diploma in Actuarial Science

November, 2010

DECLARATION AND CERTIFICATION

Candidate

This is to certify that this research project is my original work and that it has not been presented elsewhere for an award

Name O. A. WEILESA ..... signature ..... date 24/11/2010

Supervisor

This is to certify that this research project has been submitted with my approval as the supervisor

Name PROFESSOR R. U. SIMWA ..... signature ..... date 23/11/2010

ACKNOWLEDGEMENT

I wish to recognize the contribution of Prof. Simwa my supervisor for his thoughtful guidance and patience in supporting me through the research project. To be appreciated is also management of Cooperative Bank for accepting to allow access to their loans data.

TABLE OF CONTENTS

CHAPTER ONE

INTRODUCTION

1.1 Background to the Study

Over the past decade, a number of the world's largest banks have developed sophisticated systems in an attempt to model the credit risk arising from important aspects of their business lines. Such models are intended to aid banks in quantifying, aggregating and managing risk across geographical and product lines. The outputs of their models, also play increasing important roles in bank's risk management and performance measurement process, including performance – based compensation, customer profitability analysis, risk based pricing and to a lesser degree, active portfolio management and capital structures decisions.

In retail banking, credit risk models aid the decision of whether to grant credit to an application or not. Traditionally, this is done by estimating the probability that an applicant will default. This aim has been changing in recent years towards choosing the customers of highest profit. That change means it now becomes important not only if but when a customer will default (Thomas et. al 1990). It is possible that if the time to default is long, the acquired interest will compensate or even exceed losses resulting from default. Traditionally, credit scoring aimed at distinguishing good payers from bad payers at the time of the application. The timing when customers default is also interesting to investigate since it can provide the bank with the ability to do profit scoring. Analysing when customers default is typically tackled using survival analysis.

It has been shown previously by Thomas et. al (1999) and Narain (1992) that survival analysis can be applied to estimate the time to default or to early repayment. The major strength of survival analysis is that it allows censored data to be incorporated into the model. This translates in the consumer credit context as a customer who never defaults, or never pays off early, so an event of interest is not observed. Clearly, there is a great amount of such data because, luckily, most of the customers are "good".

This approach to using survival analysis to estimate time to default has also been used to model credit risk in the pricing of bonds and other financial investments. In his Ph. D. thesis, Lando (1994) introduced a proportional hazards survival-analysis model to estimate the time until a bond defaults, the aim being to use economic variables as covariates.

In credit scoring we look for differences in application characteristics for customers with different survival times. Also, it is possible that there are two or more types of failure outcome. In consumer credit we are interested, in several possible outcomes when concerned with profitability: early repayment, default, closure etc.

The idea of employing survival analysis for building credit-scoring models was first introduced by Narain (1992) and then developed further by Thomas *et al.* (1999). Narain (1992) applied the accelerated life exponential model to 24 months of loan data. The author showed that the proposed model estimated the number of failures at each failure time well. Then a scorecard was built using multiple regression, and it was shown that a better credit-granting decision could be made if the score was supported by the estimated survival times. Thus it was found that survival

analysis adds a dimension to the standard approach. The author noted that these methods can be applied to any area of credit operations in which there are predictor variables and the time to some event is of interest.

Thomas *et al.* (1999) compared performance of exponential, Weibull and Cox's nonparametric models with logistic regression and found that survival-analysis methods are competitive with, and sometimes superior to, the traditional logistic-regression approach. Furthermore, the idea of competing risks was employed when two possible outcomes were considered: default and early payoff.

It was noted by Thomas *et al.*, (1999) that there were several possible ways of improving the performance of the simplest survival-analysis models, such as Weibull's, exponential, or Cox's proportional hazards models.

Due to lack of detailed updated information about the counterparty, the traditional approaches such as Merton's firm- value model are not applicable. This motivates a statistical model based on survival analysis under extreme censoring for the time- to- default variable.

A common definition of the risk of default is that a borrower is unable to meet a specific financial obligation. Mathematically this may be quantified as a probability that a certain event occurs. Let i be a borrower and D_i the default indicator at time t of the borrower i, defined by

$$D_i(t) = \begin{cases} 1 & \textit{if borrower default} \\ 0 & \textit{otherwise} \end{cases}$$

The risk of default at time t of borrower i is the probability $P\{D_i(t) = 1\}$. The time t is called time to default or failure time.

The model incorporates the stochastic nature of default and is based on incomplete information. In survival analysis, one must consider a key analytical problem called censoring. In essence, censoring occurs when we have some information about an individual's survival time, but do not know the exact survival time. There are a number of types of censoring, such as random, interval, left, and right censoring. In credit scoring application, most of the cases are right censoring.

For example, suppose we follow a group of borrowers for 3 years. If we observe borrower A fails to repay at 15^{th} month, he is certainly classified as a default case and his default time is 15. On the other hand, consider borrower B, who repays on time during the whole observed period. We do not know his exact default time but are sure that it must be greater than 36. For such case, borrower B is known as a right censored observation. Another example of right censoring could be when borrower C repays on time from the 1^{st} month to the 12^{th} month. At the 12^{th} month, we do not have future repayment pattern of borrower C. As borrower B, we do not know the exact default time of borrower C, we only know that it must be greater than 12. This is also a right censoring example.

1.2 Statement of the Problem

Since the year 2003, the Kenyan financial market has experienced growing liquidity, which has caused banks to rigorously market various loan products. This has given rise to the need to review the banks' credit granting criteria to reflect the growing volume of loan portfolio and to respond to the current global credit crunch. However, research on credit risk has surprisingly received insignificant attention from both practitioners and scholars. Over the years, banks have perpetually used traditional credit scoring techniques the rate loan applicants.

A number of studies have been carried out on the issue of credit risk modelling using different approaches. A limited number of studies have applied survival analysis techniques but none has used product-limit method to analyse credit risk. To this end, the research intended to model probability of servicing loans and hazard rates for various risk groups using this method. Furthermore, existing credit scoring models classify borrowers into different risk categories but cannot provide any information on when the borrower is likely to default. It is more informative for the lender not only to know the probability of defaulting but also when the default is likely to happen. This helps to fairly price risks and improve the focus on ultimate profitability. For instance if the lender knows that a group of loan applicants are bad type, instead of rejecting their applications, it may grant loans to them at higher interest rate, as long as the term of the loan is shorter than the likely time to default. Thus some "bad" applicants can also be viewed as profitable propositions. The traditional structural models currently used by most institutions are unable capture this. It is also worth noting that the banks have traditionally and consistently categorized borrowers in terms of some risk groups. Accordingly, there is an apparent need to test whether these classifications constitute homogeneous risk groups.

1.3 Objectives of the Study

1.3.1 General objective

The broad objective of this research was to use product- limit survival model to generate default probabilities at various points in time. The study also intended to perform a test of homogeneity on the various risk groups.

1.3.2 Specific objectives

The specific objectives included:

1. To estimate time to default using product-limit estimator for each risk group (male and female).
2. To determine hazard rate for each risk group on the basis of product-limit estimator.
3. To test the statistical significance of the differences in the survival curves for each risk group based on log-rank tests.

1.4 Justification of the Study

Survival analysis is a relatively new application that offers an advantage of predicting time to the event of interest, and therefore lays the foundation for estimating the applicants' profitability. This is superior to the traditional logistic regression approach which assumes that accounts which that do not experience default are 'good', non defaulting accounts whilst survival analysis treats such accounts in a more conservative way, as those that proved to be 'good' so far.

The outcome of the research is expected to put to light the reliability and consistency or otherwise of the survival methods of data analysis. Likewise, credit risk analysts may draw from the research on better approaches to classifying loan applicants based on risk characteristics.

Besides, the research outcome may also serve as the basis for setting risk premium to be loaded on the base rates.

Survival analysis approach has the following strengths:

1. Survival analysis is able to account for censoring, unlike the other techniques
2. Unlike linear regression, survival analysis has a binary outcome, which more realistic
3. It analyses time to default rather than mere probability of defaulting.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This section reviews some of the most important works about failure prediction methodologies. After first analyzing the most popular alternative statistical techniques that can be used to develop credit risk models, focus switches to the works that have investigated the problem of modelling credit risk for personal loans using survival techniques. The review ranges from historical background in 2.2 to survival techniques in 2.20

In the literature, various techniques have been described. Most of them are Classification Models in which the main purpose is to distinguish good type borrowers from the bad type. Another technique is to make use of survival model. Its major advantage is the ability to predict the time-to-default, which is never known under Classification Models. Since survival model is the basis of the proposed modelling, it will be deeply investigated after brief reviews of Classification Models.

2.2 History of Credit Scoring

Credit scoring is essentially a way of recognising the different groups in a population when one cannot see the characteristic that separates the groups but only related ones. This idea of discriminating between groups in a population was introduced in statistics by Fisher (Fisher 1936). He sought to differentiate between two varieties of iris by measurements of the physical size of the plants and to differentiate the origins of skulls using their physical measurements. David Durand (Durand 1941) in 1941 was the first to recognise that one could use the same

techniques to discriminate between good and bad loans. His was a research project for the US National Bureau of Economic Research and was not used for any predictive purpose. At the same time some of the finance houses and mail order firms were having difficulties with their credit management. Decisions on whether to give loans or send merchandise had been made judgementally by credit analysts for many years. However these credit analysts were being drafted into military service and there was a severe shortage of people with this expertise. So the firms got the analysts to write down the rules of thumb they used to decide to whom to give loans (Johnson 1992). These rules were then used by non-experts to help make credit decisions – one of the first examples of expert systems. It did not take long after the war ended for some folk to connect these two events and to see the benefit of statistically derived models in lending decisions. The first consultancy was formed in San Francisco by Bill Fair and Earl Isaac in the early 1950s and their clients at that time were mainly finance houses retailers and mail order firms

The arrival of credit cards in the late 1960s made the banks and other credit card issuers realise the usefulness of credit scoring. The number of people applying for credit cards each day made it impossible both in economic and manpower terms to do anything but automate the lending decision.

When these organisations used credit scoring they found that it also was a much better predictor than any judgmental scheme and default rates would drop by 50% or more – see (Myers 1963) for an early report on such success or Churchill *et al.* (Churchill, Nevin, Watson 1977) for one from a decade later.

The only opposition came from those like Capon (Capon 1982) who argued "that the brute force empiricism of credit scoring offends against the traditions of our society". He felt that there should be more dependence on credit history and it should be possible to explain why certain characteristics are needed in a scoring system and others are not. The event that ensured the complete acceptance of credit scoring was the passing of the Equal Credit Opportunity Acts (ECOA 1975, ECOA 1976) in the US in 1975 and 1976. These outlawed discriminating in the granting of credit unless the discrimination could be statistically justified. It is not often that lawmakers provide long term employment for any one but lawyers but this ensured that credit scoring analysis was to be a growth profession for the next 25 years. This has proved to be the case and still is the case. So the number of analysts in the UK has doubled even in the last four years.

In the 1980s the success of credit scoring in credit cards meant that banks started using scoring for their other products like personal loans, while in the last few years scoring has been used for home loans and small business loans. Also in the 1990s the growth in direct marketing has led to the use of scorecards to improve the response rate to advertising campaigns. In fact this was one of the earliest uses in the 1950s when Sears used scoring to decide to whom to send its catalogues (Lewis 1992).

Advances in computing allowed other techniques to be tried to build scorecards. In the 80s logistic regression and linear programming, the two main stalwarts of today's card builders, were introduced.

More recently Artificial Intelligence techniques like expert systems and neural networks have been piloted.

At present the emphasis is on changing the objectives from trying to minimise the chance a customer will default on one particular product to looking at how the firm can maximise the profit it can make from that customer. Moreover the original idea of estimating the risk of defaulting has been augmented by scorecards which estimate response (how likely is a consumer to respond to a direct mailing of a new product), usage (how likely is a consumer to use a product), retention (how likely is a consumer to keep using the product after the introductory offer period is over), attrition (will the consumer change to another lender) and debt management (if the consumer starts to become delinquent on the loan how likely are various approaches to prevent default).

2.3 The methods used for credit scoring

Originally credit was based on a purely judgmental approach. Credit analysts read the application form and said yes or no. Their decisions tended to be based on the view that what mattered was the 3Cs or the 4Cs or the 5Cs. (Capon 1982)

Credit scoring nowadays is based on statistical or operational research methods. The statistical tools include discriminant analysis which is essentially linear regression, a variant of this called logistic regression and classification trees, sometimes called recursive partitioning algorithms. The Operational Research techniques include variants of linear programming. Most scorecard

builders use one of these techniques or a combination of the techniques. Credit scoring also lends itself to a number of different

non-parametric statistical and modelling approaches. Ones that have been piloted in the last few years include the ubiquitous neural networks, expert systems, genetic algorithms and nearest neighbour methods. It is interesting that so many different approaches can be used on the same classification problem. Part of the reason is that credit scoring has always been based on a pragmatic approach to the credit granting problem. The object is to predict who will default not to give explanations for why they default or answer hypothesis on the relationship between default and other economic or social variables. That is what Capon (Capon 1982) considered to be one of the main objections to credit scoring in his critique of the subject.

A sample of previous applicants is taken, which can vary from a few thousand to as high as hundreds of thousands, (not a problem in an industry where firms often have portfolios of tens of millions of customers). For each applicant in the sample, one needs their application form details and their credit history over a fixed period - say 12 or 18 or 24 months.

One then decides whether that history is acceptable, i.e. are they bad customers or not, where a definition of a bad customer is commonly taken to be someone who has missed three consecutive months of payments. There will be a number of customers where it is not possible to determine whether they are good or bad because they have not been customers long enough or their history is not clear. It is usual to remove this set of "intermediates" from the sample.

One question is what is a suitable time horizon for the credit scoring forecast – the time between the application and the good/bad classification. The norm seems to be twelve to eighteen months. Analysis shows that the default rate as a function of the time the customer has been with the organisation builds up initially and it is only after twelve months or so (longer usually for loans) that it starts to stabilise.

Thus any shorter a horizon is underestimating the bad rate and not reflecting in full the types of characteristics that predict default. A time horizon of more than two years leaves the system open to population drift in that the distribution of the characteristics of a population change over time, and so the population sampled may be significantly different from that the scoring system will be used on.

One is trying to use what are essentially cross-sectional models, i.e. the ones that connect two snapshots of an individual at different times, to produce models that are stable when examined longitudinally over time. The time horizon – the time between these two snapshots - needs to be chosen so that the results are stable over time. Another open question is what proportion of goods and bads to have in the sample. Should it reflect the proportions in the population or should it have equal numbers of goods and bads. Henley discusses some of these points in his thesis (Henley 1995).

Credit scoring then becomes a classification problem where the input characteristics are the answers to the application form questions and the results of a check with a credit reference bureau and the output is the division into 'goods' and 'bads'. One wants to divide the set of

answers A into two subsets $X \in A_B$ the answers given by those who turned out bad, and $X \in A_G$, the set of answers of those who turned out to be good. The rule for new applicants would then be – accepted if their answers are in the set A_G; reject if their answers are in the set A_B. It is also necessary to have some consistency and continuity in these sets and so we accept that we will not be able to classify everyone in the sample correctly. Perfect classification would be impossible anyway since, sometimes, the same set of answers is given by a 'good' and a 'bad'. However we want a rule that misclassifies as few as possible and yet still satisfy some reasonable continuity requirement.

2.4 Behavioural Scoring

Behavioural scoring systems allow lenders to make better decisions in managing existing clients by forecasting their future performance. The decisions to be made include what credit limit to assign, whether to market new products to these particular clients, and if the account turns bad how to manage the recovery of the debt. The extra information in behavioural scoring systems compared with credit scoring systems is the repayment and ordering history of this customer. Behavioural scoring models split into two approaches - those which seek to use the credit scoring methods but with these extra variable added, and those which build probability models of customer behaviour. The latter also split into two classes depending on whether the information to estimate the parameters is obtained from the sample of previous customers or is obtained by Bayesian methods which update the firm's belief in the light of the customer's own behaviour. In both cases the models are essentially Markov chains in which the customer jumps from state to state depending on his behaviour.

In the credit scoring approaches to behavioural scoring one uses the credit scoring variables and includes others which describe the behaviour. These are got from the sample histories by picking some point of time as the observation point. The time preceding this -say the previous 12 months - is the performance period and variables are added which describe what happened then- average balance, number of payments missed. etc. A time some 18 months or so after the observation point is taken as the performance point and the customer's behaviour by then is assessed as good or bad in the usual way. Hopper and Lewis (Hopper Lewis 1992) give a careful account of how behavioural scoring systems are used in practice and also how new systems can be introduced. They advocate the Champion v Challenger approach where new systems are run on a subset of the customers and their performance compared with the existing system. This makes the point yet again that it takes time to recognise whether a scoring system is discriminating well.

The choice of time horizon is probably even more critical for behavioural scoring systems than credit scoring systems. Behavioural scoring is trying to develop a longitudinal forecasting system by using cross-sectional data, i.e. the state of the clients at the performance period end and at the end of the outcome period. Thus the time between these periods will be crucial in developing robust systems.

Experimentation (and data limitations) usually suggests a 12 or 18 month period. Some practitioners use a shorter period, say 6 months, and then build a scoring system to estimate which sort of behaviour at six months will lead to the client eventually defaulting and define this six month behaviour as "bad".

One can use older data for the second scorecard while using almost current data for the main scorecard. The probability models classify the different states the consumer can be in using variables from the application form and variables describing current and recent behaviour, for example – balance outstanding, number of periods since a payment was made, average balance. The following example takes this approach to a revolving account where a customer is both paying for previous orders and ordering new items.

Let the states, which describe the customers account be given by u= (b,n,i) where b is the balance outstanding, n is the number of periods since the last payment and i is any other relevant information. Suppose the action is which credit limit, L, to set and we assume the performance of the account may be affected by the credit limit set.. It is necessary to estimate $p^L(u,u')$ and $r^L(u)$, which are the probability of the account moving from state u to u' under a credit limit L in the next period and the chance the reward obtained in that period is $r^L(u)$. These can be obtained by estimating $t^L(u,a)$, the probability that an account in state u with credit limit L repays a next period;

$q^L(u,o)$, the probability that an account in state u with credit limit L orders o next period;

$w^L(u,i')$, the probability that an account in state u with credit limit L changes its information state to i' and defining transition probabilities by

$$p^L(b,n,i;\ b+o-a,0,i') = t^L(u,a)q^L(u,o)w^L(u,i'),\ \text{provided } b+o-a \leq L,\ \text{and } a>0.$$

$$p^L(b,n,i;\ b-a,0,i') = t^L(u,a)w^L(u,i')(\ q^L(u,0)+\sum_{0}\ o.L-b+a\ q^L(u,o)),\ \text{where } a>0.$$

$$p^L(b,n,i;\ b+o,n+1,i') = t^L(u,0)q^L(u,o)w^L(u,i'),\ \text{provided } b+o \leq L.$$

$$p^L(b,n,i;\ b-a,n+10,i') = t^L(u,0)\ wL\ (u,i')(\ q^L(u,0)+\sum_{0}\ o.L-b+a\ q^L(u,o)).$$

If f is the fraction of a purchase that is profit for the company and the company has a policy of writing

off bad debt after N periods of non-payment that the reward function would be

$$r^L(b,n,i) = f \sum_o oq^L(u,o) - bt^L(u,0)\delta(n-(N-1))$$

One can then use dynamic programming to find Vn (u) the expected profit over n periods given the account is in state u and the optimal credit limit policy by solving the optimality equation

$$V_n(u) = max_L\{ r^L(u) + u \sum_{u'} p^L(u,u')V_{n-1}(u')\}$$

The first published account of this type of model was by Cyert, Davidson and Thompson (Cyert *et al.* 1962), where the units were dollars not accounts and the state was how overdue the account was. Their approach had some difficulties with accounting conventions – an account with £10 three months overdue and £10 one month overdue would become four months overdue if only £10 is paid in the next month. This pioneering paper was followed by several which modified the basic model. Kuelen (1981) suggested a modification of the approach that overcame the difficulty with defining partial payments of overdue accounts while Corcoran (1978) pointed out that the system would be even more stable if different transition matrices were used for accounts of different characteristics such as size of the accounts. The question on how many segments of the population should have different scoring systems is important in credit scoring as well as behavioural scoring. Banasik *et al.* (1996) point out that segmentation does not always give an improved scorecard in practice, if the segments are not distinctive enough.

An alternative Bayesian based probability model was pioneered by Bierman and Hausman (Bierman Hausman 1970). In this the probability of paying was not given from a sample of previous customers but was taken to be a Bernoulli random variable whose parameter satisfied a Beta distribution. The parameters of the Beta distribution were updated by the payment

performance of the individual customer, so if initially they were (r,n) than after n' payments periods in which the customer paid r' times they became (r+r', n+n'). The authors assumed that once credit had been refused no more credit

was granted, unlike the model described earlier in this section. Dirickx and Wakeman (1976) relaxed this assumption while Srinivasan and Kim (1987) allowed the simple extension of payments and orders being possible in the same period. Thomas (1994) extended the model by allowing not only the probability of repayment but also the maximum affordable repayment amount to be random variables which are updated in a Bayesian fashion according to the amount of repayments made.

2.5 Survival analysis approach to profit scoring

The Markov chain models describe the dynamics of a consumers movement through a number of delinquency states or scoring bands. If one is only interested in when they reach the default state and not their intermediate behaviour then one can use survival analysis approaches to estimate when this will occur. So instead of just asking which consumers will default as in behavioural scoring one asks when will they default. Using survival analysis to answer the "when" question has several advantages namely:

i. it deals easily with censored data, where customers cease to be borrowers (either by

paying back the loan, death, changing lender) before they default

ii. it avoids the instability caused by having to choose a fixed period to measure satisfactory performance which is inherent in behavioural and credit scoring

iii. estimating when default occurs is a major step towards calculating the profitability of an applicant

iv. it makes it easier to incorporate estimates of changes in the economic climate into the 'scoring' system.

Narain (1992) was one of the first to suggest that survival analysis could be used in credit scoring. Banasik et.al (1999) compared the survival analysis approach with logistic regression based scorecards and showed how competing risks can be used in the credit scoring context. Stepanova and Thomas (1992) and Hand and Kelley (1993) developed the ideas further and introduced tools for building survival analysis scorecards as well as introducing survival analysis ideas into behavioural scoring.

If T is the time until a loan defaults then there are three standard ways of describing the randomness of T in survival analysis :

survival function $S(t) = Prob\{T \geq t\}$ where $F(t) = 1\text{-}S(t)$ is the distribution function density function $f(t)$ where $Prob\{t \leq T \leq t+\square\ t\} = f(t)\delta t$

hazard function $h(t) = f(t)/S(t)$ so $h(t)\delta t = Prob\{\{t \leq T \leq t+\delta t | T \geq t\}$

In the survival analysis approach, we want models, which allow the application and behavioural characteristics to affect the probability of when a customer defaults. Two models connect the explanatory variables to failure times in survival analysis – proportional hazard models and accelerated life models. If $\boldsymbol{x} = (x_1, \ldots . x_p)$ are the explanatory characteristics, then an accelerated life model assumes

$$S(t) = S_0(e^{w x z} t) \text{ or } h(t) = e^{w.x} h_0(e^{w.x} t)$$

where h_0 and S_0 are baseline functions so the **x** can speed up or slow down the 'ageing' of the account. The proportional hazard models assume

$$h(t) = e^{w.x} h_0(t)$$

so the characteristics **x** have a multiplier effect on the baseline hazard. One can use a parametric approach to both the proportional hazards and acceleration life models by assuming $h_0(.)$ belongs to a particular family of distributions. It turns out that the negative exponential and the Weibull distributions are the only ones that are both accelerated life and proportional hazard models. The difference between the models is that in proportional hazards the applicants most at risk of defaulting at any one time remain the ones most at risk of defaulting at any other time.

Cox (1972) pointed out that in proportional hazards one can estimate the weights **w** without knowing $h_0(t)$ using the ordering of the failure times and the censored times. If t_i , $\mathbf{x}_i$ are the failure (or censored) times and the application variables for each of the items under test, then the conditional probability that customer i defaults at time t_i given R(i) are the customers still operating just before t_i is given by:

$$\exp\{W.X_i\}h_0(t) / \sum_{k \in R_{(i)}} \exp\{W.X_i\} / \sum_{k \in R_{(i)}} \exp\{W.X_k\}$$

which is independent of h_0. This approach which does not prejudge the form of the baseline hazard function is the one that has been most closely explored in the credit context. One of the disadvantages of the proportional hazards assumption is that the relative ranking among the applicants of the risk (be it of default or early repayment) does not vary over time. This can be overcome by introducing time-dependent characteristics. So suppose $x_1=1$ if the purpose of the loan is refinancing and 0 otherwise. One can introduce a second characteristic $x_2=x_1t$. In one model (Stepanova and Thomas 2001) with just x_1 involved, the corresponding weight was $w_1=0.157$, so the hazard rate at time t for refinancing loans was $e^{0.157}h_0(t)=1.17h_0(t)$ and for other loans $h_0(t)$. When the analysis was done with both x_1 and x_2, the coefficients of the proportional hazard loans were $w_1=0.32$, $w_2=-0.01$. So for refinancing loans the hazard rate at time t was $e^{0.32-0.01t}h_0(t)$ compared with others $h_0(t)$. Thus in month 1, the hazard from having a refinancing loan

was e0.31=1.36 times higher than for a non- refinancing loan, while after months, the hazard rate for refinancing was $e^{-0.04} = 0.96$ of the hazard rate for not refinancing.

Thus time-by-characteristic interactions in proportional hazard models allow the flexibility that the effect of a characteristic can increase or decrease with the age of the loan.

Survival techniques can also be applied in the behavioural scoring context, though a little more care is needed. Suppose it is u periods since the start of the loan and $\underline{b}(u)$ are the behavioural characteristics in period u , then a proportional hazard model says the hazard rate for defaulting in another t periods time, i.e. t+u since the start of the loan, is

$e^{w(u).b(u)}h_0u(t)$. At the next period u+1, the comparable hazard rate would be that for t- more periods to go, i.e. $e^{w(u+1).b(u+1)}h_0{}^{u+1}{}_{(t-1)}$. Thus the coefficients w(u) have to be estimated separately for each period u, using only the data in the data set that has survived up to period u. As it stands these coefficients could change significantly from one period to the next. One way of smoothing out these changes would be to make the behavioural score at the last period, one of the characteristics for the current period. Another way is to fit a simple curve to explain the time variation in each coefficient $b_i(u)$ so in the linear case one seeks to fit $b_i(u)$ by $a_i + b_iu$. Details of such an analysis can be found in Stepanova and Thomas (2001).

2.6 Default prediction studies

The literature about default prediction methodologies is substantial. Many authors during the last 40 years have examined several possible realistic alternatives to predict customers' default or business failure. The seminal works in this field were Beaver (1967) and Altman (1968), who developed univariate and multivariate models to predict business failures using a set of financial

ratios. Beaver (1967) used a dichotomous classification test to determine the error rates a potential creditor would experience if he classified firms on the basis of individual financial ratios as failed or non-failed. He used a matched sample consisting of 158 firms (79 failed and 79 non-failed) and he analyzed 14 financial ratios. Altman (1968) used a multiple discriminant analysis technique (MDA) to solve the inconsistency problem linked to the Beaver's univariate analysis and to assess a more complete financial profile of firms. His analysis drew on a matched sample containing 66 manufacturing firms (33 failed and 33 non-failed) that filed a bankruptcy petition during the period 1946-1965. Altman examined 22 potentially helpful financial ratios and ended up selecting five as providing in combination the best overall prediction of corporate bankruptcy9. The variables were classified into five standard ratios categories, including liquidity, profitability, leverage, solvency and activity ratios.

MDA is based on two restrictive assumptions: 1) the independent variables included in the model are multivariate normally distributed; 2) the group dispersion matrices (or variance-covariance matrices) are equal across the failing and the non-failing group. See Barnes (1982), Karels and Prakash (1987) and McLeay and Omar (2000) for further discussions about this topic. Zmijewski (1984) was the pioneer in applying probit analysis to predict default, but, until now, logit analysis has given better results in this field.

For many years thereafter, MDA was the prevalent statistical technique applied to the default prediction models. It was used by many authors (Deakin (1972), Edmister (1972), Blum (1974), Eisenbeis (1977), Taffler and Tisshaw (1977), Altman *et al.*. (1977), Bilderbeek (1979), Micha (1984), Gombola *et al.*. (1987), Lussier (1995), Altman *et al.*. (1995)). However, in most of these studies, authors pointed out that two basic assumptions of MDA are often violated when applied to the default prediction problems10. Moreover, in MDA models, the standardized coefficients cannot be interpreted like the slopes of a regression equation and hence do not indicate the relative importance of the different variables. Considering these MDA's problems, Ohlson (1980), for the first time, applied the conditional logit model to the default prediction's study11. The practical benefits of the logit methodology are that it does not require the restrictive assumptions of MDA and allows working with disproportional samples. Ohlson used a data set with 105 bankrupt firms and 2,058 non-bankrupt firms gathered from the COMPUSTAT database over the period 1970-1976. He based the analysis on nine predictors (7 financial ratios and 2 binary variables), mainly because they appeared to be the ones most frequently mentioned in the literature. The model's performance, in terms of classification accuracy, was lower than

that reported in the previous studies based on MDA (Altman, 1968 and Altman *et al.*., 1977). But reasons were provided as to why logistic analysis was preferable.

From a statistical point of view, logit regression seems to fit well the characteristics of the default prediction problem, where the dependant variable is binary (default/non-default) and with the groups being discrete, non-overlapping and identifiable. The logit model yields a score between zero and one which critics of the logit technique, have pointed out the specific functional form of a logit regression can lead to bimodal (very low or very high) classification and probabilities of default.

conveniently gives the probability of default of the client12. Lastly, the estimated coefficients can be interpreted separately as the importance or significance of each of the independent variables in the explanation of the estimated PD. After the work of Ohlson (1980), most of the academic literature (Zavgren (1983), Gentry *et al.*. (1985), Keasey and Watson (1987), Aziz *et al.*. (1988), Platt and Platt (1990), Ooghe *et al.*. (1995), Mossman *et al.*. (1998), Charitou and Trigeorgis (2002), Lizal (2002), Becchetti and Sierra (2002)) used logit models to predict default. Despite the theoretic differences between MDA and logit analysis, studies (see for example Lo (1985)) show that empirical results are quite similar in terms of classification accuracy. Indeed, after careful consideration of the nature of the problems and of the purpose of this study, we have decided to choose the logistic regression as an appropriate statistical technique. For comparison purposes, however, we also analyze results using MDA.

Determining the probability of default, *PD*, in consumer credits, loans and credit cards is one of the main problems to be addressed by banks, savings banks, savings cooperatives and other

credit companies. This is a first step needed to compute the capital in risk of insolvency, when their clients do not pay their credits, which is called *default*. The risk coming from this type of situation is called *credit risk*, which has been the object of research since the middle of last century. The importance of credit risk, as part of financial risk analysis, comes from the New Basel Capital Accord (Basel II), published in 1999 and revised in 2004 by the Basel Committee for Banking Supervision (BCBS). This accord consists of three parts, called pillars. They constitute a universal theoretical framework for the procedures to be followed by credit companies in order to guarantee minimal capital requirements, called *statistical provisions for insolvency* (SPI).

Pillar I of the new accord establishes the parameters that play some role in the credit risk of a financial company. These are the probability of default, *PD*, the exposition after default, *EAD*, and the loss given default, *LGD*. The quantitative methods that financial entities can use are those used for computing credit risk parameters and, more specifically, for computing *PD*. These are the standard method and the internal ratings based method (IRB). Thus, credit companies can elaborate and use their own credit qualification models and, by means of them, conclude the Basel implementation process, with their own estimations of SPI.

There is an extensive literature on quantitative methods for credit risk, since the classical *Z*-score model introduced by Altman (1968). Nowadays there exist plenty of approaches and perspectives for modelling credit risk starting from *PD*. Most of them have provided better predictive powers and classification error rates than Altman's discriminant model, for credit solicitors (*application scoring*), as well as for those who are already clients of the bank

(*behavioural scoring*). This is the case of logistic regression models, artificial neural networks (*ANN*), support vector machines (*SVM*), as well as hybrid models, as mixtures of parametric models and *SVM*. For the reader interested in a more extended discussion on the evolution of these techniques over the past 30 years we mention the work by Altman and Saunders (1998), Saunders (1999), Crouhy *et al.*. (2000), Hand (2001), Hamerle *et al.*. (2003), Hanson and Schuermann (2004), Wang *et al.*. (2005), and Chen *et al.*. (2006).

The idea of using survival analysis techniques for constructing credit risk models is not new. It started with the paper by Narain (1992) and, later, was developed by Carling *et al.*. (1998), Stepanova and Thomas (2002), Roszbach (2003), Glennon and Nigro (2005), Allen and Rose (2006), Baba and Goko (2006), Malik and Thomas (2006) and Beran and Djaïdja (2007). A common feature of all these papers is that they use parametric or semiparametric regression techniques for modelling the time to default (*duration models*), including exponential models, Weibull models and Cox's proportional hazards models, which are very common in this literature. The model established for the time to default is then used for modelling *PD* or constructing the scoring discriminant function.

2.7 Conditional survival analysis in credit risk

The use of survival analysis techniques to study credit risk, and more particularly to model *PD*, can be motivated via Figure 2.1. It presents three common situations that may occur in practice when a credit company observes the "lifetime" of a credit. Let us consider the interval (0, 1) as the horizon of the study. Case (a) shows a credit with default before the endpoint of the time under study (*1*). In this case, the lifetime of the credit, *T*, which is the time to default of the

credit, is an observable variable. Cases (b) and (c) show two different situations. In both of them it is not possible to observe the time instant when a credit enters into default, which causes a lack of information coming

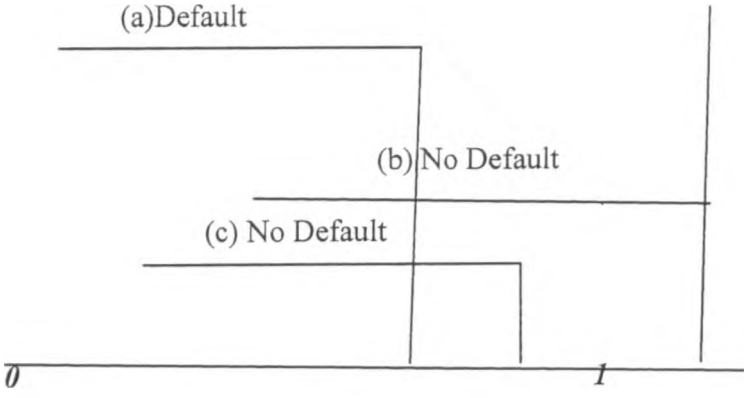


Figure2. 1: *Time to default in consumer credit risk.*

from right censoring. In case (b) it is only the time from the start of the credit to the end of the study, while (c) accounts for a situation where anticipated cancellation or the end of the credit occurs before default. The available information to model the *PD* is a sample of n iid random variables $\{(Y_1, X_1, \delta_1), \ldots, (Y_n, X_n, \delta_n)\}$, of the random vector $\{Y, X, \delta\}$, where $Y = \min\{T, C\}$ is the observed maturity, T is the time to default, C is the time to the end of the study or anticipated cancellation of the credit, $\delta = I(T \leq C)$ is the indicator of non censoring (default) and X is a vector of explanatory covariates. In this survival analysis setting we will assume that there exists an unknown relationship between T and X. We will also assume that the random variables T and C are conditionally independent given X.

In the previous setup it is possible to characterize completely the conditional distribution of the random variable T using some common relations in survival analysis. Thus the conditional survival function, $S(t|x)$, the conditional hazard rate, $\lambda(t|x)$, the conditional cumulative hazard

function, $\Lambda(t|x)$, and the conditional cumulative distribution function, $F(t|x)$, are related as follows:

$$S(t \setminus x) = P(T > t \setminus X = x) = \int_t^{\infty} f(u \setminus x)du$$

$$\lambda(t \setminus x) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \setminus T \ge t, X = x)}{\Delta t} = \frac{f(t \setminus x)}{S(t \setminus x)}$$

$$\Lambda(t \setminus x) = \int_0^t \lambda(u \setminus x)du = \frac{f(t \setminus x)}{S(t \setminus x)}du$$

$$S(t \setminus x) = e^{-\Lambda(t \setminus x)}$$

$$F(t \setminus x) = 1 - S(t \setminus x)$$

The conditional survival function used for modelling credit risk opens an interesting perspective to study default. Rather than looking at default or not, we look at the time to default, given credit information of clients (endogenous covariates) and considering the indicators for the economic cycle (exogenous covariates). Thus, the default risk is measured via the conditional distribution of the random variable time to default, T, given a vector of covariates, X. The variable T is not fully observable due to the censoring mechanism.

In practice, since the proportion of defaulted credits is small, the proportion of censored data is large, which may cause poor performance of the statistical methods. On the other hand, the sample size is typically very large. This alleviates somehow the problem of the large proportion of censoring.

In order to estimate empirically the conditional distribution function of the time to default, we use the generalized product-limit estimator by Beran (1981). This estimator has been extensively studied by Dabrowska (1987), Dabrowska (1989), Gonz´alez- Manteiga and Cadarso-Su´arez

(1994), Van Keilegom and Veraverbeke (1996), Iglesias- P´erez and Gonz´alez-Manteiga (1999), Li and Datta (2001), Van Keilegom *et al.*. (2001) and Li and Van Keilegom (2002), among other authors.

2.8 Probability of default in consumer portfolio

In the literature devoted to credit risk analysis there are not many publications on modelling the credit risk in consumer portfolios or personal credit portfolios. Most of the research deals with measuring credit risk by *PD* modelling in portfolios of small,medium and large companies, or even for financial companies. There exist, however, several exceptions. In the works by Carling et al.. (1998), Stepanova and Thomas (2002) and Malik and Thomas (2006), the lifetime of a credit is modelled with a semi-parametric regression model, more specifically with Cox's proportional hazards model.

In the following we present three different approaches to model the probability of default, *PD*, using conditional survival analysis. All the models are based on writing *PD* in terms of the conditional distribution function of the time to default. Thus *PD* can be estimated, using this formula, either by (i) Cox's proportional hazards model, where the estimator of the survival function is obtained by solving the partial likelihood equations in Cox's regression model, which gives $\widehat{PD}^{PHM}$, by (ii) a generalized linear model, with parameters estimated by the maximum likelihood method, which gives $\widehat{PD}^{GLM}$, or by (iii) using the nonparametric conditional distribution function estimator by Beran, which gives the nonparametric estimator of the default probability, $\widehat{PD}^{NPM}$.

2.9 Modelling the probability of default via the conditional distribution function

Following Basel II, credit scoring models are used to measure the probability of default in a time horizon $t+b$ from a maturity time, t. A typical value is $b = 12$ (in months). Thus, the following probability has to be computed:

$$\begin{aligned} PD(t \setminus x) &= P(t \le t < t+b \setminus T \ge t, X = x) \\ &= \frac{P(T < t+b \setminus X = x) - P(T \le t \setminus X = x)}{P(T \ge t \setminus X = x)} \\ &= \frac{F(t+b \setminus x) - F(t \setminus x)}{1 - F(t \setminus x)} = 1 - \frac{S(t+b \setminus x)}{S(t \setminus x)} \end{aligned} \qquad 2.1$$

Where t is the observed maturity for the credit and x is the value of the covariate vector, X, for that credit.

2.10 Proportional hazards model

In this section, a semiparametric approach to perform the study of PD is given. Here we use Cox's proportional hazards approach to model the conditional survival function $S(t|x)$. The key in this method rests on the estimation of the cumulative conditional hazard function, $\Lambda(t|x)$, using maximum likelihood.

We follow the idea introduced by Narain (1992) for the estimation of $S(t|x)$, but we apply it in the definition of PD, as we have stated above in formula (2.1). The objective is to build a conditional model for the individual $PD(t|x)$, which is defined in terms of $\Lambda(t|x)$. In order to describe $\widehat{PD}^{PHM}$, we define the following expressions relative to Cox's regression theory. The estimator of the conditional hazard rate function is defined as:

$$\hat{\lambda}(t \setminus x) = \hat{\lambda}_0(t) \exp(x^T \hat{\beta}),$$

where $\lambda_0(t)$ is an estimator of the hazard rate baseline function, and β^* is an estimator of the parameter vector. Thus, under the assumption of a proportional hazards model, *PD* is estimated by:

$$\widehat{PD}^{PHM}(t \setminus x) = \frac{\hat{F}_\beta(t+b \setminus x) - \hat{F}_\beta(t \setminus x)}{1 - \hat{F}_\beta(t \setminus x)} = 1 - \frac{\hat{S}_\beta(t+b \setminus x)}{\hat{S}_\beta(t \setminus x)},$$

where

$$1 - \hat{F}_\beta(t \setminus x) = \hat{S}_\beta(t \setminus x) = \exp(-\hat{\Lambda}(t \setminus x)$$

The estimation method for this model consists of two steps. In the first step the cumulative baseline hazard function, $\hat{\Lambda}_0(t)$, is estimated by:

$$\hat{\Lambda}_0(t) = \sum_{i=1}^{n} \frac{1\{Y_i \le t, \delta_i = 1\}}{\sum_{j=1}^{n} 1\{Y_j \ge Y_i\}},$$

then the parameter β is estimated by

$$\hat{\beta}^{PHM} = \arg\max_{\beta} L(\beta),$$

where the partial likelihood function is given by

$$L(\beta) = \prod_{i=1}^{n} \frac{\exp(x_i^T \beta)}{\left\{\sum_{j=1}^{n} 1_{Y_j > Y_i} \exp(x_j^T \beta)\right\}}$$

Thus, the conditional cumulative hazard function estimator is given by

$$\hat{\Lambda}(t \setminus x) = \int_0^t \hat{\lambda}(s \setminus x) ds = \exp(x^T \hat{\beta}^{PHM}) \hat{\Lambda}_0(t).$$

2.11 Generalized linear model

A generalized linear model can be assumed for the lifetime distribution:

$$P(T \le t \setminus X = x) = F_\theta(t \setminus x) = g(\theta_o + \theta_1 t + \theta^T x),$$

where $\theta = (\theta_2, \theta_3 ..., \theta_{p+1})^T$ is a p-dimensional vector and g is a known link function, like the logistic or the probit function. Thus, this model characterizes the conditional distribution of the lifetime of a credit, T, in terms of the unknown parameters. Once this parameters are estimated, an estimator of the conditional distribution function is obtained, $F_{\hat{\theta}}$, and, finally, an estimator of PD can be computed by plugging this estimator in equation (2.1), i.e.

$$\widehat{PD}^{GLM}(t \setminus x) = \frac{\hat{F}_\beta(t+b \setminus x) - \hat{F}_\beta(t \setminus x)}{1 - \hat{F}_\beta(t \setminus x)} = 1 - \frac{\hat{S}_\beta(t+b \setminus x)}{\hat{S}_\beta(t \setminus x)},$$

Where $\hat{\theta} = \hat{\theta}^{GLM}$ is the maximum likelihood estimator of the parameter vector.

Let us consider the one-dimensional covariate case. Then $\theta = \theta_2$ and the conditional distribution given by the model is $F(t \setminus x) = g(\theta_o + \theta_1 t + \theta_2 x)$, with density

$$f(t \setminus x) = \theta_1 g'(\theta_0 + \theta_1 t + \theta_2 x).$$

Since we are given a random right censored sample, the conditional likelihood function is a product of terms involving the conditional density, for the uncensored data, and the conditional survival function, for the censored data:

$$L(Y, X, \theta) = \prod_{i=1}^{n} f(Y_i \setminus X_i)^{\delta_i} (1 - F\{Y_i \setminus X_i\})^{1-\delta_i},$$

where Y_i is the maturity of the i-th credit and δ_i is the indicator of default for the i-th credit. Thus, the log-likelihood function is

$$\ell(\theta) = \ln\left(L(Y,X,\theta)\right) = \sum_{i=1}^{n}\left[\delta_i \ln(f(Y_i \setminus X_i)) + 1 - \delta_i) \ln(1 - F(Y_i \setminus X_i))\right]$$

$$= \sum_{i=1}^{n}\left[\delta_i \ln(\theta_1 g'(\theta_0 + \theta_1 Y_i = \theta_2 X_i)) + (1-\delta_i)\ln(1 - g(\theta_0 + \theta_1 Y_i + \theta_2 X_i))\right]$$

$$= \sum_{i=1}^{n}\delta_i\left[\ln(\theta_1) + \ln(g'(\theta_0 + \theta_1 Y_i + \theta_2 X_i))\right] + \sum_{i=1}^{n}(1-\delta_i)\ln(1 - g(\theta_0 + \theta_1 Y_i + \theta_2 X_i))$$

Finally, the estimator is found as the maximizer of the log-likelihood function:

$$\hat{\theta}^{GLM} = \arg\max_{\theta} \ell(\theta).$$

The works by Jorgensen (1983) and McCullagh and Nelder (1989) deal with generalized linear models in a regression context. These models can be adapted to the conditional distribution function setup.

2.12 Nonparametric conditional distribution estimator

First of all a nonparametric estimator of the conditional distribution function is obtained. This estimator, say $\hat{S}_h(t \setminus x)$, is used to derive an estimator of $PD(t|x)$, say $\widehat{PD}^{NPM}(t \setminus x)$, for the desired values of t and x.

Since we have a sample of right censored data for the lifetime distribution of a credit, we use the estimator proposed by Beran (1981) for the conditional survival function of T given $X = x$:

$$\hat{S}_h(t \setminus x) = \prod_{i=1}^{n}\left(1 - \frac{1_{\{Y_i \le t, \delta_i = 1\}} B_{ni}(x)}{1 - \sum_{j=1}^{n} 1_{\{Y_i < Y_j\}} B_{nj}(x)}\right),$$

where Y_i is the observed lifetime of the i-th credit, δ_i is the indicator of observing default of the i-th credit (uncensoring) and X_i is the vector of explanatory covariates for the i-th

credit. The terms $B_{ni}(x)$ are Nadaraya-Watson nonparametric weights:

$$B_{ni}(x) = \frac{K\left((x - X_i)/h\right)}{\sum_{j=1}^{n} K\left((x - X_i)/h\right)}, 1 \le i \le n,$$

and $h \equiv h_n$ is the smoothing parameter that tends to zero as the sample size tends to infinity.

To estimate the probability of default at time t given a covariate vector x, we replace, in (2.1), the theoretical value of the conditional survival function by its estimator $\hat{S}_h$:

$$\widehat{PD}^{NPM}(t \setminus x) = \frac{\hat{F}_h(t+b \setminus x) - \hat{F}_\beta(t \setminus x)}{1 - \hat{F}_h(t \setminus x)} = 1 - \frac{\hat{S}_h(t+b \setminus x)}{\hat{S}_h(t \setminus x)},$$

2.13 Linear Regression

Ordinary linear regression (Reg) is the simplest, compared all other techniques. Using the dummy variable for the dependent variable of good/bad indicator (say, define it as 1 if borrower is good, 0 if he/she is bad) and regressing it on a set of characteristics of borrowers by the standard least square approach will produce the estimated "probability" of being good. It is well known as the Linear Probability Model. Its main drawback is that there is no guarantee that the estimated probability would happen within the interval of [1, 0]. Orgler (1970) has applied regression analysis in a model for commercial loans and Orgler (1971) used it for evaluating existing loans.

2.14 Logistic Regression

Because of theoretical reasons, logistic regression (Logit) is a more appropriate statistical tool than linear regression if there are two discrete classes of the dependent variable. Logit approach tries to estimate the probability of a borrower being good as follows:

$$\Pr(good \setminus x) = \Pr(u = 1 \setminus x) = \frac{1}{1 + \exp(x\omega)}$$

where x is an input vector, ω is a vector of logistic parameters. The parameters are typically estimated by the maximum likelihood procedure. Because of its simplicity, logistic regression is now the most common approach for estimating default risk (Thomas *et al.*. 2002).

Wiginton (1980) was one of the first to publish credit scoring results using logistic regression; he compared it with discriminant analysis. Leonard (1993) also applied logistic regression to a commercial loan evaluation.

2.15 Discriminant Analysis

Discriminant analysis (DA) is a technique for first identifying the "best" characteristics of the debtors, known as discriminator variables, which provide the maximum discrimination between high and low default risk borrowers. Generally, assumption of multivariate normality of the variables is required. Durand (1941) considered the use of discriminant analysis for the scoring system. Another account of its application in credit scoring is given by Myers and Forgy (1963).

2.16 Decision Tree and Rule

The Decision Tree (D. tree), also known as the classification tree and recursive partitioning, tries to split the population into two sub-groups which are more homogeneous by making use of the possible characteristics of the debtors. It keeps applying this procedure until one has a number of groups identified as either good or bad debtors. Application of such a method in credit scoring is given in Makowski (1985) and Mehta (1968).

2.17 K-Nearest Neighbour Classifiers

K-Nearest neighbour (KNN) classifiers classify a data instance (i.e. borrower) by considering only the k-most similar data instances. The class label is then assigned according to the class of the majority of the k-nearest neighbours. To measure the distances among the data instances, it is common to choose the Euclidean distance, in which the characteristics of the borrower are taken into account. This approach was applied in the credit-scoring context by Chatterjee and Barcun (1970) and Henley and Hand (1996).

2.18 Bayesian Network Classifiers

Basically, with the class-conditional probabilities $\Pr(x_i \setminus u)$ of each input $x_i, i = 1,2,...,N$, given the class label u, a new case (i.e. new debtor) is then classified by using Bayes' rule to compute the posterior probability of each class u, given the vector of observed attribute value

$$\Pr(u \setminus x) = \frac{\Pr(x \setminus u)\Pr(u)}{\Pr(x)}$$

The assumption behind the method is that the attributes are conditionally independent, given the class label. Hence, $\Pr(u \setminus x) = \prod_{I=1}^{N} \Pr(x_i \setminus u)$.

Bayesian network classifier (B. net) is relatively rare in application to credit scoring, in the existing literature. Baesens *et al.*. (2003) made a careful study to compare different classifiers and found that this classifier is statistically (significantly) worse than the others (Table 2.1).

2.19 Linear Programming

Suppose there are N_G good and N_B bad borrowers and a set of q characteristics variables, so the borrower i has variable values $(x_{i1}, x_{i2}, ..., x_{iq})$. One seeks to develop a linear scorecard where all the good ones will have a value above the cut-off point τ and all the bad ones have a score below the cut-off. This cannot happen for all the cases; thus, new variable ε_i is introduced for allowing possible errors. Linear programming (LP) is a technique for finding the weights $(\kappa_1, \kappa_2, ...\kappa_q)$ that minimize the sum of the absolute values of the errors. That is,

Minimize $\varepsilon_1 + \varepsilon_2 + ... + \varepsilon_{N_G+N_B}$

Subject to $\kappa_1 x_{i1} + \kappa_2 x_{i2} + ... + \kappa_q x_{iq} \geq \tau - \varepsilon_i \qquad 1 \leq i \leq N_G$

$\kappa_1 x_{i1} + \kappa_2 x_{i2} + ... + \kappa_q x_{iq} \geq \tau - \varepsilon_i \qquad N_G + 1 \leq i \leq N_G + N_B$

$\varepsilon_{I \geq 0} \qquad 1 \leq i \leq N_G + N_B$

Hardy and Adrian (1985) compared linear programming with other statistical approaches and found that it classifies as well as them. Another famous application in credit scoring is given in Kolesar and Showers (1985).

2.20 Support Vector Machines

Support Vector Machine (SVM) is closely related to linear programming. The major difference is that it not only minimizes the value of errors, but also maximizes the marginal difference between the good ones and the bad ones. As a Bayesian network classifier, SVM is uncommon in credit scoring literature. Recently, Baesens *et al.*. (2003) found that SVM performs very well.

2.21 Neural Networks

Neural networks (NN) are mathematical representations inspired by the functioning of the human brain. As Hand and Henley (1997) stated, the type of NN that is normally applied to credit scoring problems can be viewed as a statistical model involving linear combinations of nested sequences of non-linear transformations of linear combinations of variables. In brief, NN can be considered as a form of non-linear regression. Rosenberg and Gleit (1994) described applications of NN to corporate credit decisions and fraud detection and Davies *et al.*. (1992) compared it with other scorecards.

2.22 Comparisons of Classification Models

A number of methods for developing credit scoring systems have been applied. The intuitive question is which one is the best. Several comparisons of the scorecards have been implemented

in the literature. Table I shows the results of four comparisons using credit-scoring data (Thomas *et al.*. 2002), in terms of percentage correctly classified, i.e. either good ones being classified as good or bad ones being classified as bad. The numbers should be compared across the rows but not between rows since they considered different data settings.

Table 2.3.1: Percentage correctly classified by different classification models

Author	**Reg.**	**Logit**	**D.Tree**	**LP**	**NN**
Henley (1995)	43.4	43.3	**43.8**	--	--
Boyle et.al.(1992)	**77.5**	--	75	74.7	--
Srinivasan and Kim (1987)	87.5	89.3	**93.2**	86.1	--
Desai *et al.*. (1997)	66.5	**67.3**	--	--	66.4

Source: Thomas *et al.*. (2002)

The highest for each row is denoted in bold face. In the studies of Henley (1995) and Srinivasan and Kim (1987), the Decision Tree is the best scorecard. Linear regression is the winner in the comparison of Boyle *et al.*. (1992), while logistic regression classifies the best in the paper by Desai *et al.*. (1997). It seems there is no uniformly best scorecard.

One may argue that the above comparisons are not reliable since each one considers only one data set. The best scorecard may have been so due to the pattern of its particular data set only. In response of this argument, Baesens *et al.*. (2003) evaluated a number of scorecards by eight real-life credit scoring data sets.

Table 2.3.2: Average ranking of different classification methods

Method	DA	Logit	LP	SVM	NN	B.net	D.Tree	KNN
Average ranking	6.9	6.1	6.5	**3.6**	5.2	15.1	6.7	7.9

Source: Baesens *et al.*. (2003)

In the Wilcoxon signed rank test, the authors tested the significant differences among the ranks. Although support vector machine is the best one, five other methods are not significantly different from it. They are linear and logistic regressions, linear programming, neural networks and decision tree. The Bayesian network classifier is statistically and significantly worse than the others. It was concluded by Baesens *et al.*. (2003) that with the statistical test, there is no uniformly best model in the credit scoring context, which agrees with the conclusion of Thomas *et al.*. (2002) as per Table 2.3.1.

2.23 Probability Density Function of Credit Losses

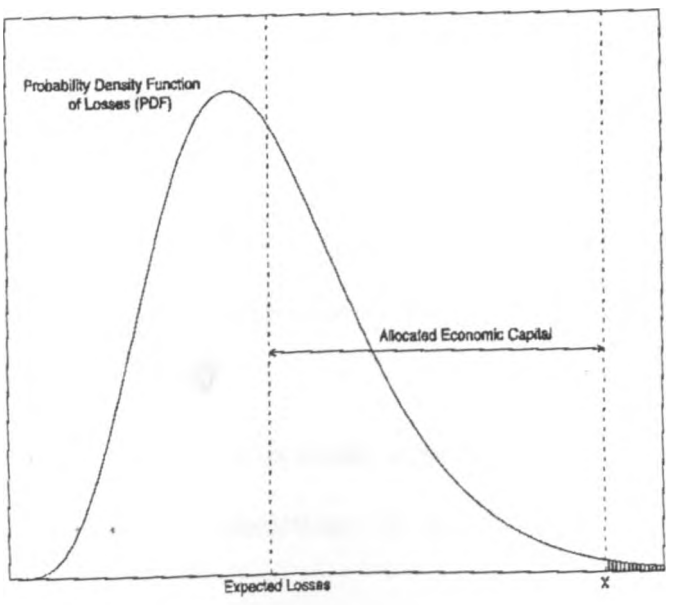
When estimating the amount of economic capital needed to support their credit risk activities, many large sophisticated banks employ an analytical framework that relates the overall required economic capital for credit risk to their portfolio's ***probability density function of credit tosses (PDF),*** which is the primary output of a credit risk model (Cooper and Martin 1996). Exhibit 1 illustrates this relationship. A bank would use its credit risk modelling system to estimate such a PDF. An important property of a PDF is that the probability of credit losses exceeding a given amount X (along the x-axis) is equal to the (shaded) area under the PDF to the right of X. A risky portfolio, loosely speaking, is one whose PDF has a relatively long and fat tail. The *expected credit loss* (shown as the left-most vertical line) shows the amount of credit loss the bank would

expect to experience on its credit portfolio over the chosen time horizon (Hurd 2007). Banks typically express the risk of the portfolio with a measure of *unexpected credit loss* (i.e. the amount by which actual losses exceed the expected loss) such as the standard deviation of losses or the difference between the expected loss and some selected target credit loss quantile.

The estimated economic capital needed to support a bank's credit risk, exposure is generally referred to as its required economic capital for credit risk. The process for determining this amount is analogous to *value at risk (VaR)* methods used in allocating economic capital against market risks.

Specifically, the economic capital for credit risk is determined so that the estimated probability of unexpected credit loss exhausting economic capital is less than some target insolvency rate. (Schoenfeld 1982).

Exhibit 1



Capital allocation systems generally assume that it is the role of reserving policies to cover expected credit losses, while it is that of economic capital to cover unexpected credit losses Thus, required economic capital is the additional amount of capital necessary to achieve the target insolvency rate, over and above needed for coverage of expected losses. In Exhibit 1, for a target insolvency rate equal to the shaded area, the required economic capital equals the distance between the two dotted lines.2 Broadly defined, a ***credit risk model*** encompasses all of the policies, procedures and practices used by a bank in estimating a credit portfolio's PDF (Collett 1994).

2.24 Survival Modelling

Because of the limitation of Classification Models, the credit scoring model is extended to estimate the time-to-default, instead of whether the borrower will default or not. As suggested by Banasik *et al.*. (1999), it has now become important to know not only if but also when the borrower would default. It is similar to the ideas of survival analysis in mortality and equipment reliability. Since Narain (1992) applied survival analysis for credit scoring, it has been widely investigated for credit risk management. Banasik *et al.*. (1999) analyze the time-to-default and time-to-early-repayment by semi-parametric proportional hazards model (Cox model) and two parametric proportional hazards models (with exponential and Weibull baseline hazards).

Stepanova and Thomas (2002) adopt the Cox model to personal loan data by coarse-classifying of characteristics and by including interactions of time-by-characteristics. Stepanova and

Thomas (2001) further develop survival analysis techniques in credit risk modelling by estimating the expected profit of personal loans. Most of them not only estimate the probability of default of the loan over time, but also classify the borrowers into either "good" or "bad". Concerning the accuracy of classification, the survival analysis is comparable with logistic regression, the most common approach for credit risk modelling (Thomas *et al.*. 2002)

2.25 Descriptive Methods of Time-to-event

Survival analysis is a statistical method for modelling the time to some events for a population of individuals. For example, events may refer to death in medical application, or recidivism of released prisoners in criminology application, or first bought of a new product by customer in marketing studies. The time to the occurrence is termed as survival time or lifetime. In application to credit risk modelling, the events refer to default of a loan and therefore its lifetime refers to time-to-default *T*.

Default times are subject to random variation and are thus random variables. To describe their randomness, there are five standard ways:

Density function (PDF), $f_{20}(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \le T \le t + \Delta t)}{\Delta t}$

Distribution function (CDF), $F_{20}(t) = \int_0^t f_{20}(u)du = \Pr(T \le t)$

Survivor function, $S_{20}(t) = \int_t^\infty f_{20}(u)du = 1 - F_{20}(t) = \Pr(T \ge t)$

Hazard function

$$h_{20}(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \le T \le t + \Delta t \setminus T \ge t}{\Delta t}$$

$$= \frac{f_{20}(t)}{S_{20}(t)} = \frac{-d \ln S_{20}(t)}{dt}$$

Cumulative hazard function $H_{20}(t) = \int_0^{\infty} h_{20}(t)dt = -\ln S_{20}(t)$

These five formulations are mathematically equivalent but they highlight different aspects of the default time. The distribution function tells us the probability that default occurs at or before time *t*. Conversely, survivor function is the probability that default does not occur at or before time *t*; in other words, the loan survives (non-default), at least, to time *t*. The interpretation of hazard function is slightly tricky. It is the "rate" that borrower defaults at time t, conditional on his staying on the books up to that time. Note that hazard is not a probability and thus can be greater than one.

In survival analysis, one must consider a key analytical problem called censoring. In essence, censoring occurs when we have some information about an individual's survival time, but do not know the exact survival time. There are a number of types of censoring, such as random, interval, left, and right censoring. In credit scoring application, most of the cases are right censoring.

For example, suppose we follow a group of borrowers for 3 years. If we observe borrower A fails to repay at 15th month, he is certainly classified as a default case and his default time is 15. On the other hand, consider borrower B, who repays on time during the whole observed period. We do not know his exact default time but are sure that it must be greater than 36. For such case,

borrower B is known as a right censored observation. Another example of right censoring could be when borrower C repays on time from the 1^{st} month to the 12^{th} month. At the 12^{th} month, we do not have future repayment pattern of borrower C. As borrower B, we do not know the exact default time of borrower C, we only know that it must be greater than 12. This is also a right censoring example.

2.26 General Issues in Credit Risk Modelling

The field of credit risk modelling has developed rapidly over the past few years to become a key component in the risk management systems at financial institutions. In fact, several financial institutions and consulting firms are actively marketing their credit risk models to other institutions. In essence, such models permit the user to measure the credit risk present in their asset portfolios. This information can be directly incorporated into many components of the user's credit portfolio management, such as pricing loans, setting concentration limits and measuring risk-adjusted profitability (Stepanova and Thomas 2002).

As summarized by the Federal Reserve System Task Force of USA on Internal Credit Risk Models (FRSTF, 1998) and the Basle Committee on Banking Supervision (BCBS, 1999), there exists a wide variety of credit risk models that differ in their fundamental assumptions, such as their definition of credit losses; i.e., default models define credit losses as loan defaults, while market-to-market or multi-state models define credit losses as rating migrations of any magnitude.

However, the common purpose of these models is to forecast the probability distribution function of losses that may arise from a bank's credit portfolio (Stepanova and Thomas 2002). Such loss distributions are generally not symmetric. Since credit defaults or rating changes are not common events and since debt instruments have set payments that cap- ture possible returns, the loss distribution is generally skewed toward zero with a long right-hand tail (Treacy and Carey 1998).

Although an institution may not use the entire loss distribution for decision-making purposes, credit risk models typically characterize the full distribution. A credit risk model's loss distribution is based on two components: the multivariate distribution of the credit losses on all the credits in its portfolio and a weighting vector that characterizes its holdings of these credits. This ability to measure credit risk clearly has the potential to greatly improve banks' risk management capabilities. With the forecasted credit loss distribution in hand, the user can decide how best to manage the credit risk in a portfolio, such as by setting aside the appropriate loan loss reserves or by selling loans to reduce risk. Such developments in credit risk management have led to suggestions, such as by ISDA (1998) and IIF (1998) that bank regulators permit, as an extension to risk-based capital standards, the use of credit risk models for determining the regulatory capital to be held against credit losses. Currently, under the Basle Capital Accord, regulated banks must hold 8% capital against their risk-weighted assets, where the weights are determined according to very broad criteria. For example, all corporate loans receive a 100% weight, such that banks must hold 8% capital against such loans. Proponents of credit risk models for regulatory capital purposes argue that the models could be used to create risk-weightings more closely aligned with actual credit risks and to capture the effects of portfolio

diversification. These models could then be used to set credit risk capital requirements in the same way that VaR models are used to set market risk capital requirements under the MRA.

However, as discussed by FRSTF (1998) and BCBS (1999), two sets of important issues must be addressed before credit risk models can be used in determining risk-based capital requirements. The first set of issues corresponds to the quality of the inputs to these models, such as accurately measuring the amount of exposure to any given credit and maintaining the internal consistency of the chosen credit rating standard. For example, Treacy and Carey (1998)

CHAPTER THREE

METHODOLOGY

3.1. Introduction

Researchers use the survival analysis in a variety of contexts that share a common characteristic: interest centres on describing whether or when events occur. It is necessary to use the survival analysis if we are interested in whether and when an event occurs (Allison, 1984). In this context the event occurrence represents a borrower's transition from one state, loan "in bonis" that is not in default, to another state, the default. To introduce survival approach to loans we assume that:

Assumption 1: a generation (or cohort) of loans is formed by loans granted by the banks in the same year;

Assumption 2: the death of the loan occurs with the default (the definition of default given by the Central Bank of Kenya is adopted in this study);

Assumption 3: the death of a loan is an uncertain event both "*when*" and "*if*";

Assumption 4: loan survival is the difference between two times: the time when a loan has been granted and the time when a loan becomes default;

Assumption 5: a loan is censored when, in the period of study (named follow-up), it is not in default or it goes out of the study to verify an event different of default. Thus, a loan is censored when: 1) a loan is *in bonis* so it is survived in all time, 2) the loan has been repaid.

3.2 Study Sample (Number at Risk)

The loan borrowers included in the study were randomly picked from a banks databank comprising 70 branches. The sample was based on personal loans whose maturity was 30 months. Thus the study cohort included loans taken in the month of January, 2007.

250 male applicants

250 female applicants

Window of observation was 30 months (January 1, 2007 to June 30, 2010).

Number who made early loan settlements:

Males: 19 customers

Females: 8 customers

Number who defaulted:

Males: 8

Females: 12

Times in (months) at which borrowers made early settlement of their loan accounts or defaulted were follows:

Early settlements:	Males: 3,3, 6,6,6,6,8,9, 12,12,12,15,15,18,18,18,18,24,25
	Females: 4, 4,12,18,18,20,20,22
Defaults:	Males: 4, 5, 5,7,10,10,13,16,
	Females: 2,3,3,7,11,11,16,21,21,25,25,27

3.3 Research Design

The data for analysis has been provided by the credit reference bureau based on the leading five commercial banks in Kenya. The focus is on personal loans whose maturity is three years and above. The performance of the accounts was observed during 36 months from January 2006 to January 2009. The study considered loans taken within the month of January 2006.

The life of the account is measured from the month it was opened until the account becomes 'bad' or it is closed or until the end of observation. The account is considered bad if payment is not made for two consecutive months in line with the industry practice. If the account does not miss two payments and is closed or survives beyond the observation period, it is considered to be censored. The data set consisted of the application information of 50 successful personal loan applicants randomly picked from the applications data for each stratum, together with the repayment status of each month of the observed period.

Table 3.1: Key application characteristics used by banks under study:

No.	Characteristic
1	Customer age
2	Years with current employer
3	Customer gender
4	Number of dependants
5	Marital status
6	Home ownership

This research made use of only one attribute due to its level and scope.

The research sample was stratified into two risk groups namely:

1. Males
0. Females

3.4 The Product - Limit Method

This function estimates survival rates and hazard from data that may be incomplete.

The survival rate is expressed as the survivor function (S):

$$S(t) = \frac{number\ of\ individuals\ surviving\ longer\ than\ t}{total\ number\ of\ individuals\ studied}$$

- where t is a time period known as the survival time, time to failure or time to event (such as death); e.g. 5 years in the context of 5 year survival rates. Some texts present S as the estimated probability of surviving to time t for those alive just before t multiplied by the proportion of subjects surviving to t.

This is univariate method which generates the characteristic "stair step" survival curves. It also called Kaplan-Meier estimator. The survival curves for the risk groups are compared using log-rank test which is a better measure than wilcoxon which places greater weights on events near time 0.

Hypothesis test:

H_0: the curves are statistically different

H_1: the curves are statistically the same. The test statistic is compared to χ^2 distribution.

Data was analysed using SPSS Version 10.

3.5 Model derivation

Suppose $t_{(1)} < t_{(2)} < ... < t_{(r)}$ are the ordered failure times. Let n_j denote the number of individuals alive (at risk of failure) just before time $t_{(j)}$, including those who will fail at time $t_{(j)}$. If an observation is censored at the same time $t_{(j)}$,that one or more failures occurs, the censoring is assumed to occur after any failures and n_j includes the censored observation s. Let d_j denote the number of failures at time $t_{(j)}$. The conditional probability that an individual fails in the time interval from $t_{(j)} - \Delta$ to $t_{(j)}$, given survival upto time $t_{(j)} - \Delta$,is estimated as

$$\frac{d_j}{n_j}$$

The conditional probability that an individual survives beyond $t_{(j)} - \Delta$, given survival upto time $t_{(j)} - \Delta$, is estimated as

$$\frac{n_j - d_j}{n_j}$$

In the limit as $\Delta \to 0$,

$$\frac{n_j - d_j}{n_j}$$

becomes an estimate of the conditional probability of surviving beyond $t_{(j)}$ given survival upto $t_{(j)}$.

For $t_{(k)} \le t < t_{(k+1)}$, the probability of surviving beyond time t is

$$S(t) = P\{T > t\} = P\{T > t \text{ and } T > t_{(k)}\}$$

$$
\begin{aligned}
&= P\left\{T > t \setminus T > t_{(k)}\right\}.P\left\{T > t_{(k)}\right\} \\
&= P\left\{T > t \setminus T > t_{(k)}\right\}.P\left\{T > t_{(k)} \setminus T > t_{(k-1)}\right\}.P\left\{T > t_{(k-1)}\right\} \\
&= P\left\{T > t \setminus T > t_{(k)}\right\}.P\left\{T > t_{(k)} \setminus T > t_{(k-1)}\right\}.P\left\{T > t_{(k-1)} \setminus t_{(k-2)}\right\}\cdots \\
&\quad ....P\left\{T > t_{(1)} \setminus T > t_{(0)}\right\}.P\left\{T > t_{(0)}\right\} \\
&\approx \prod_{j=1}^{k} P\left\{T > t_j \setminus T > t_{(j-1)}\right\}
\end{aligned}
$$

Where $t_{(0)} = 0$ and $t_{(r+1)} = \infty$.

The Product-limit estimator of the survivor function at time t for $t_{(k)} \le t < t_{(k+1)}$ is

$$\hat{S}(t) = \prod_{j=1}^{k} \frac{n_j - d_j}{n_j}$$

Also

$$\hat{S}(t) = \prod_{t_j \le t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{t_j \le t} \left(1 - \frac{d_j}{n_j} \right)$$

Variance of the Product-limit estimator (Greenwood's formula)

For $t_{(k)} \le t < t_{(k+1)}$,

$$\widehat{Var}\left(\hat{S}(T)\right) = \left(\hat{S}(t)\right)^2 \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}$$

Derivation

Recall

$$\hat{S}(t) = \prod_{j=1}^{k} \frac{n_j - d_j}{n_j}$$

$$\log\left(\hat{S}(t)\right) = \log\left(\prod_{j=1}^{k} \frac{n_j - d_j}{n_j}\right)$$

$$= \sum_{j=1}^{k} \log\left((n_j - d_j)/n_j\right)$$

$$= \sum_{j=1}^{k} \log(p_j)$$

$$Var\left(\log(\hat{S}(t)\right) = Var\left(\sum_{j=1}^{k} \log(p_j)\right)$$

$$= \sum_{j=1}^{k} Var\left(\log(p_j)\right)$$

Appling delta estimation technique,

$$Var\left(\log(p_j)\right) \approx \left(\frac{1}{\pi_j}\right)^2 \frac{\pi_j(1-\pi_j)}{n_j}$$

$$= \left(\frac{1}{\pi_j}\right)\frac{1-\pi_j}{n_j}$$

Thus,

$$Var\left(\log(\hat{S}(t)\right) \approx \sum_{j=1}^{k}\left(\frac{1}{\pi_j}\right)\frac{1-\pi_j}{n_j}$$

$$\Rightarrow Var\left(\hat{S}(t)\right) \approx \left[S(t)\right]^2 Var\left(\log(\hat{S}(t))\right)$$

$$= \left[\hat{S}(t)\right]^2 \sum_{j=1}^{k}\left(\frac{1}{\pi_j}\right)\frac{1-\pi_j}{n_j}$$

Substituting $p_j = (n_j - d_j)/n_j$ for π_j, is

$$\widehat{Var}\left(\hat{S}(t)\right) = \left(\hat{S}(t)\right)^2 \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}$$

Standard error of estimation (for large sample) is given by

$$se\left(\hat{S}(t)\right) = \hat{S}(t)\sqrt{\sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}}$$

Hazard function estimation for $t_{(j)} < t < t_{(j+1)}$ is given by

$$\hat{h}(t) = \frac{d_j}{n_j(t_{(j+1)} - t_{(j)})}$$

Hazard function estimation

Product- limit: For

$$t_{(j)} < t < t_{(j+1)}$$

$$\hat{h}(t) = \frac{d_j}{n_j(t_{(j+1)} - t_{(j)})}$$

Notation

d_j= number of defaults at time t_j which can take values 0,1,2,…,n

So d_j is a binomial distribution with parameters;

p_j = the probability of defaulting at time t_j

n_j = number of borrowers at risk just before time t_j

In terms of this study, the data set can be represented as follows:

Table 3.2: Raw Data for modelling

Females				Males			
t_i	d_i	n_i	c_i	t_i	d_i	n_i	c_i
0	0	250	0	0	0	250	0
2	1	250	0	3	0	250	2
3	2	248	0	4	1	247	0
4	0	245	2	5	2	243	0
7	1	241	0	6	0	238	4
11	2	234	0	7	1	236	0
12	0	223	1	8	0	229	1
16	1	211	0	9	0	221	1
18	0	195	2	10	2	212	0
20	0	177	2	12	0	202	3
21	2	157	0	13	1	190	0
22	0	136	1	15	0	177	2
25	2	114	0	16	1	162	0
27	1	89	0	18	0	146	4
				24	0	128	1
				25	0	104	1

CHAPTER FOUR

RESEARCH FINDINGS AND INTERPRETATION OF RESULTS

4.1 Introduction

In this chapter we give a summary of research findings and explanations of the results in the context of the research area. Details of the model output are at the appendix

4.2 The Research Findings

The model outputs were as follows:

For gender factor 0 (female):

Table 4.1: Survival Analysis for female

Number of Cases: 13 Censored: 6 (46.15%) Events: 7

	Survival Time	Standard Error	95% Confidence Interval
Mean:	16	3	(10, 23)
Median:	13	4	(6, 20)

	Percentiles		
	25.00	50.00	75.00
Value	27.00	13.00	10.00
Standard Error	.	3.73	2.50

For gender factor 1(males):

Table 4.2: Survival analysis for male

Number of Cases: 15 Censored: 4 (26.67%) Events: 11

	Survival Time	Standard Error	95% Confidence Interval
Mean:	15	2	(11, 20)
Median:	16	5	(6, 26)

	Percentiles		
	25.00	50.00	75.00
Value	25.00	16.00	7.00
Standard Error	.	4.92	3.04

Gender 1 refers to male borrowers and 0 refers to female borrowers. Event 1 denotes loan default and 0 denotes censored state.

The survival data output on gender 1 implies that out of the 250 male loan applicants for loans maturing in 30 months, 11 defaulted and 4 settled their loan accounts before maturity. Mean survival time of 15 means that on average, a male applicant will take 15 months to default. Same interpretation can be attributed to data on female applicants. The following summary was also generated:

Table 4.3: combined output
Survival Analysis for TIME

		Total	Number Events	Number Censored	Percent Censored
GENDER	0	13	7	6	46.15
GENDER	1	15	11	4	26.67
Overall		28	18	10	35.71

Test Statistics for Equality of Survival Distributions for GENDER

	Statistic	df	Significance
Log Rank	.17	1	.6780

By observation of survival curves, it can be seen that the two curves are similar. This is confirmed by the test statistic (log rank 0.17), which shows that the two survival distributions are statistically the same. Thus it is not meaningful to classify borrowers on the basis of gender.

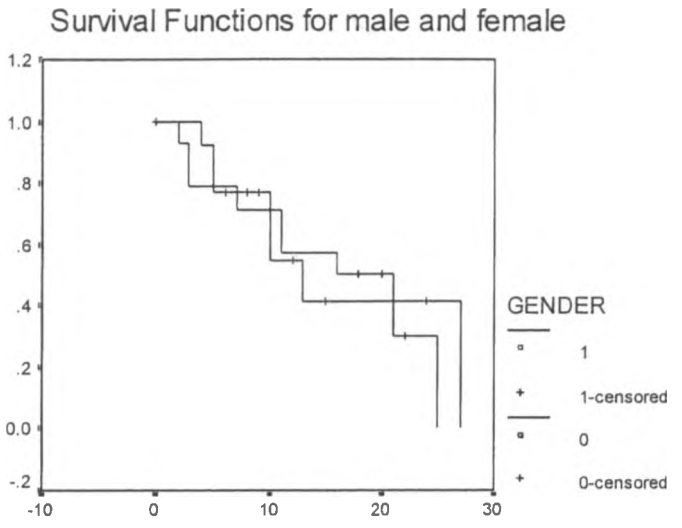


Fig.4.1

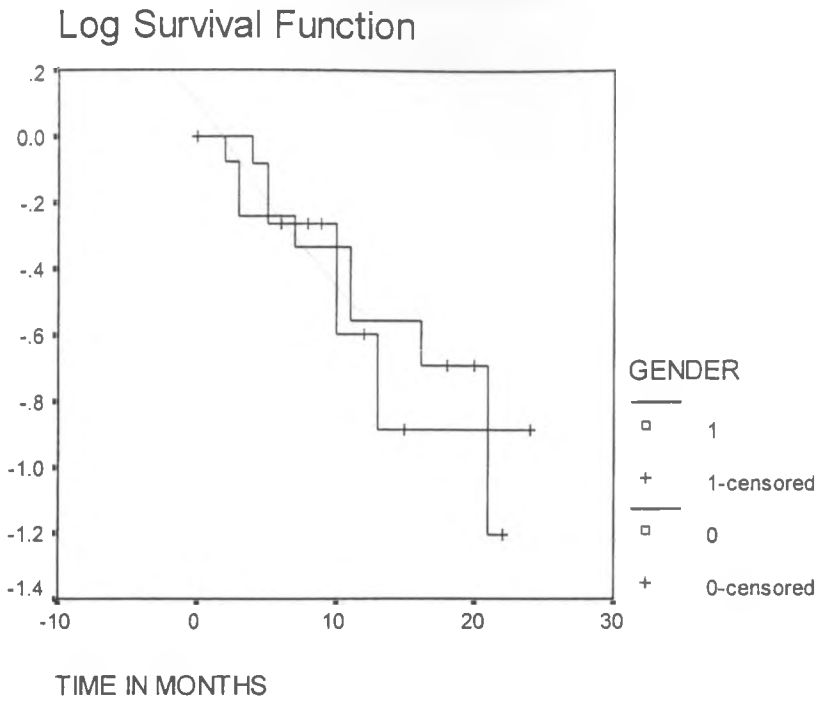

Log Survival Function
.2
0.0
-.2
-.4
-.6
-.8
-1.0
-1.2
-1.4
-10
0
10
20
30
TIME IN MONTHS
GENDER
1
1-censored
0
0-censored

Fig.4.2

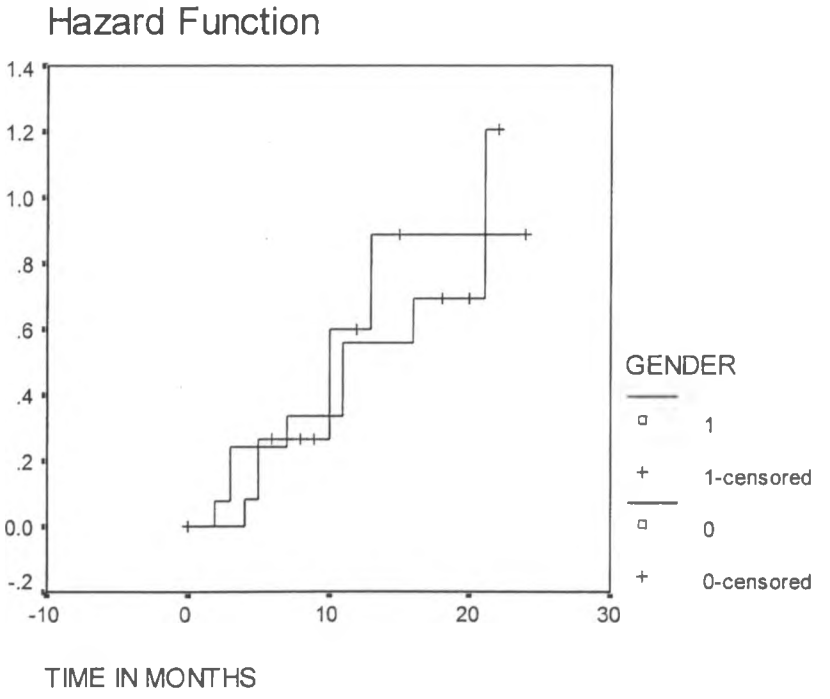

Hazard Function
1.4
1.2
1.0
.8
.6
.4
.2
0.0
-.2
-10
0
10
20
30
TIME IN MONTHS
GENDER
1
1-censored
0
0-censored

Fig.4.3

The survival curves generated also give the same indication that there is no significant difference in the survival curves for male and female borrowers.

CHAPTER FIVE

CONCLUSIONS AND RECOMMENDATIONS

5.1 Introduction

This chapter gives the position arrived at based on the research outcome and relates the outcome to the world of practice. Opportunity for further research is also proposed.

5.2 Conclusions

The research findings show that there is no significant difference between male and female borrowers in terms of their time to default on loan obligations. This implies that gender does not affect credit risk. Mean survival times would guide credit granting process on the average maturity for loans that may minimize on default losses and optimize profitability.

5.3 Recommendations and Suggestions for Further Research

This method of credit risk modelling is quite reliable as it does not make assumptions about loan default distribution unlike parametric methods. However, given that product-limit is a univariate method, it may be more informative to adopt multivariate techniques like Cox model to model credit risk. Thus further research can be conducted on the same data set using other survival techniques.

REFERENCES

Agresti, A. 2002. *Categorical Data Analysis*. Wiley, New York.

Allison, P. D. 1999. *Logistic Regression Using SAS System : Theory and Application*. SAS Institute Inc., Cary, NC.

Baesens, B., T. Van Gestel, M. Stepanova, D. Van den Poel, J. Vanthienen. 2005. Neural network survival analysis for personal loan data. *Journal of the Operational Research Society* **56** 1089-1098.

Baesens, B., T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, J. Vanthienen. 2003. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society* **54** 627-635.

Bamber, D. 1975. The area above the ordinal dominance graph and the area below the receiver operating graph. *Journal of Mathematical Psychology* **12** 387-415.

Banasik, J., J. N. Crook, L. C. Thomas. 1999. Not if but when will borrowers default. *Journal of the Operational Research Society* **50**(12) 1185-1190.

Boyle, M., J. N. Crook, R. Hamilton, L. C. Thomas. 1992. Methods for credit scoring applied to slow payers. L. C. Thomas, J. N. Crook, D. B. Edelman, ed. *Credit Scoring and Credit Control*. Oxford Unversity Press: Oxford, 75-90.

Chatterjee, S., S. Barcun. 1970. A nonparametric approach to credit screening. *Journal of the American Statistical Association* **65** 150-154.

Copas, J. B., F. Heydari. 1997. Estimating the risk of reoffending by using Exponential mixture models. *Journal of the Royal Statistical Society Series. A* **160** 237-252.

Cox, D. R. 1975. Partial Likelihood. *Biometrika* **62**(2) 269-276.

D'Agostino, R. B., B. H. Nam. 2004. Evaluation of the performance of survival analysis models: discrimination and calibration measures. N. Balakrishnan, C. R. Rao, ed. *Advances in Survival Analysis*. Elsevier, London, 1-25.

Davis, R. H., D. B. Edelman, A. J. Gammerman. 1992. Machine-learning algorithms for credit-card applications *IMA Journal of Management Mathematics* **4**(1) 43-51.

Desai, V. S., D. G. Conway, J. N. Crook, G. A. Overstreet, Jr. 1997. Credit-scoring models in the credit-union environment using neural networks and genetic algorithms. *IMA Journal of Management Mathematics* **8**(4) 323-346.

Durand, D. 1941. *Risk Elements in Consumer Instalment Financing*. National Bureau of Economic Research, New York.

Farewell, V. T. 1986. Mixture models in survival analysis: are they worth the risk? *The Canadian Journal of Statistics* **14**(3) 257-262.

Gamel, J. W., I. W. McLean, S. H. Rosenberg. 1990. Proportion cured and mean log survival time as functions of tumour size. *Statistics in Medicine* **9** 999-1006.

Hand, D. J., W. E. Henley. 1997. Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society Series. A* **160** 523-541.

Hanley, J. A., B. J. McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143** 29-36.

Hardy, W. E., Jr., J. L. Adrian, Jr. 1985. A linear programming alternative to discriminant analysis in credit scoring. *Agribusiness* **1**(4) 285-292.

Henley, W. E. 1995. *Statistical Aspects of Credit Scoring*. Ph.D. thesis. Open

83 University, Milton Keynes, U.K.

Henley, W. E., D. J. Hand. 1996. A k-nearest-neighbour classifier for assessing consumer credit risk. *Statistician* **65** 77-95.

Henley, W. E., D. J. Hand. 1997. Construction of a k-nearest-neighbour credit-scoring system. *IMA Journal of Management Mathematics* **8** 305-321.

Kolesar, P., J. L. Showers. 1985. A robust credit screening model using categorical data. *Management Science* **31**(2) 123-133.

Kuk, A. Y. C., C. H. Chen. 1992. A mixture model combining logistic regression with proportional hazards regression. *Biometrika* **79**(3) 531-541.

Larson, M. G., G. E. Dinse. 1985. A mixture model for the regression analysis of competing risks data. *Applied Statistics* **34** 201-211.

Leonard, K. J. 1993. Empirical Bayes analysis of the commercial loan evaluation process. *Statistics & Probability Letters* **18**(4) 289-296.

Makowski, P. 1985. Credit scoring branches out. *Credit World* **75** 30-37.

Maller, R. A., S. Zhou. 1994. Testing for sufficient follow-up and outliers in survival data. *Journal of the American Statistical Association* **89** 1499-1506.

Maller, R. A., X. Zhou. 1996. *Survival Analysis with Long-Term Survivors*. Wiley, New York.

Mehta, D. 1968. The formulation of credit policy models. *Management Science* **15**(2) 30-50.

Myers, J. H., E. W. Forgy. 1963. The development of numerical credit evaluation systems. *Journal of the American Statistical Association* **58** 799-806.

Narain, B. 1992. Survival analysis and the credit granting decision. L. C. Thomas, J. N. Crook, D. B. Edelman, ed. *Credit Scoring and Credit Control*. Oxford 84 Unversity Press: Oxford, 109-121.

Orgler, Y. E. 1970. A credit scoring model for commercial loans. *Journal of Money, Credit & Banking* **2**(4) 435-445.

Orgler, Y. E. 1971. Evaluation of consumer loans with credit scoring models. *Journal of Bank Research* 31-37.

Rosenberg, E., A. Gleit. 1994. Quantitative methods in credit management: a survey. *Operations Research* **42**(4) 589-613.

Srinivasan, V., Y. H. Kim. 1987. Credit granting: a comparative analysis of classification procedures. *Journal of Finance* **42**(3) 665-681.

Stepanova, M., L. Thomas. 2002. Survival analysis methods for personal loan data. *Operations Research* **50**(2) 277-289.

Stepanova, M., L. C. Thomas. 2001. PHAB scores: proportional hazards analysis behavioural scores. *Journal of the Operational Research Society* **52**(9) 1007-1016.

Thomas, L. C., D. B. Edelman, J. N. Crook. 2002. *Credit Scoring and Its Applications*. Society for Industrial and Applied Mathematics, Philadelphia.

Thomas, L. C., R. W. Oliver, D. J. Hand. 2005. A survey of the issues in consumer credit modelling research. *Journal of the Operational Research Society* **56**(9) 1006-1015.

Viaene, S., R. A. Derrig, B. Baesens, G. Dedene. 2002. A comparison of state-of-art classification techniques for expert automobile insurance claim fraud detection. *The Journal of Risk and Insurance* **69** 373-421.

Wiginton, J. C. 1980. A note on the comparison of logit and discriminant models of consumer credit behavior. *Journal of Financial and Quantitative Analysis* 85 **15**(3) 757-770.

Yamaguchi, K. 1992. Accelerated failure-time regression models with a regression model of surviving fraction: An application to the analysis of "Permanent Employment" in Japan. *Journal of the American Statistical Association* **87**(418) 284-292.

Kaplan-Meier

Survival Analysis for TIME

Factor GENDER = 0

Time	Status	Cumulative Survival	Standard Error	Cumulative Events	Number Remaining
4	1	.9231	.0739	1	12
5	1			2	11
5	1	.7692	.1169	3	10
6	0			3	9
8	0			3	8
9	0			3	7
10	1			4	6
10	1	.5495	.1556	5	5
12	0			5	4
13	1	.4121	.1667	6	3
15	0			6	2
24	0			6	1
27	1	.0000	.0000	7	0

Number of Cases: 13 Censored: 6 (46.15%) Events: 7

	Survival Time	Standard Error	95% Confidence Interval	
Mean:	16	3	(10,	23)
Median:	13	4	(6,	20)

Survival Analysis for TIME

Factor GENDER = 1

Time	Status	Cumulative Survival	Standard Error	Cumulative Events	Number Remaining
0	0			0	14
2	1	.9286	.0688	1	13
3	1			2	12
3	1	.7857	.1097	3	11
7	1	.7143	.1207	4	10
11	1			5	9
11	1	.5714	.1323	6	8
16	1	.5000	.1336	7	7
18	0			7	6
20	0			7	5
21	1			8	4
21	1	.3000	.1358	9	3
22	0			9	2
25	1			10	1
25	1	.0000	.0000	11	0

Number of Cases: 15 Censored: 4 (26.67%) Events: 11

	Survival Time	Standard Error	95% Confidence Interval	
Mean:	15	2	(11,	20)
Median:	16	5	(6,	26)

Survival Analysis for TIME

		Total	Number Events	Number Censored	Percent Censored
GENDER	0	13	7	6	46.15
GENDER	1	15	11	4	26.67
Overall		28	18	10	35.71

Test Statistics for Equality of Survival Distributions for GENDER

	Statistic	df	Significance
Log Rank	.17	1	.6780