EXPLORATORY LOAN DATA ANALYSIS & MODELLING TIME TO DEFAULT USING SURVIVAL ANALYSIS TECHNIQUES

NOVAH BERYL OULA

I56/69227/2011

School of Mathematics

University of Nairobi

A research project submitted in partial fulfillment of the requirement for the degree of Master of Science in Social Statistics of the University of Nairobi.

JULY 2013

DECLARATION

I declare that ***Exploratory Loan Data Analysis & Modelling Time to Default Using Survival Analysis Techniques*** is my own original work. This research project has never been presented for examination at any of the learning institution/University whether in Kenya or elsewhere as per my own knowledge and understanding. All the sources quoted have been indicated and acknowledged with complete reference.

STUDENT

NOVAH BERYL OULA

SIGN: ---------- DATE: ----------

This project has been submitted for examination with the approval of the following as University supervisor.

SUPERVISOR

Prof. JAM OTTIENO

SIGN: ---------- DATE: ----------

ACKNOWLEDGEMENT

All thanks to the almighty God for this far he has brought me with my studies. Special thanks in particular to my supervisor Prof. JAM OTTIENO, School of Mathematics for his priceless support without which this project would not have been successful. I also wish to thank DR. NELSON OWUOR for regularly assisting whenever there was need especially on Survival analysis approaches using SAS statistical analysis software and the entire School of Mathematics fraternity for any support accorded to me. May GOD bless you all.

DEDICATION

This project is with love dedicated to my daughter Nana Gracie, husband Abill Nerry for their immeasurable support and sacrifice during the entire period of my studies. Special thanks to my husband Abill Nerry for the consistent support he accorded me. May God abundantly bless you.

EXECUTIVE SUMMARY

The financial sector in Kenya has recorded double-digit growth in profits for most of the past decade, with the loans portfolio recording the highest growth while the economic growth has averaged at about 5%. Of particular concern is that the banking sector has been growing faster than the rest of the economy and would result in institutions and households that are not able to repay their debts leading to the increase of non-performing loans and as a result banks would be required to hold higher capital buffers to absorb possible shocks considering scenario like the global financial crisis experienced in late 2008. Banks are required to set aside some amounts for the non-performing loans and this impact on the profits as it is a deductable expense. This therefore means that loans portfolio should be effectively managed to ensure that credit risk is at manageable levels. Effective management of the growing portfolio requires frequent review of the credit granting process to ensure that only credit worthy individuals are granted loans.

Banks have traditionally employed the use of credit scoring to differentiate 'bad' customers from 'good' customers in their credit granting process however the idea of markov chain where borrowers' move from one credit state to another brings to light that borrower's status is dynamic and not static. Credit scoring puts a static element to this dynamism and the study focus now is *not if but when will the borrowers default.*

With this identified dynamicity, lending institutions need to review their credit granting criteria to be robust so that they not only score for risk but for profitability. This would ensure they choose customers whose time to default is long hence resulting in maximized profits since interest charged will compensate or even exceed losses resulting from default.

This paper therefore explores the loan data and uses survival analysis techniques specifically the Kaplan Meier and Cox Proportional Hazard Model approach to model time to default using various borrowers' application characteristics that include gender, age, income, term of loan, income commitment and banking history. Both the log rank and Wilcoxon tests are used to assess whether there is difference in the survival curves of the categorical variables.

The explanatory variables found to be significant in the univariate analysis are then assessed for time dependency and a multivariate Cox PH model with time independent covariate fitted. The results showed that out of the 6 application variables, only income and banking history were significant. It was therefore

not meaningful to classify borrowers on the basis of their gender, age, term of loan and commitments to the bank as these application variables did not affect risk of default.

As customers move from low income (< KES 100,000) to high income (≥ KES 300,000), rate of default decreases by 51%, when all other variables are held constant. Customers with banking history <6 months experienced default that is 2.3 times higher than those who have banked >24 months. Customers with banking history of 6-12 months have a default rate that is 96% higher than those who have banked > 24 months.

Table of Contents

CHAPTER 1: GENERAL INTRODUCTION

1.1 BACKGROUND

Various research papers have been published on the use of statistical methods to model consumer credit risk. Banks in response have developed credit scoring models that differentiate good customers from bad customers hence aid in making decision of whether to grant credit to an applicant or not. The methodology used in credit scoring is to take a sample of borrowers; assign weight to the different borrowers' application characteristics and classify them into either 'goods' or 'bads' depending on their repayment performance (probability of default) over a given period of time. In particular, logistic regression has become a standard method for this task (Thomas et al 2002).

There are also other behavioural scoring systems that look at how likely a borrower with a given current performance pattern is likely to perform in a given period of time in the future. The idea of markov chain where borrowers move from one state to another has extensively been used in behavioural scoring and this approach brings to light that credit status of borrowers has a dynamic element and that borrowers default at different times during their credit history.

The traditional credit scoring techniques approach puts a static element to this dynamism since the methodology looks at the borrowers' status after a fixed period of time.

Since credit status has therefore been perceived to be dynamic and not static (Banasik, Crook, & Thomas, 1999) in ***Not if but when will borrowers default***, the issue of dynamicity has become a key research question with most studies focusing on not if but when will the borrowers default. Lending institutions would want to choose customers who would help them in maximizing their profits and this therefore means that if time to default is long, interest income will compensate or even exceed losses resulting from default.

Various advantages of studying time to default have already been highlighted (Thomas et al 1999) and these include:

(i) Estimates of when an applicant defaults will give a better view of the likely profitability of the applicant and hence is a first step on the road to profit scoring.

(ii) That such estimate will give a forecast of the default levels as a function of time. This would be useful for firms' debt provisioning.

(iii) The estimates may guide the decision making on how long a credit facility ought to be granted.

(iv) That such an approach may make it easier to incorporate estimates of future changes in economic environment and future default estimates can be obtained

Narain (1992) was the first author to employ use of survival analysis specifically the Kaplan-Meier method in building credit scoring models. This model was further developed by Thomas et al (1999) and showed that survival analysis can be used to estimate time to default and early repayment. Several other research works have been done and these will be discussed in the literature review section.

1.2 STATEMENT OF THE PROBLEM

Kenya's financial industry is currently one of the fastest growing not only in the East African region but in the continent. The banking sector is a very important sector to the Kenyan economy with key highlights of performance over the last 10 years (2002-2012) consolidating the importance of the banking as follows; Assets grew from Ksh.456.7 billion to Ksh.2.35 trillion; Total deposits grew from Ksh.360.6 billion to Ksh.1.76 trillion. ***Net advances increased from Ksh.222.8 billion to Ksh.1.27 trillion,*** Profit before tax of Ksh.5.8 billion increased to Ksh.107 billion. The number of bank accounts has increased from 1.9 million accounts to 17.6 million.

This phenomenal growth has been supported by the expansion of banks into new market segments, prudent risk management and enhanced economic prospects underpinned by a stable macroeconomic environment. Ongoing reforms and initiatives by the Government under the Vision 2030 agenda and Central Bank of Kenya (Credit Reference Bureau for credit information sharing, prudential guidelines and Risk Management guidelines) will serve to further propel the banking sector to new frontiers of financial inclusion for more Kenyans to access these services

The sector has recorded double-digit growth in profits for most of the past decade, when the economic growth has averaged at about five per cent and of particular concern is that a banking sector growing faster than the rest of the economy could result in institutions and households that are not able to repay their debts leading to growth of non-performing loans and as a result could require banks to hold higher capital buffers to absorb possible shocks considering scenario like the global financial crisis experienced in late 2008.

Credit risk is measured by the amounts banks set aside for the non-performing loans. This in financial terms is known as provisioning for bad debts. The provisioning impact on a bank's profits as it is a deductable expense. This therefore means that loans portfolio should be effectively managed to ensure that credit risk is at manageable levels. Effective management of the growing portfolio requires frequent review of the credit granting process to ensure that only credit worthy individuals are granted loans. Banks over the years have used the traditional credit scoring models based on logistic regression that differentiates bad borrowers from good borrowers over a given period of time and does not take into account how long it takes before the borrowers default since default has been confirmed to be a dynamic event and having a robust view of a customer would help in effective risk management.

This paper tests the hypothesis that probability of default is affected by various application characteristics of borrowers.

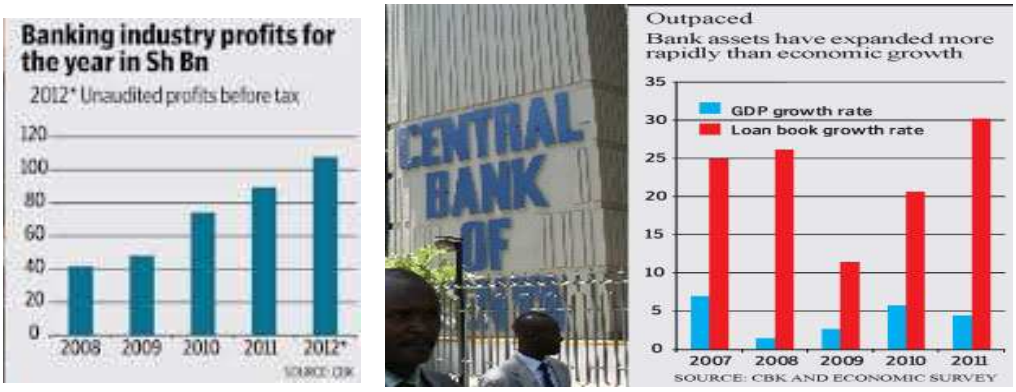


Figure 1

1.3 OBJECTIVES

Main Objective

The overall objective of this study is to use survival analysis techniques to generate default trends at various points in lifetime of a loan using various borrowers' application characteristic.

Specific objectives

The specific objectives of this study are;

(i) Identifying which application variables affect default.
(ii) Estimating time to default using Kaplan Meier method for each risk group.
(iii) Testing difference in the survival curves for each risk group using the non-parametric tests of log-rank and Wilcoxon test.

1.4 SIGNIFICANCE OF THE STUDY

This study seeks to help banks in;

- Identifying which application variables affect default.
- Applying profit scoring
- Making decision on how long a credit facility ought to be granted.

CHAPTER 2: THEORY OF SURVIVAL ANALYSIS

2.1 INTRODUCTION

This chapter looks at the history of survival analysis, types of censoring, functions of survival time, survival analysis techniques and hypothesis testing of survival curves

2.2 BACKGROUND OF SURVIVAL ANALYSIS

The origin of survival analysis goes back to centuries ago from the mortality tables however its rebirth emerged after World War II stimulated by interest in reliability (or failure time) of military equipment. New statistical methods then emerged from the strict mortality data research to failure time research. These ideas quickly spread through private industry and customers' response was need for more safer and reliable products.

Survival analysis is a branch of statistics dealing with study of "TIME UNTIL AN EVENT OCCURS''. Several events might be observed in the course of the study but only one event might be of interest. When more than one event of interest is to be considered then the problem becomes a recurrent event or competing risk problem. (Prinja, Gupta, & Verma, 2010)

Based on the field of study, survival analysis has also been referred to as lifetime data analysis, reliability theory in engineering and event history analysis.

Survival analysis application has been extended to various fields of study and in the banking industry the events include time to credit default, time to early repayment; in the engineering field the events are failure time of machines and parts, life of bulbs etc. In medical fields we have, survival time after treatment, incubation period i.e. time of infection to time of disease occurrence, incidence period e.g. high dose given to mice and time to death due to effect of the dose. In demography the events are age at death, birth, first marriage etc.

In most studies, there will be subjects who choose to quit participating, who move too far away to follow, or who will die from some unrelated events. Such kind of exit is called censoring. Survival analysis gives such studies a major breakthrough since it allows researchers to include the information of these individual up to the point they exit/censored.

Survival analysis is therefore viewed as the analysis of **censored data.**

2.3 CENSORING

Censoring occurs when we have some information about individual survival time, but we don't know the survival time exactly. Censoring is attributed to some random cause during the course of study.

There are 3 major classification of censoring and these are: *right censoring, left censoring and interval censoring.*

i). Right censoring

This is the most common type of censoring. In this scenario, the study participants are tracked until a time when they take no further part in the study, but the event of interest has yet to occur. Such occurrence can be attributed to:

- The study comes to an end while the participant has not experienced the event of interest (also known as withdrawal alive as the individual is still surviving);
- Participant leaves the study for some other reason which is independent of the interest of study (This is also called loss to follow up) ;

ii). Left censoring

This is a scenario whereby some of the study participants have already experienced the event of interest yet the study has not yet begun and as such it is not exactly known the occurrence time of the event. Left censoring is an uncommon type of censoring and is not usually a problem in various studies, since beginning point is defined by an event e.g. entry of participants in study and as such participants who have experienced the event of interest are likely to be excluded from the study.

Examples of left censoring include: Borrower has already defaulted at the beginning of the study.

iii). Interval censored data

Interval censoring occurs when monitoring of happening of events is done periodically. The exact time of occurrence of an event is not accurately known but a time interval bound of the happenings is well known.

Usually, if the assessment intervals are very short e.g. monthly, then it is common to pick one end point consistently for the study e.g. month end status. This therefore implies that interval censoring can be dealt with as point censoring if the frequency of assessment is justifiable.

Examples of interval censoring include credit defaults during any given month etc.

2.4 FUNCTIONS OF SURVIVAL TIME

Survival analysis has 3 functions to be studied. This in particular are the ***Survival function, Probability Density Function*** and the ***hazard function.***

Let *T* be the length of time to an event under consideration. T is therefore a random variable and there are 3 different ways of describing the randomness of T in survival analysis **S**

a. Survival Function {S(t)}

The survival function gives the probability that a subject will survive past time t.

$$S(t) = Prob\{T>t\}$$

As t ranges from 0 to 1, the survival function has the following properties:

- It is non-increasing
- At time t = 0, S (t) = 1. In other words, the probability of surviving past time 0 is 1.
- At time t = 1, S (t) = S (1) = 0. As time goes to infinity, the survival curve goes to 0.
- In theory, the survival function is smooth. In practice, we observe events on a discrete time scale (days, weeks, etc.).

$F(t) = Prob\{T \leq t\} = 1 - Prob\{T > t\}$

$= 1 - S(t)$

$F(t)$ is the Cumulative Distribution Function (CDF)

b. Probability Density Function {f(t)}

$$f(t) = \lim_{\Delta t \to 0} \frac{prob(t < T \leq t + \Delta t | T > t)}{\Delta t}$$

$$= \lim_{\Delta t \to 0} \frac{F(t+\Delta t|T) - F(t)}{\Delta t}$$

$$= \frac{d}{dt} F(t) = \frac{d}{dt} \{1 - S(t)\}$$

$$= \frac{d}{dt} \{1 - S(t)\}$$

$$= \frac{-dS}{dt}$$

c. The hazard function {h(t)}

The hazard function, $h(t)$ is the instantaneous rate at which events occur i.e. it gives the rate of change of probability of failure at a time t:

$$h(t) = \lim_{\Delta t \to 0} \frac{prob(t < T \leq t + \Delta t | T > t)}{\Delta t}$$

From Bayes theorem:

$$p(A/B) = \frac{Prob(A \cap B)}{Prob\ (B)}$$

Therefore;

$$h(t) = \lim_{\Delta t \to 0} \frac{prob(t < T \leq t + \Delta t\ \cap T > t)}{\Delta t\ .Prob(\ T > t)}$$

$$h(t) = \frac{1}{Prob(\,T > t)} \lim_{\Delta t \to 0} \frac{prob(t < T \leq t + \Delta t \ \cap T > t)}{\Delta t}$$

$$h(t) = \frac{1}{S(t)} \lim_{\Delta t \to 0} \frac{prob(t < T \leq t + \Delta t)}{\Delta t}$$

$$h(t) = \frac{f(t)}{s(t)}$$

The cumulative hazard describes the accumulated risk up to time t,

$$H(t) = \int_0^t h(u)du$$

Functions relationships

Since the above 3 functions are mathematically equivalent, knowledge of any one function means the other 2 functions can be derived.

$$h(t) = \frac{f(t)}{s(t)} = \frac{1}{s(t)}[-\frac{dS}{dt}]$$

$$h(t) = -\frac{1}{s(t)}.\frac{dS}{dt} = -\frac{d}{dt}\log S(t)$$

$$but\ \frac{d}{dt}\{\log s(t)\} = -h(t)$$

$$\log s(t)|_0^u = -\int_o^u h(t)dt$$

$$\log s(u) - \log\ s(0) = -\int_o^u h(t)dt$$

But $s(0) = Prob(T > 0) = 1 - Prob(T \leq 0)$

$$= 1 - F(0) = 1 - 0 = 1$$

Therefore

$$\log s(u) = -\int_{o}^{u} h(t)dt$$

$$s(u) = \exp\{-\int_{o}^{u} h(t)dt\}$$

In general;

$$S(t) = \exp\{-\int_{0}^{t} h(u)du\} \ = \exp\{-H(u)\}$$

<u>Relationship summary</u>

i) $f(t) = \frac{-dS}{dt}$

ii) $h(t) = \frac{f(t)}{s(t)}$

iii) $S(t) = \exp\{-\int_{0}^{t} h(u)du\} \ = \exp\{-H(u)\}$

2.5 SURVIVAL ANALYSIS TECHNIQUES

There are 3 broader classification of the survival analysis methods used for dealing with censored data. These are ***parametric, non-parametric and semi-parametric methods.***

2.5.1 Parametric Survival Analysis:

Parametric survival models make assumption on the distribution of the outcome (survival time). The outcome is assumed to follow some family of probability distributions of similar form with unknown parameters. Since even the parameters are unknown, their exact distribution is also unknown and data is used to estimate the parameters. Once the value of the parameter is known by way of estimation then can the distribution be fully specified.

Given that the outcome distribution is assumed to follow some distribution whose probability density function *f (t)* can be expressed in terms of unknown parameters, survival and hazard functions can then be determined once the pdf is specified for survival time.

If we have two Poisson distribution with one of them having a mean of 8 and variance of 8 and another one having a mean of 10 with a variance of 10 then these two distributions are from the same family (normal) but they are not exactly the same distribution.

Exponential, Weibull and log-logistic distribution are some of the distributions commonly used in survival analysis. Their survival and hazard functions are given below.

Distribution	***S*(*t*)**	***h*(*t*)**	***f* (*t*)= *S*(*t*). *h*(*t*)**
Exponential	$\exp(-\lambda t)$	Λ	$\lambda\exp(-\lambda t)$
Weibull	$\exp(-\lambda t^p)$	$\lambda p\, t^{p-1}$	$\lambda p\, t^{p-1} . \exp(-\lambda t^p)$
Log-logistic	$\frac{1}{1+\lambda t^p}$	$\frac{\lambda p\, t^{p-1}}{1+\lambda t^p}$	$\frac{1}{1+\lambda t^p}.\frac{\lambda p\, t^{p-1}}{1+\lambda t^p}$

From the table above, the survival time is expressed as a function of the explanatory variables thus most of the parametric models are acceleration failure time models.

There are various methods for parameter estimation. These are;

i. The Maximum Likelihood Estimation (MLE) method

ii. Least Square Method

iii. Methods of moments

iv. Cramer Rao method

The mostly commonly used is the MLE method and this is described below.

The Maximum Likelihood Estimation (MLE) method

The MLE method takes care of 2 different scenarios i.e. with complete data (uncensored) and the other involves censored data.

Case 1: With complete data

Consider a random (independent) sample of size n, with observations $x_1, x_2, x_3, \ldots x_n$ coming from a larger population with pdf $f(x,\theta)$, that is;

x_1 has pdf $f(x_1,\theta)$; x_2 has pdf $f(x_2,\theta)$;…; x_n has pdf $f(x_n,\theta)$

The joint pdf of the sample is the product of their respective pdf's. This joint pdf is what is called the Likelihood Function denoted by L.

$$L = f(x_1,\theta).\, f(x_2,\theta)\ldots f(x_n,\theta)$$

$$L = \prod_i^n f\ (x_i,\theta)$$

To solve for the estimate of θ , the equation $\frac{dL}{d\theta} = 0$ is solved.

However in most cases *log L* instead of L is used hence the estimation is;

$$\frac{d}{d\theta} log\ L\ = 0$$

Case 2: With incomplete (censored) data

Let $d_i = \begin{cases} 1;\ if\ the\ ith\ life\ is\ uncensored \\ 0;\ \ if\ the\ ith\ life\ is\ censored \end{cases}$

Then;

$$L = \prod_i^n [f(t_i)]^{d_i}\ [S(t_i)]^{1-d_i} \qquad (i)$$

Alternatively;

$$L = \prod_i^r [f(t_i)]\ \prod_{i=r+1}^n [S(t_i)] \qquad (ii)$$

where ***r*** are uncensored and ***n-r*** are censored.

From *(i);*

$$L = \prod_i^n [f(t_i)]^{d_i}\ \frac{S(t_i)}{[S(t_i)]^{d_i}}$$

$$L = \prod_i^n \left\{\frac{f(t_i)}{S(t_i)}\right\}^{d_i} [s(t_i)]$$

$$L = \prod_i^n [h(t_i)]^{d_i}\ [s(t_i)]$$

Accelerated Failure Time (AFT) Models

Parametric methods of survival analysis assume distribution of hazard rate as a function of time.

The AFT model is a general model for survival data in which the explanatory variables measured

on an individual are assumed to act multiplicatively (proportionately) with respect to survival time, and so affect the rate at which the individual proceeds along the time axis, that is, the model is interpreted in terms of the speed of progression of an event.

(Kleinbaum, n.d.) compared the survival functions among smokers $S_1(t)$ and non-smokers$S_2(t)$. They expressed the AFT assumption as $S_1(t) = S_1(\gamma t)$ for $t \geq 0$, where γ is a constant called the *acceleration factor* comparing smokers to nonsmokers. In a regression framework the acceleration factor γ could be parameterized as *exp (α)* where *α* is a parameter to be estimated from the data. With that parameterization, the AFT model assumption was expressed as$S_1(t) = S_1(exp\ (\alpha)\ t)$.

The general form of AFT model for the hazard function is given by;

$$h_i(t) = \emptyset^{-x_i} h_o(t/\emptyset^{x_i})$$

Where $\emptyset^{-1}$ is the acceleration factor, x_i is the value of X variable for the *ith* individual in the study and when $x_i = 0$ then $h_o(t/\emptyset^{x_i})$ reduces to $h_o(t)$ which is the baseline hazard function.

Since $\emptyset$ must be non-negative, it is then set to $\emptyset = e^{\propto}$. Therefore we have that;

$$h_i(t) = e^{-\propto x_i} h_o(t/e^{\propto x_i})$$

When we have *p* explanatory variables recorded for each individual in a study, then the hazard function for i^{th} individual at time *t is* given by;

$$h_i(t) = e^{-n_i} h_o(t/e^{n_i})$$

Where;

$$n_i = \propto_1 x_{1i} + \propto_2 x_{2i} + \cdots + \propto_p x_{pi}$$

The corresponding survivor function is given by;

$$S_i(t) = s_o(t/e^{n_i})$$

2.5.2 Non-Parametric Survival Analysis:

Non parametric methods compute the instantaneous probability of an event of interest occurring at a certain point in time given that an individual has survived up to that point. Non parametric methods thus have a major advantage of maximizing the use of individual's available information up to the point when the individual is censored rather than excluding all information on a censored person.

Under non-parametric estimation we have: ***Kaplan Meier (Product Limit) Method, Nelson Aalen approach*** *and* ***Delta Method.***

a) <u>Kaplan – Meier (Product Limit) Approach</u>

Let:

1. N = The population /sample size
2. t_j = The time at death for $j = 1,2,\ldots,k$ such that $t_1 < t_2 < \cdots < t_k$
3. d_j = The no. of deaths at time t_j where $d_1 + d_2 + \cdots + d_k = m$
4. C_j = *The no. of individuals censored between time* t_j *and* t_{j+1} *for* $j = 1,2,\ldots,k$

$$Total\ number\ censored = N - m$$

5. n_j = *The no. of individuals at risk just before time* t_j

Note: $n_{j+1} = n_j - (d_j + c_j)$

The Kaplan – Meier estimator is given by;

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

$$Var\ \hat{s}(t) = [\hat{s}(t)]^2 \sum_{t_j \le t} \frac{d_j}{n_j(n_j - d_j)}$$

95% Confidence Interval;

$$\hat{s}(t) \pm 1.96\sqrt{Var\ \hat{s}(t)}$$

b) <u>Nelson-Aalen Method of Estimation</u>

From the geometric series we have that;

$$-\log(1 - x) = x + \frac{x^2}{2} + \frac{x^3}{3} + \ldots$$

For small x we ignore the powers of x, therefore;

$$\text{Log}(1 - x) \approx -x \qquad (1)$$

The Kaplan Meier estimate is given by:

$$\hat{s}(t) = \prod_{t_j \le t}\left(\frac{n_j - d_j}{n_j}\right) = \prod_{t_j \le t}\left(1 - \frac{d_j}{n_j}\right) \qquad (2)$$

Also $S(t) = \exp\{-\int_0^t h(u)du\}$

$$ln\ S(t) = -\int_0^t h(u)du \qquad (3)$$

From (2);

$$ln\ \hat{s}(t) = ln \prod_{t_j \le t}\left(1 - \frac{d_j}{n_j}\right) = \sum_{t_j \le t} ln\left(1 - \frac{d_j}{n_j}\right) \qquad (4)$$

Equating (3) and (4);

$$-\int_0^t h(u)du = \sum_{t_j \le t} ln\left(1 - \frac{d_j}{n_j}\right) \qquad (5)$$

Using (1) we know that; $ln(1-\frac{d_j}{n_j}) \approx -\frac{d_j}{n_j}$(5) becomes;

$$\int_0^t h(u)du = \sum_{t_j \le t} \frac{d_j}{n_j} = \hat{H}(t)$$

$H(\hat{t}) = = \sum_{t_j \le t} \frac{d_j}{n_j}$ is called the **Nelson – Aallen Estimator of *H (t).***

H (t) is called the integrated or cumulative function.

$$S(\hat{t}) = exp\ \{-H(\hat{t})\}.$$

$Also;\ \ d_j \sim Bin\ (n_j, p_j)$Where d_j is the no. of deaths at time t_j

$$E\ (d_j) = n_j\ p_j$$

$$Var\ (d_j) = n_j\ p_j q_j$$

In this case;

$$p_j = 1 - \frac{d_j}{n_j} \quad \text{and } q_j = \frac{d_j}{n_j}$$

$$\hat{s}(t) = \prod_{t_j \le t}(1 - \frac{d_j}{n_j}) \quad = \prod_{t_j \le t} p_j$$

$$E\{H(\hat{t})\} = \sum_{t_j \le t}\ E(\frac{d_j}{n_j})$$

$$= \sum_{t_j \le t}\ (\frac{n_j p_j}{n_j}) = \sum_{t_j \le t}\ p_j = \sum_{t_j \le t}\ \{1 - \frac{d_j}{n_j}\}$$

$$Var\ \{H(\hat{t})\} = \sum_{t_j \le t} \frac{q_j p_j}{n_j} = \sum_{t_j \le t} \frac{1}{n_j} (\frac{n_j - d_j}{n_j})(\frac{d_j}{n_j})$$

$$= \sum_{t_j \le t} \frac{(n_j - d_j)\ d_j}{n_{j^3}}$$

c) <u>The delta method of estimation:</u>

If $\hat{\theta} \sim N(\theta, \delta_\theta^2)$ then $f(\hat{\theta}) \sim N\{f(\theta), \delta_f^2\}$

Where $\delta_f^2 = \delta_\theta^2 \left|\frac{df}{d\theta}\right|^2$

Giving an example of Bernoulli distribution:

$$f(x) = p^x(1-p)^{1-x}; x = 0,1$$

The delta method is as follows;

Given that $\hat{p} \sim N\{E(\hat{p}) = p,\ Var(\hat{p}) = \frac{p(1-p)}{n}\}$

Then $f(\hat{p}) \sim N\{f(p),\ Var\, f(p)\}$

2.5.3. Semi-Parametric Survival Analysis Method:

The commonly used semi-parametric survival analysis method is the Cox Proportional Hazard Model and is described below

Cox-Proportional Hazard Method (CPHM)

Cox model is called a semi parametric model since it is a product of two functions that depends partly on the baseline hazard function and partly on a vector of coefficients β that are linear multiples of the covariates.

Cox model is called proportional hazards (PH) model because of the baseline hazard function feature that is a function of t and does not involve the X's however the exponential function does not involve t and the X's are called ***time independent X's***. This means that the value of the variables for a given individual does not change over time

Cox (1972) proposed the following model;

$$h_{\underline{x}}(t) = h_0(t).\ e^{\underline{B}'\underline{X}}$$

Where;

$\boldsymbol{h_{\underline{x}}(t)}$= The hazard at time t for an individual with a set of explanatory variables denoted by $\underline{x}$

$\boldsymbol{h_0(t)}$ = The baseline hazard functions at time t when the values of all the explanatory variables equal 0.

$\boldsymbol{\underline{B}}$ = The vector for the regression coefficient

$\boldsymbol{\underline{x}}$ = The vector of covariates (Predictor variables that is modeled to predict an individual's hazard.

The model has it that;

$$h_{\underline{x}}(t) \propto h_0(t)$$

therefore $\boldsymbol{e^{\underline{B}'\underline{X}}}$ **is the constant of proportionality** and is always greater than 0. For the Cox PH the proportionality assumption must hold.

If there are no X's in the model, the model reduces to baseline hazard function and this is why $h_0(t)$is called the baseline (starting) function. Again since the $h_0(t)$ is an unspecified function, this property makes the cox model to be called semi-parametric model as opposed to the parametric model whose functional form is fully specified with the exception of the values of the unknown parameters

The Cox Proportional Hazard Model also has an advantage in that it allows X's which involve *t* called the ***Time-varying Covariates*** (CPHM with TVCs) and is given by

$$h\{t, \underline{x}(t)\} = h_0(t).\ e^{\underline{B}'\underline{X}(t)}$$

The above model of Cox with TVCs gives hazard at time t for observation x, given parameters β. Variables whose values change over time are known as ***Time Dependent Variables***. There are 2 types of time dependent variables i.e. internal and external variables.

Internal variables relate to a particular individual in a study and can only be measured when an individual is alive i.e. repeated measurement of certain characteristics is made on an individual over time e.g. blood pressure while external variables are time dependent variables that do not necessarily require the survival of an individual for their existence e.g. inflation rate, age e.t.c. Of importance is that $\underline{B}'\underline{X}(t)$ can be rewritten as $\underline{B}'(t)\underline{X}$ as $\underline{X}(t) = \underline{X}t$ is a time dependent variable

Cox contribution was to come up with a method for estimating $\underline{B}'s$ for arbitrary$h_0(t)$. The method is known as the partial likelihood approach. This means that the cox model has a flexibility to introduce time dependent covariates described above.

The partial likelihood Function

Considering data with no Ties and no Censoring

Let;

- Times at occurrence of an event of interest (t_i)be arranged in an ascending order i.e. $t_1 < t_2 < t_3 < \cdots < t_k$ for *i=1,2,…,k*
- Risk sets $R(t_i)$ be the set of all members at risk just before time t_i
- Probability of individual *i* dying at time t_i given the risk set $R(t_i)$ be given by;

$$\frac{Hazard\ of individual\ i\ dying\ at\ time t_i}{The\ sum\ of\ hazards\ at\ time\ t_i}$$

Then the Cox partial likelihood is given by

$$L(B) = \prod_i \{\frac{h_i(t_i)}{\sum_{j \in R(t_i)} h_j(t_j)}\} \qquad (1)$$

Since $h\underline{x}(t) = h_0(t).\ e^{\underline{B}'\underline{X}}$, the $h_0(t_i)$ will cancel each other at the numerator and denominator of equation (1) and this will give;

$$L(B) = \prod_i \{\frac{e^{\underline{Bi}'\underline{X}}}{\sum_{j\in R(t_i)} e^{\underline{Bj}'\underline{X}}}\} \quad (2)$$

The maximum Likelihood estimates of *B* are then found by maximizing the logarithm of equation (2).

Considering data with Ties and Censoring

In certain instances we get data that are recorded after a given period of time and as such several failures can be at a particular time. In such cases, the above model is modified such that

$$L(B) = \prod_i^k \{\frac{\exp(S'_{D_i}\beta)}{\sum_{R\in R(t_i;d_i)} \exp(S'_R\beta)}\} \quad (3)$$

Where;

d_i = *The number of failures at time* t_i

D_i = *The set of* d_i *individuals failing at* t_i

$R(t_i; d_i)$ = *The set of all subsets of* d_i *individuals taken from the risk set* $R(t_i)$

$R \in R(t_i; d_i)$ = *The set of* d_i *individuals who might have failed at time* t_i

$S_R = \sum_{l\in R} x_l$ *is the sum of the covariate vectors* ***X*** *over the individuals in set R*

$S_{D_i} = \sum_{l\in D_i} x_l$ = *The sum of covariate vectors of these individuals*

2.6 COMPARISON OF GROUPS OF SURVIVAL DATA

This section involves hypothesis testing of the survival curves of two or more groups to check if they are statistically equivalent. There are 2 methods of this hypothesis testing i.e. the parametric and the non- parametric methods.

i). PARAMETRIC METHODS

Parametric analysis relies on the data being normally (or nearly) distributed so as to test whether estimates on a given population parameter is equal for two samples.

Parametric hypothesis tests the null hypothesis against an alternate hypothesis, example, whether or not the population mean is equal to a certain value, and then using an appropriate statistic to calculate the probability that the null hypothesis is true. You then reject or accept the null hypothesis based on this calculated probability. Parametric analysis can only be used on quantitative data since it is only quantitative data that can have normal distribution.

Below are the parametric Hypothesis Tests that are available

Comparison of means	Parametric (means)
Differences between the means of two independent groups	Independent t-test
Differences between paired (matched) samples e.g. weight before and after a diet for each subject	Paired t-test
Differences in the means of 3+ independent groups for one variable Differences between 3+ measurements on the same subject	One-way ANOVA Repeated Measures

Comparison of means	Parametric (means)
	ANOVA

ii). NON-PARAMETRIC METHODS

Hypothesis testing involves assessing whether survival curves for two or more groups are statistically equivalent.

The problem is to test;

$$H_o : S_1(t) = S_2(t)$$

{No difference between the survival curves}

vs. $H_1 : S_1(t) \neq S_2(t)$

The commonly used tests for non-parametric methods are the;

i). Log-rank test statistics

ii). Generalized Wilcoxon Test Statistics.

The above test statistics are both based on the hyper geometric distribution given by;

	No. dead	No. Alive	Total
Group 1	X	m-x	m
Group 2	r-x	n - r + x	n
Total	R	m + n -r	m + n

$$Prob\ (X = x) = \frac{\binom{m}{x}\binom{n}{r-x}}{\binom{m+n}{r}} \quad ; \quad x = 0, 1, 2, \ldots, r$$

and is called the hyper-geometric Probability Mass Function.

$$E\ (X = x) = \frac{mr}{m+n}$$

$$Var\big(X = d_{1j}\big) = \frac{mnr(m+n-r)}{(m+n)^2(m+n-1)}$$

a). THE LOG RANK TEST STATISTIC

This is a large sample chi-square test that makes use of observed versus expected cell computation over groups of outcomes.

The log–rank statistic groups are defined by each of the ordered failure times for the entire set of data being analyzed.

Let

t_j be time to an event *for j =1,2,…,k* and $t_{1<}t_{2<\cdots<}t_k$

d_j = The total number of events of interest occurring at time t_j

n_j = Individuals at risk of event of interest just before time t_j

d_{ij} = The no. of events occurring for group *i =1,2* at time t_j

n_{ij}=Individuals at risk of event of interest in group *i =1,2* before timet_j

The above information can be summarized in a 2*2 table as follows

Group	No. dead	No. Alive	Individuals at risk in each group just before time t_j
Group 1 Group 2	d_{1j} d_{2j}	$n_{1j} - d_{1j}$ $n_{2j} - d_{2j}$	n_{1j} n_{2j}
Total	d_j	$n_j - d_j$	n_j

$$Prob\left(X = d_{1j}\right) = \frac{\binom{n_{1j}}{d_{1j}}\binom{n_{2j}}{d_{2j}}}{\binom{n_j}{d_j}}; \qquad d_{ij} = O, 1,2, \ldots, d_j$$

$$E\left(X = d_{1j}\right) = \frac{d_j n_{1j}}{n_j}$$

$$Var\left(X = d_{1j}\right) = \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}$$

Recalling from basic statistics that;

$$\text{If } x \sim N(\mu, \delta^2) \text{ then } Z = \frac{x-\mu}{\delta} \sim N(0,1)$$

$$\text{And } Z^2 = \left(\frac{x-\mu}{\delta}\right)^2 \sim \mathrm{X}^2(1)$$

$$\text{Let Y} = \sum_{t_j} d_{1j}$$

Standardizing Y we get;

$$Z = \frac{Y - E(Y)}{\sqrt{Var\, Y}}\ ; E(Z) = 0 \; and \; Var(Z) = 1$$

$$Z^2 = \left\{\frac{\left(\sum_{t_j} d_{1j} - \sum_{t_j} E\left(d_{1j}\right)\right)}{\sqrt{\sum_{t_j} Var\ (d_{1j})}}\right\}^2 \sim \mathrm{X}^2(1)$$

$$Z^2 = \frac{[\sum_{t_j}\left\{d_{1j} - \frac{d_j n_{1j}}{n_j}\right\}]^2}{\sum_{t_j}\left\{\frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}\right\}} \sim \mathrm{X}^2(1) \quad \text{Under } H_o$$

This is called the log-rank test statistics (Under the null hypothesis, the log–rank statistic is approximately chi-square with one degree of freedom)

If the calculated $X^2(1)$ is greater than the tables X^2 at $\propto -$level of significance then H_o is rejected i.e. $S_1(t) \neq S_2(t)$

Log rank test statistic can also be used to compare several survival curves (≥ 3 groups); the null hypothesis is that all the survival curves are the same. The log–rank statistic has approximately a large sample chi-square distribution with G − 1 degrees of freedom.

$$Z^2 \sim X^2(G-1) \text{ df}$$

A similar table used for the 2 groups can be used however the test statistic gets complicated since it involves calculation of covariance and variances of summed observed minus expected values for each group. This calculation has however been made much easier since there are several computer programs that can calculate the statistic.

b). Generalized Wilcoxon Test Statistics.

This test is a modification of the log rank test. The log rank test gives similar weights to each failure time while the Wilcox test apply different weights to each failure time.

The Wilcox test (also called Breslow test) statistic is given by;

$$Z^2 = \frac{[\sum_{t_j} W(t_j)\left\{ d_{1j} - \frac{d_j n_{1j}}{n_j}\right\}]^2}{\sum_{t_j} W(t_j)\left\{\frac{n_{1j}n_{2j}d_j(n_j-d_j)}{n_j^2(n_j-1)}\right\}} \sim X^2(1)$$

$W(t_j)$ = weight at jth failure time

$W(t_j) = n_j(number\ at\ risk)$ for Wilcoxon test but equals to 1 for the log rank test statistic.

In the Wilcoxon scenario, the observed versus expected values at time *tj* are subjected to weighting by considering the number at risk *nj*, over all groups at time tj.

This therefore means that more weight is assigned at the start of the survival curve since the number exposed to risk is great compared to later stages i.e. more weight applied to early failures than later time failures.

It is appropriate to use this kind of weighting if treatment effects is more pronounced at the early stages of administration than later stages

CHAPTER 3: LITERATURE REVIEW

3.1 INTRODUCTION

This chapter looks at the previous works done which are related to the topic of study. It gives the names of the author(s), topic of study, year of publication and the journals used. It then summarizes the reviews in form of a table to identify gaps and finally group the findings.

1. **Wekesa, Okumu Argan; Samuel, Mwalili; Peter, Mwita (2012*). Modeling Credit Risk for Personal Loans Using Product- Limit Estimator.* International Journal of Financial Research; Vol. 3 Issue 1, p22-32**

The purpose of this study was to estimate default probabilities at various points in time using product limit estimator and to test the statistical significance of the differences in the survival curves for 2 risk groups, namely male and female applicants based on log-rank tests.

Methodology

250 female and 250 male applicants were randomly chosen from a Kenyan bank portfolio of personal loans whose maturity was 30 months. This was a group (cohort) of loans taken in the month of January, 2007 and were observed for a period of 30 months (January 1, 2007 to June 30, 2010).

An account was considered bad if it missed payments for 2 consecutive months. If an account did not miss 2 consecutive months and was closed or survived beyond the observation period then the study considered it as having been censored. Cases of early settlements or repayment were considered as censored in this study.

The life of the account was thus being measured from the time it was opened up to the time it got a bad status or was censored for any reason.

Time in months at which the borrowers made early repayment (censored) or defaulted were then obtained differentiating for females and males and arranging the times in ascending order.

The Product limit estimator was then used to asses survival probability for the gender based study.

Mean survival time and median survival times were also obtained for the 2 groups.

The survival curves for males and females were then compared using the log rank test. The hypothesis tested was;

H_o: The male and female curves are statistically different

H_1: The male and female curves are statistically the same.

Result

Out of 250 male loan applicants for loans maturing in 30 months, 11 defaulted and 4 settled their loan accounts before maturity. Mean survival time was 15 meaning that on average, a male applicant would take 15 months to default while a female would take on averagely 16 months to default.

The survival curves of the 2 groups were found to be similar. This was confirmed by test statistic (log rank 0.17) that gave a significance value of 0.678 that showed that the 2 survival distributions were statistically the same at 95% Confidence interval. If the significance value was <0.05 then the null hypothesis would have been accepted.

Conclusions

From the data used, there was no significant difference between male and female borrowers in their time to default. It was thus not meaningful to classify borrowers on the basis of gender as this did not affect credit risk. Mean survival times would guide underwriting on average age for loans so as to minimize losses emanating from loan defaults thereby optimizing profit returns.

Recommendation and Suggestions for Further Research

Product limit method was reasonably reliable to use compared to the parametric methods as it did not make assumptions about loan default distribution. However, since this method is a univariate method, it may be more informative to adopt multivariate techniques like Cox model to model credit risk. Thus further research could be conducted on the same data set using other survival techniques.

2. **Lm J.K, Apley DW, Qi C, Shan X (2012)** ***A time-dependent proportional hazards survival model for Credit Risk Analysis*****. Journal of the Operational Research Society 63(3): 306-321**

Purpose:

This paper uses a modification of the proportional hazards survival model that includes a time dependency mechanism for capturing temporal phenomena (dynamic economic conditions) and develops a maximum likelihood algorithm for fitting the model.

Methods

A set of credit card customer performance data was used. Data consisted of (1) All customers card requests approved between January 2003 to July 2008 and who defaulted at some point within this period (2) A 1% random sample of all customers who were approved between January 2003 and July 2008 and who did not default. Thus, data set obtained by random under-sampling of the majority class (those who did not default) and by using the entire minority class (those who did default). Data set was balanced to improve classification performance, Batista *et al* (2004) and Chawla *et al* (2004)

A random sample of $2/3$ was used for training and the rest $1/3$ for testing out of a total of 212,742 customers. There were several predictor variables but only 75 were chosen by field expert to be potentially significant. Only 10 rather significant variables were chosen for the study out of 75 predictor variables by simply using a forward sequential feature selection method. ***The 10 variables could not be disclosed due to confidentiality issues.***

The MLEs algorithm was used to estimate the values of the time dependent function (variable) γ and the values of the parameters of the exponential function $\Psi(x)$and the log normal base hazard function $h_o(t)$ model were derived.

Time Dependent Proportional Hazard (TDPH) $\widehat{\boldsymbol{\beta}}$ were estimated and compared with the PH and LR models. The modeling approach was used to assess the effectiveness in representing the time to default distribution under dynamic market conditions and in scoring customers for credit risk.

Results

TDPH model fitted the data well. As a scoring method that takes into account dynamic market conditions, the LR model using the TDPH as an additional predictor achieved roughly a 3.2% improvement in KS Statistic over regular LR when 2 different vintage windows of data are considered under quite different conditions.

Conclusion

-This result is consistent with the results of Teng et al (2007) and Bellotti and Crook (2009), who found modest but statistically significant improvements in predictive performance using macroeconomic variables with substantial dynamic variability.

- Overall, incorporating the TDPH into either the LR or the PH approaches improves the performance of these methods.
- Additional benefits of the TDPH approach is that it provides an inherent mechanism for adjusting the customer acceptance threshold to keep constant the collective default rate of accepted customers in the face of dynamic market conditions.

3. Bellotti T and Crook JN (2009). ***Credit Scoring with Macroeconomic Variables using Survival Analysis. J Opl Res Soc*** **60:1699-1707**

Objective

- To show that survival analysis is competitive for prediction of default in comparison with logistic regression.
- To explore the hypothesis that probability of default is affected by general conditions in the economy over time i.e. inclusion of the macroeconomic variables provides a statistically significant improvement in predictions of default.

Data & Methods:

Data

- Credit card application and monthly performance data from a UK bank was used. The card accounts were opened between 1997 to mid-2005. Accounts opened between 1997 and 2001 were used as a training data set, and those opened between 2002 and 2005 were used as a test data set. Each data set contained over 100,000 accounts with application variables such as *income, age, housing and employment status* along with a bureau score **taken** at the time of application.

- An account is in default status if it goes 3 months down or more within the first 12 months for this particular study. An account that defaults is referred to as a bad case and a non-defaulting account is referred to as a good case**. *For this data set, using this definition, the proportion of bad cases in the data was small.***

Macroeconomic variables.

- The following macroeconomic variables were used Interest Rates (IR), Earnings, FTSE, Unemployment (Unemp), Production (Prod), House Price Index (House) and Consumer Confidence Index (CC). These variables were selected as the most likely to affect default. A positive value meant that as the value of the macroeconomic variable rises, this was linked to a rise in risk of default and vice versa e.g. interest rate had a positive value meaning that increase in interest rate is expected to place further stress in the economy resulting into increase in default while production that has a negative value is an indicator of improving economy providing conditions for reduced risk of default.

Methods

- Since the data was skew in terms of good to bad cases, ***greater weight was given to the bad cases***. This is possible for both Cox PH and logistic regression (LR) models since both use Maximum Likelihood Estimation for which bad cases can be included in the likelihood function multiple times.

  Training data was modelled using Cox PH survival model to model time to default with each macroeconomic variable. Cox PH model was used since it allows for inclusion of

macroeconomic variables as Time Varying Covariates (TVCs). This was contrasted with the logistic regression (LR) which is a standard model for scoring.

A Cox PH Model without macroeconomic variables was also built to determine whether any uplift in performance was due to the use of Cox PH Model or the inclusion of macroeconomic variables.

- Each macroeconomic variable was then interacted with an application variable and added to the basic model. ***It was expected that some categories of credit consumers would be more prone to changes in economic conditions than others***. The uplift of the model was then measured using the Log Likelihood Ratio (LLR) derived from the Maximum Likelihood procedure used to estimate the model. The interaction giving the lowest p-value for its LLR is included in the optimal macroeconomic Cox PH model.

Assessment:

- The optimal model was assessed in terms of both its explanatory power on the training data and its predictive power on the independent test set.
- The Cox model was assessed as an explanatory model by reporting its fit to the training data with and without macroeconomic variables using LLR. The significance of each coefficient in the model is determined using a Wald statistic derived from MLE. The Wald statistic follows a chi square statistic, so a p-value can be computed for the null hypothesis that the coefficient value is Zero.

Results and Conclusion:

- Interest Rates (IR), Earnings, FTSE, Unemployment (Unemp), Production (Prod), House Price Index (House) and Consumer Confidence Index (CC) were all found to be significant macro-economic variables with all having a positive correlation with default except Earnings and Production that were negatively correlated i.e. as the variable increases then there is a decrease in risk of default. Interaction with other application variables was also found to be very significant e.g. interaction of IR and Income were highly significant. Increase in interest rate was expected to place further stress in the economy resulting into increase in default while production that has a negative value is an indicator of improving economy providing conditions for reduced risk of default.

- The inclusion of macroeconomic variables consistently gave better performance over time. The Cox PH model with macroeconomic variables outperforms LR. There is also a general improvement in prediction over time using macroeconomic variables, in relation to LR.
- This method of estimation also makes this model suitable for stress testing by including macroeconomic conditions that simulate a depressed or booming economy. This makes it valuable for the implementation of the requirements of the Basel II Accord (e.g. see Basel II paragraph 415).

Recommendation:

- Future lines of research should focus on further application of these methods to other credit card and fixed loan products.
- Also, although the analysis of the explanatory model gives an understanding of how each macroeconomic variable contributes to modeling the data, further extensive experimental work is required to determine the effect of each of the macroeconomic variables on the prediction of Probability of Default.

4. Stepanova M and Thomas LC (2002). *Survival analysis methods for personal loan data.*

Opl Res 50: 277–289

Objective:

This paper identifies three developments that improve the present application of Cox's proportional hazards model to build credit-scoring models for personal loan data that assess aspects of profit as well as default.
Firstly, it develops a new coarse-classifying approach for the characteristics in credit scoring.
Secondly, it explains how the residual tools can be used for examining fitness of the model, and discusses pluses and minuses of each of these tools.
Finally, the paper expands the use to the time-dependent models to overcome the restriction of the proportional hazards.

Data and Methods:

- 50,000 personal loan data obtained from U.K. financial institution with their repayment terms varied from 6 to 60 months. The data set had repayment status for each month of the observation period up to 36 months.
- There were 16 application characteristics used to mention a few, customers age, amount of loan, account closing date, years at current address, years with current employer, gender, no. of dependent children, frequency paid, home phone no. given, Insurance premium, loan type (single/joint), marital status, account opening date, term of loan, home ownership and purpose of loan. There were 22 different purposes for loan.

Coarse-Classifying Using the Survival-Analysis Approach

- Application characteristics of continuous variables were coarse-classified into categories bands to ensure that credit scoring systems were robust i.e. predictive rather than descriptive of data.
- Continuous variable e.g. age were coarse classified as follows;

1. Split the characteristic into 15 to 20 equal bands.
2. Create a binary variable for each band.
3. Fit Cox's proportional hazard model to these binary variables.
4. Chart parameter estimates for all bands. (Parameter estimates from the PH model predicting default/early repayment are drawn in a chart/graph)
5. Choose the splits based on similarity of parameter estimates.

For discrete characteristics such as purpose of the loan, a binary variable was created for each attribute of the characteristic and then the method is the same as for a continuous characteristic. After charting the 22 purpose parameter estimates, then, three binary indicator variables are created so that one has purposes with the highest parameters, i.e., purposes with highest risk of early repayment or default, the second one has purposes with the middle values of parameters, and the third one has purposes with the lowest parameters.

Separate splits were done for every type of failure considered (early repayment and Default) since the effect of the characteristics differed substantially for default and early repayment.

Predicting default

- Based on the competing risk approach, loans that are defaulted are considered failures while all others are considered censored.

- The data is then modelled using Cox's proportional hazards model (PH) compared with a logistic regression approach (LR) under two criteria:

1. Estimating which loans defaulted within the first 12 months.
2. Estimating which loans, which are still repaying after 12 months, will default within the next 12 months.

   Two separate LR models were built on the training sample for each of these definitions. One PH model was fitted to the times until default, considering all other outcomes to be censored.

- To compare LR and PH models, the latter were measured under two criteria whose two definitions are as follows:

  1. PH model gives the ordering of relative likelihood to default, i.e., for each customer there is a "score" that reflects the estimated likelihood to default relative to others.

  2. The cutoff is then chosen in both the PH and the LR models so that number of predicted "bads" equals actual number of "bads" in some holdout sample

- The numbers of "bads" and "goods" correctly classified by the PH and the LR models for predicting default in the holdout sample are compared and the result suggest there is a little difference between the models in either the 1st or the 2nd year, and that segmentation has a less dramatic improvement on PH results under the default criteria than the early repayment criteria.

- ROC curves were also produced to compare performance of the two models under the above criteria. The results without segmenting on the term of the loan, LR and PH give very similar results in both the 1st and the 2nd years i.e. segmentation by term of loan has less effect in predicting default.
- ***Segmentation by term of the loan (24, 36 and 48 months) had less effect in predicting default because default was independent of term of the loan.***

- SAS statistical software was used to fit both the PH and the LR models with procedures PHREG and LOGISTIC, respectively. There are three options of treatment of ties available in the PHREG procedure: "Breslow," "Efron," and "discrete," which correspond to three different approximations of exact likelihood. The SAS statistical package recommends "discrete" for the data that contain large number of ties.

- The log likelihood values were then obtained by fitting the proportional hazards model to the data using the discrete method and the Breslow approximation. The smaller value of the log-likelihood statistic indicates better fit to the data. From the log-likelihood values, discrete approximation gave a much better fit in the segmentation by term, but there was almost no difference in parameter estimates and no difference in the number of correctly classified accounts between the two methods. This suggests that the Breslow approach is a good approximation, and because it is by far the fastest method of the three, it was used for the majority of calculations.

Comparison of Model Diagnostic Methods

Cox-Snell residuals were then calculated and is given by;

$$r_{c_i} = \exp(\hat{B}\, x_i)\, \widehat{H_o}(t_i) = \widehat{H_\iota}(t_i) = -\log(\widehat{S_\iota}(t_i))$$

Where $\widehat{H_o}(t_i)$ is the estimated cumulative baseline hazard, $\widehat{H_\iota}(t_i)$ is the estimated cumulative hazard for individual *ith* individual at time t_i and $\widehat{S_\iota}(t_i)$ is the estimated survivor function.

The residuals were then examined to check whether they have unit mean exponential distribution i.e. the Kaplan-Meier estimate of the survivor function was obtained and log-log transformation of these values plotted against log of the corresponding residual. A straight line with unit slope and zero intercept indicates that the fitted model is correct. The plotted points were close to the straight line, with unit slope and zero intercept if the observations with the lowest residuals were ignored thereby concluding that the model fitted the data well.

Martingale residuals given by $r_{M_i} = \delta_i - r_{c_i}$ were also plotted against rank order of time and should not exhibit any pattern if the model is adequate. The residuals interpreted as the difference between the observed number of failures for an individual in the interval $(0, t_i)$ and the expected number of failures.

The values appear in two bands, one representing uncensored observations and another representing censored ones. This is because Martingale residuals are always negative for the censored observations. The scatter of the points within a band increases with rank order of time.

Deviance residuals were also plotted and are very similar in appearance to Martingale residuals. There are no clear outliers. Because the number of observations is very large, it is doubtful that

these plots can be as useful in identifying problems with the model as in medical studies, where the number of observations is fewer.
Because of the large number of observations, the explainable patterns are clearly visible and overshadow any other systematic features or outliers.

Schoenfeld residual that is the difference between the observed value of the covariate xi and its expected value, conditional on the risk set R_i were also plotted to investigate whether any covariates needed to be transformed or whether the effect of a covariate on the survival time changes over time. The diagnostic was very laborious when the number of covariates is as large as in the data used. The plots did not show any signs of time dependency or transformation.

TIME-DEPENDENT EFFECTS OF COVARIATES

Finally, the paper expanded to the use to the time-dependent models to overcome the restriction of the proportional hazards. This is an extension to the PH model.
Since the predictor variables were large in number, they were screened for possibility of the time-dependent effect before including a time-by-characteristic interaction in the model.

Test for Time-Dependency

Harrel's test was chosen as a screening test for including the time-dependent covariate because it is close to the time-dependent covariate test in power, and is also computationally simple. It is based on Fisher's z-transform of the Pearson correlation between Schoenfeld residuals of the model and rank order of time. The statistic is a normal deviate, so its value is compared with normal distribution tables to test for significance. Harrel's Z-test suggested several possible time-by covariate interactions.

Parameter estimates, from proportional hazards regression predicting default when no time-by-covariate interactions, are compared with parameter estimates when time-by-covariate interactions are included.

It was found out that including time-by-characteristic interaction in the credit scoring for loan data adds another dimension—flexibility to reflect an increase or decrease of the effect of a characteristic with the age of the loan.

Conclusions

- Data analysis, specifically match-rate tables and ROC curves, supports the idea that survival-analysis models are competitive with the industry standard logistic regression approach when used for the traditional purpose of classifying applicants into two groups.
- Segmenting continuous-characteristic variables and regrouping discrete ones using survival-analysis techniques are more appropriate than the traditional method of using good-bad ratio, if one wishes to avoid choosing an arbitrary time horizon. It is important to do such segmentation separately for all types of failure under consideration because the attributes of the most risky individuals depend on the type of failure.
- *Segmentation by term of the loan has less effect in predicting default than early repayment because default is independent of term of the loan and early repayment is not.*
- Diagnostics methods to test the model adequacy were compared and all suggested that the model fits well. Cox-Snell residuals were the easiest to interpret when analyzing loan data.
- Several tests for time-dependency of the effect of a covariate were considered, and Harrel's Z-test was found to be the most appropriate. Time-by-characteristic interactions suggested by Harrel's test were included in the model. This extension allows the effect of a covariate on the predicted time-to-failure to increase or decrease as the loan evolves.

5. Banasik J, Crook JN, Thomas LC (1999). ***Not if but when will borrowers default.*** **J Opl Res Soc 50:1185-1190** (Banasik et al., 1999)

Objective

- The aim of this study is to show how some of the ideas of survival analysis may be applied in the credit scoring context (time to default or to early repayment) and compare the results with more standard approaches.

Data & Methods:

The data consisted of application information of 50 000 UK financial loans accepted between June 1994 and March1997 together with their monthly performance description for the period up to July 1997.

The sample was split in two random groups where 70% was used to build the systems and the remaining 30% (15,018 cases) was used as a holdout sample.

The initial application characteristics included information on age, marital status, employment, residency type, electoral role information as well as loan specific information such as the purpose of the loan and its term. Variables that had several categories were split into attributes by combining answers e.g. purpose of loan that had 25 categories were reduced to four by putting together ones with similar purposes where the default rate was not dissimilar to form categories like vehicle purchase whether for new or old cars, motorbikes or other vehicles were put together. Continuous variables like age were made categorical to have <21, 22-24, 25-32, 33-42, 43-51 and over 51. Banding of variable like age and income is a standard procedure because even for continuous variables, the default risk is not monotone.

The number of months until the loan defaulted, paid off early or paid off on time was recorded. Therefore for each loan one had a survival time, whether it was censored or not. In building the models for time until default, all the cases of early or normal payoff as well as those which were still active by July 1997 were considered censored.

The data was analyzed using the non-parametric proportional hazards model (no baseline hazard assumption i.e. the Cox PH model), two parametric proportional hazards models using exponential ($h_o(t) = \lambda$) and Weibull baseline hazards ($h_o(t) = k \lambda^k t^{k-1}$) and an ordinary logistic regression scorecard approach. The exp and Weibull approaches give probability functions of the avoidance of default for all durations of the loan, while Cox's proportional hazard approach gives an ordering of relative likelihood to default for each loan.

In all three cases the ordering of the likelihood of default of the loans stays the same at all times, this is the proportional hazard assumption. So the same group is considered most at risk for all ages of the loans. This is not necessarily the case for the standard logistic regression approach

Two measures of how these survival approaches compared with the logistic regression approach were used, namely

(i) How likely are the loans to default in their first 12 months?

(ii) How likely are loans that survive 12 months to default in the subsequent 12 months?

Two separate logistic regression scorecards (LR) were therefore built for each of these approaches. For (i) 'bads' are failures in the first 12 months, 'goods' are all others; for (ii) only loans that are still being repaid at 12 months were considered and 'bads' were the ones that defaulted before 24 months.

This gave the LR approach quite an advantage, in that two separate logistic regression-based scorecards were built and each one was tailored to be the best classifier for the criterion under which it was used.

The same proportional hazard model was measured under the two criteria in turn. Moreover because they are proportional hazards models with constant coefficients it was the same applicants who will be considered most likely to default whatever time period is taken. The aim here was to present the simplest type of survival analysis models however there are generalizations of proportional hazards like allowing the coefficients to be time dependent or taking accelerated life models with other distributions which would have overcome this problem.

The estimators and logistic regression scorecards are built on the training sample and the resulting functions applied to the holdout sample. In each case the cut-off was chosen so that the predicted number of 'bads' equals the actual number of 'bads' in the holdout sample under each of these criteria, which removes any effect of the cut-off.

Two competing risks approaches were identified i.e. either the borrower pays off early or defaults. Exactly the same methods above were used to model time until early payoff. In this case, the loans that were paid off early were the ones that were considered as 'failures' while the repayment times of all the others were considered as censored times. The Cox, Exp and Weib methods were then applied to the early repayment data

Results & Conclusion

The proportional hazard models (exponential, Weibull and Cox's nonparametric models) were competitive with the logistic regression approach in identifying those who default in the first year, and may be superior to that approach for looking at who will pay-off early in the first year.

The proportional hazard results for the second year, where there were fewer defaulters were not so encouraging and suggested that more sophisticated models would have been appropriate. The superior performance on early repayment compared with default might have been because the sample used had already been credit scored. Therefore there are very few bad cases under the default criterion compared with the early repayment criterion, and the survival analysis approach benefits more from a large sample of 'bads' than does the logistic regression approach.

To overcome the fixed for all time risk ordering of proportional hazards, it was proposed to allow the b-coefficients to be time dependent and to use accelerated life models with other distribution families than the Weibull one. Both of these extensions would allow the risk ordering to vary over time. The developments of multi-stage models in survival analysis to deal with these problems could also prove useful in the credit scoring context

6. Stepanova M and Thomas LC (2001). ***PHAB scores: proportional hazards analysis behavioural scores*.** J Opl Res Soc 52: 1007–1016.

Objectives:

Behavioural scoring is a type of credit scoring that is performed on existing customers to assist lenders in decisions like increasing the balance or promoting new products. The paper shows how using survival analysis tools specifically Cox's PH regression, allows one to build behavioural scoring models. Their performance is compared with that of logistic regression.

Data and Methods:

- Data from a UK financial institution containing 11 500 customers with their application characteristics and subsequent performance variables for 36 months was used.
- Application characteristics used included Amount of loan ,Term of loan, Months with current employer, Months at current address, Net income, Living with parents or renting, no. of dependent children, Married, High risk purposes(such as refinance), Low risk occupation code
- The data was split into three samples (two training and one holdout) of approximately equal size. The two training samples were used to simulate the real-life situation and the holdout sample was then scored with the application score and all the PHAB scores so that their performance were evaluated and compared.
- The application score was first built using only application characteristics based on the Cox PH on the first training sample .Time to default was thus computed and the loan duration of customers who did not default was considered censored.
- Since Stepwise proportional hazards regression on the first sample was used to build an application score, this was an exploratory analysis to see if PHABS made sense, no transformation was done neither to the variables nor coarse-classifying the first five variables which were continuous. The remaining five variables were binary and were attributes that were found to be of importance by the stepwise procedure.
- The second step was to build a PHAB score on the second training samples for each month of the life of the loan from 4 to 32. The dependent variable was remaining time to default at the month of observation. The model for the ith month was fitted to the remaining time to default, e.g. if a customer defaulted in month 12 and the model is for month 4, remaining time to default was 8 months. The application score from the first step and behavioural (performance) variables were used as predictors variables for the ith month
- Several performance variables and their combinations were tried out to see which ones result in the better fit. The log-likelihood statistic, which indicates how well a model fits the data, was plotted over time for each of the models.

- The predictive power of PHAB scores using the logistic regression model and of the proportional hazards based application score were compared over time using ROC curves with two different definitions of `good' and `bad'. The ROC curve is a plot of the percentage accepted `bad' versus percentage accepted `good' customers. It shows how well the scoreboard discriminates between `good' and `bad'.

Results:

As opposed to what would have been expected intuitively, early default could not be predicted well by application score since, if it could, these applications would have been rejected. However, these early defaulters are identified quite easily using behavioural variables as customers with large balance difference, i.e. customers who are falling behind at the early stages.

ROC curve analysis has shown that Proportional Hazards Analysis Behaviour (PHAB) scores were competitive with the traditional logistic regression scores, especially after about 2 years into the loan.

Furthermore the use of survival analysis enables one to estimate the `survival' probability of the loan over time, i.e. the probability of receiving each of the monthly repayments. This allows one to estimate the profit from the loan, which is an important addition to scoring techniques since lenders are now moving from scoring only for risk to scoring for profitability.

Profit plots versus scores showed that the effect risk has on the return are consistent with what would be expected intuitively. Profit increases as the score goes from high to low risk. The fact that the profit curves cross for different terms of the loans of similar amounts suggests that one has to look at both term and behaviour score when ranking loans of similar amount.

Conclusion and Recommendation:

Early default could not be predicted well by application score since, if it could, these applications would have been rejected. However, early defaulters are identified quite easily using behavioural

variables as customers with large balance difference, i.e. customers who are falling behind at the early stages.

Application and behaviour information complement each other, but their importance changes over time: at the early stages behaviour variables are more important and at the later stages application information becomes more predictive.

Plotting expected profit versus application score, it was found that profit increased as risk decreased meaning that underwriters must look at both term and application score when ranking loans of similar amount.

Proportional Hazards Analysis Behaviour (PHAB) models are competitive with the traditional logistic regression model.

The formula used to calculate profit can be altered to include time-dependent interest rates and hence incorporate economic conditions into the model. Alternatively interest rates can be included as a covariate when estimating survival function.

7. Neural Network Survival Analysis for Personal Loan Data

Objective

The purpose of the paper was to discuss and contrast statistical and neural network approaches for survival analysis in a credit-scoring context. The paper contrasted the performance of a neural network survival analysis model with that of the well-known proportional hazards model for predicting both loan default and early repayment.

Data and Methods:

- The statistical and neural network survival analysis techniques were applied to personal loan data from a major U.K. financial institution. The data set consisted of

the application information of 50,000 personal loans, together with the repayment status for each month of the observation period of 36 months.

- Application characteristics were available in the data set and the status variable indicated which loans were bad, paid off to term, paid off early, or still open. A subsample of 15000 observations was taken and only considered loans having duration of less than 36 months. Missing values were imputed using the mean for the continuous attributes and the most frequent category for the categorical attributes.
- The data was randomly split into a training set (10000 observations) and a test set (5000 observations).
- For the statistical approaches, the experiment was conducted with the logistic regression model, the Cox proportional hazards model while for the neural network analyses, a variant of the approach suggested by Mani was adopted.

For Mani, every observation in the training set, Tmax output units are computed. These output units represent the hazard rate instead of the survival probabilities. The outputs are then computed as follows:

$$s(t) = \begin{cases} 1, & 1 \leq t \leq L \\ -1, & D = 1 \; and \; L < t \leq \text{Tmax} \\ s(t-1) * \big(1 - h(t)\big), & D = 0 \; and \; L < t \leq \text{Tmax} \end{cases}$$

Again, Tmax represents the maximum number of periods involved in the study, L the subject life-time or censoring time, and D indicates if the subject is censored (D = 0) or not (D = 1). For uncensored observations, the hazard is set to zero until the time of death and 1 thereafter. For censored observations, the hazard is set to zero until censoring time and to the Kaplan-Meier estimate thereafter. The survival probabilities may then be estimated by using Kaplan-Meier estimator. The generated survival curves will thus be monotonically decreasing which simplifies the interpretation and increases robustness. However, the topic of time-varying inputs has been left unaddressed.

Results:

The results for predicting default in the first 12 months on the oversampled data set indicated that the logistic regression classifier yielded the best performance followed by the Cox model and the neural network. The performance differences were however not statistically significant.

For loan default between 12 and 24 months, the neural network was superior and yielded a classification accuracy of 78.58% whereas the logistic regression classifier gave 78.24% and the Cox model 77.50%. The performance difference between the NN model and the Cox model was significant.

Conclusion:

It was found that, for early repayment, the suggested neural network approach outperformed the proportional hazards model. For predicting default, the superiority of the neural network model was somewhat less pronounced.

3.2 TABLE OF LITERATURE REVIEW

Authors	(Argan, Corresponding, Samuel, & Peter, 2012)	(Im, Ã, Qi, & Shan, 2012)	(Banasik, Crook, & Thomas, 1999)	Stepanova M and Thomas LC (2001).	(Stepanova & Thomas, 2002)	(Baesens, Van Gestel, Stepanova, Van den Poel, & Vanthienen, 2005)	(Bellotti & Crook, 2007)
Title	Modeling Credit Risk for Personal Loans Using Product-Limit Estimator	Time-dependent PH survival model for Credit Risk Analysis	Not if but when will borrowers default	PHAB scores: proportional hazards analysis behavioural scores	Survival analysis methods for personal loan data.	Neural Network Survival Analysis for Personal Loan Data	Credit Scoring with Macroeconomic Variables using Survival Analysis
Variables	Gender	Not disclosed due to confidentiality issues	Age, marital status, employment, residency type, electoral role , purpose of the loan and term	Amount ,term, Months with current employer, Months at current address, Net income, Living with parents/renting, dependent children, Married, High risk purposes(such as refinance), Low risk occupation code	Age, loan amount, yrs. at current address, yrs. with current employer, gender, no. of dep children, frequency paid, home phone no, Insurance premium, loan type (single/joint), marital status, term, home ownership and loan purpose	Age, amount, Yrs. at Current Address, Yrs. with Current Employer, Gender, No. of Dep. Children, Frequency paid, Home Phone Number Given, Insurance Premium, Loan type (single or joint), Marital Status, Term, Home Ownership & loan purpose	Income, age, housing, employment status, bureau score. Macroeconomics; Interest Rates (IR), Earnings, FTSE, Unemployment (Unemp), Production, House Price Index (House) &Consumer Confidence Index (CC)
Data Description	250 female & 250 males. Loans taken in Jan 2007. Loan term 30 months with observation period of 30 months	212,742 card customers approved between Jan 2003 to July 2008	50 000 UK loans accepted btwn June 1994 and March1997 Observed period up to July 1997	11,500 UK ***loan*** customers. Performance variables for 36 months used.	50,000 personal loan with term varying from 6 to 60 months Observation period - 36 months	50,000 loans data availed. Observation period of 36 months used. Missing values imputed using the mean & mode for the categorical attributes	Over 100,000 card accounts opened between 1997 to mid-2005.
Data collection (Sampling)	Random selection from a databank of personal loans in a Kenyan bank.	(1) All customers who defaulted (2) A 1% random sample of all who did not default	Sample split into 2 random groups. 70% to build the systems and the remaining 30% used as a holdout sample.	Data split into 3 samples (2 training and 1 holdout) of approximately equal size.	Two separate LR models were built on the training sample.	15000 data with term less than 36. Randomly split into a training set (10000) and a test set (5000 observations).	Databank of personal loans in a Kenyan bank. Greater weight given to bad cases as data was skewed.
Analysis	The Product limit estimator used to asses survival probability for the gender based study & log rank test for comparison of survival curves.	(TDPH) β̂ estimated and compared with the PH and LR models	Cox PH model, parametric-exponential & Weibull baseline hazards compared with logistic regression scorecards (LR).	Training & holdout sample scored with the application score &performance of the 2 compared .Cox's PH Reg used to build behavioural scoring models then performance compared with LR.	Continuous variables categorized. Cox's PH compared with LR approach. Model Diagnostic Methods using various residuals compared. Time-dependent models (Harrel's test) used to overcome the restriction of PH	For the statistical approaches, LR & Cox PH model used while for the NN analyses, a variant of the approach suggested by Mani was adopted.	Cox PH model with TVCs contrasted with LR and Cox PH Model with & without macroeconomic variable. Macroeconomic variables interacted with application variables.
Tests/Output	Log rank test, Mean Survival Times	β̂ for TDPH, PH and LR, Median KS	proportional hazard assumption	LLR & ROC curves-discriminate between `good' and `bad'.	LLR,ROC Curves, match rate tables, residuals	Classification accuracy, Survival function, variables significance.	LLR-model uplift, Wald Statistic -significance of each coefficient.

Results	Mean survival time was 15 months for male applicant while female would take on averagely 16 months to default. Log rank test not significant	TDPH model fitted the data well. The LR model using the TDPH as an additional predictor achieved roughly a 3.2% improvement in KS Statistic over regular LR and PH .	The PH models (exponential, Weibull & Cox's) competitive with the LR in identifying those who default in the first year, and may be superior to LR for looking at who will pay-off early in the first year. Loan amount , Refinance purpose & employment greatly effect on default.	Early default not predicted by application score. ROC curve analysis showed that (PHAB) scores competitive with LR scores, especially after about 2 years into the loan. Profit increases as application score goes from high to low risk. Profit curves cross for different terms of loans of similar amounts.	Segmentation by term (24, 36 and 48 months) had less effect in predicting default in 1st & 2nd yrs. Harrel's Z-test suggested possible time-by covariate interactions. Time-by-characteristic interaction adds flexibility dimension— to reflect an increase or decrease of the effect of a characteristic with the age of the loan	Differences in classifiers for LR, Cox & NN not statistically significant in predicting 1st yr default. Btwn 12 & 24 months, NN was superior yielding a classification accuracy of 78.58%, LR classifier - 78.24% & Cox 77.5%. Performance difference btwn NN and the Cox model was significant.	All the variables were significant to default, all having positive correlation except Earnings and Production. Interacting the application variables with macroeconomi variables show significance. The Cox PH model with macroeconomic variables outperforms LR.
Conclusion on Variables	It was not meaningful to classify borrowers on the basis of gender as this did not affect credit risk.	Incorporating the TDPH into either the LR or the PH approaches improves the predictive performance of these methods.	Amount of loan (8k+), Refinance purpose and employment (17+ yrs) had the highest significance to default.	Application characteristics i.e. Amount,term, Months with current employer, Months at current address, Net income, Living with parents/renting, Married, High risk purposes(such as refinance)and Low risk occupation code were found to be highly significant to default. Behavioural variables i.e balance differences also affect default and are used to identify early defaulters.	Default rate is independent of term of the loan but that early repayment also takes into account how much longer the loan is to exist and how much more would be needed to pay it off now.	Years employed, purpose of loan and insurance premium were found to be most significant in predicting default.	Both the macroeconomic variables (IR), Earnings, FTSE, (Unemp),(Prod), (House) and (CC) and interaction with the application variables Income, age, housing and employment status, bureau score are significant to default.
Overall Conclusion	Gender does not affect credit risk however Mean survival times would guide underwriting on average age for loans so as to minimize losses emanating from loan defaults thereby optimizing profit returns.	This result is consistent with the results of Teng et al (2007) and Bellotti and Crook (2009), who found modest but statistically significant improvements in predictive performance using macroeconomic variables with substantial dynamic variability	The PH models were competitive with LR approach in identifying those who default in the first year and more sophisticated models required for 2nd yr where there were fewer defaulters. Credit risk to be measured on loan amount, employment history and purpose of loan.	At early stages behaviour variables are more important while later stages application information is more predictive. Loan profits can be estimated, which is an important addition to scoring techniques since ***lenders are now moving from scoring only for risk to scoring for profitability.*** PHAB models competitive with LR model.	COX PH competitive with LR approach in scoring. It's not meaningful to purely classify borrowers on the basis of term of loan as term doesnt affect default. Cox-Snell residuals show that model fits data well.	For predicting default, the superiority of the neural network model was somewhat less pronounced. It was found that, for early repayment, the suggested neural network approach outperformed the proportional hazards model.	These results demonstrate that survival analysis is competitive in comparison with logistic regression as a credit scoring method for prediction. The inclusion of macroeconomic variables gives a statistically significant improvement in predictive performance.

Recommendation for Policy	Mean survival times would guide underwriting on average age for loans so as to minimize losses emanating from loan defaults thereby optimizing profit returns.	TDPH provides an inherent mechanism for adjusting the customer acceptance threshold to keep constant the collective default rate of accepted customers in the face of dynamic market conditions.	Credit risk to be measured on loan amount, employment history and purpose of loan.	Application & behaviour information complement each other over time. Banks to look at both term and behaviour score when ranking loans of similar amount since profits increase as default risk decrease.	Time-by-characteristic interactions suggested by Harrel's test were included in the model. This extension allows the effect of a covariate on the predicted time-to-failure to increase or decrease as the loan evolves.	Credit risk to be measured on Years employed, purpose of loan and insurance premium since they are most significant in predicting default.	This method of estimation makes this model suitable for stress testing by including macroeconomic conditions that simulate a depressed or booming economy
Recommendation for further studies	KM is a univariate method, it may be more informative to adopt multivariate techniques like Cox model to model credit risk. Thus further research could be conducted on the same data set using other survival techniques.		To overcome the fixed for all time risk ordering of PH, it was proposed to allow the b-coefficients to be time dependent & use accelerated life models with other distribution families than the Weibull one. Development of multi-stage models in survival analysis to deal with these problems could also prove useful in the credit scoring context or consideration of Bayesian analysis for the PH Model.	The profit formula can be altered to include time-dependent interest rates thus incorporate economic conditions into the model. Alternatively interest rates can be included as a covariate when estimating survival function.		Survival analysis models suffer from a number of drawbacks: the functional form of the inputs remains linear or some mild extension thereof, non-linearity & interaction effects must be explicitly modeled by the statistician, and in the standard PH Model, the baseline hazard function is assumed to be uniform and proportional. Multilayer perceptron NN proposed as a solution to these problems since they are non-linear & universal approximators	Further application of these methods to other credit card and fixed loan. Although the analysis of the explanatory model gives an understanding of how each macroeconomic variable contributes to modeling the data, further extensive experimental work is required to determine the effect of each of the macroeconomic variables on the prediction of Probability of Default.

3.3 SUMMARY OF STUDY VARIABLES

Narain (1992) in *Survival analysis and the credit granting decision* was one of the first authors to use survival analysis method to develop credit scoring models. She compared the performance of traditional Logistic Regression approach to that of survival analysis approaches (exponential regression model) and found out the exponential regression model estimated the number of failures at each failure time well and better credit-granting decision could be made if the score was supported by the estimated survival times.

The model has further been developed by (*Banasik, Crook, & Thomas, 1999); (Stepanova M and Thomas LC (2001); (Stepanova & Thomas, 2002); (Baesens, Van Gestel, Stepanova, Van den Poel, & Vanthienen, 2005); (Bellotti & Crook, 2007); (Im, Ã, Qi, & Shan, 2012); (Argan, Corresponding, Samuel, & Peter, 2012)* who found out that survival analysis methods are competitive with, and sometimes superior to, the traditional logistic regression approach. A common feature of all these papers is that they use parametric, non-parametric or semi-parametric regression techniques for modelling the time to default.

(Banasik, Crook, & Thomas, 1999) find that amount of loan (8k+), Refinance purpose and employment (17+ yrs.) had the highest significance to default.

***(Baesens, Van Gestel, Stepanova, Van den Poel, & Vanthienen, 2005*)** found that years employed, purpose of loan and insurance premium were the most significant in predicting default.

Stepanova M and Thomas LC (2001) found that application characteristics i.e. amount of loan, term, Months with current employer, Months at current address, Net income, Living with parents/renting, Married, High risk purposes(such as refinance) and Low risk occupation code were highly significant to default. Behavioural variables i.e. balance differences also affect default and are used to identify early defaulters.

(Stepanova & Thomas, 2002) - Default rate is independent of term of the loan.

(Bellotti & Crook, 2007) - Both the macroeconomic variables i.e. Interest Rates (IR), Earnings, FTSE, Unemployment, Production, House Price Index and Consumer Confidence Index and their interaction with the application variables i.e. income, age, housing and employment status, bureau score are significant to default.

(Argan, Corresponding, Samuel, & Peter, 2012) found out that it is not meaningful to classify borrowers on the basis of gender as gender does not affect credit risk.

CHAPTER 4: METHODOLOGY

4.1 INTRODUCTION

In this study, time to default is the event of interest. This section gives the description of the data used, variables obtained, data grouping, competing risks and the methods used to achieve the objective

4.2 DATA DESCRIPTION

Sample of personal loan application data approved and disbursed in 2010 was obtained. The data set consisted of 1,712 personal loans customers whose repayment terms varied from 11 to 72 months together with their repayment status during the observation period of up to May 2013. The repayment status variable observed whether the customers defaulted, paid off their loans early, paid their loans to term or the loan was still open at the end of the observation period.

In this study, a customer is said to have defaulted if they have missed more than 90 days in repaying the amounts due to the bank as per the regulatory guidelines.

4.3 VARIABLES

The following 7 application variables were selected as the most likely to affect default by field experts and were made available for the data set;

1. Age of customer
2. Customer gender
3. Term of loan
4. Income
5. Commitments
6. Banking History

Age of customer, income and banking history were considered continous variable while gender, term of loan and Commitments were categorical variables.

4.4 DATA GROUPING

Compared to the traditional approaches like logistic regression, survival analysis main aim is to be predictive rather than descriptive and as such, the continous variables such as customers' age, commitments, term of loan were split into attributes to form categories that well define them based on expert judgement to ensure that the analysis was robust. Binary variables were then created for each category.

4.5 COMPETING RISKS

In the data set, there are 2 competing events that have been identified to be affect survival time i.e. default and early repayment. Early repayment could occur as a result of a customer requesting the bank to top-up his/her loan meaning the old account is closed and new one opened and therefore observation ends for that particular account or another bank buying off the customers' loan resulting into the closure of the loan account before the expiry of the contractual term.

Let T be the survival time in the study, T_1 to be the time until default and T_2 to be the time to early repayment/closure of the account then the predicted life time of the loan is $T = \min(T_1, T_2, term\ of\ loan)$. In our data the minimum for the competing risk is T_1 since customers who experience both events would automatically default first before they clear their arrears and the account is finally closed.

4.6 DATA ANALYSIS

Both exploratory and informatory data analysis were used on the data set.

The non-parametric Kaplan-Meier curves - was the first step to provide insight into the shape of the survival function for each of the categorical predictor. This step was to provide insight into the shape of the survival functions for each category and give preliminary indication of whether the groups are proportional or not, that is, by checking whether the survival functions are parallel to each other or not before pursuing into any complex model.

Tests of equality using the non-parametric log rank and Wilcoxon test across strata were also explored to consider whether or not to include the categorical predictors in the final model.

For the continuous variables with the very many different levels involved, it was not realistic to calculate the Kaplan-Meier curves as each predictor level would have its own curve and instead the Cox proportional hazard model with a single continuous predictor variable was used to obtain the significance of the variable so as to consider whether or not to include the continuous predictors in the final model.

4.7 PROPORTIONALITY ASSUMPTION TEST

In this study, proportionality will be checked by including **time-dependent covariates** (already described) in the model. Time dependent covariates are interactions of the predictors with time. The variables are interacted with log (survtime) because this is the most common function of time used in time-dependent covariates

4.8 FINAL MODEL

All the predictor values whose p-values were considered significant in the univariate analysis were considered in the final model and possible interactions considered for the significant variables.

CHAPTER 5: RESULTS AND DISCUSSIONS

This chapter gives the results of both the exploratory and confirmatory analysis already described in chapter 4. MS Excel 2010 was used for exploratory analysis while SAS 9.2 was used for confirmatory analysis.

The discussion of the results follows after.

5.1 EXPLORATORY RESULTS

5.1.1 Analysis by Gender

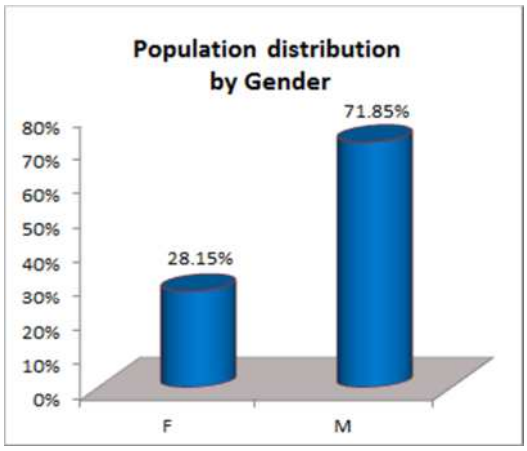


Figure 5. 1

GENDER			
STATUS	**FEMALE**	**MALE**	**GRAND TOTAL**
CENSORED	445	1120	1565
DEFAULTED	37	110	147
Grand Total	**482**	**1230**	**1712**

Table 5. 1

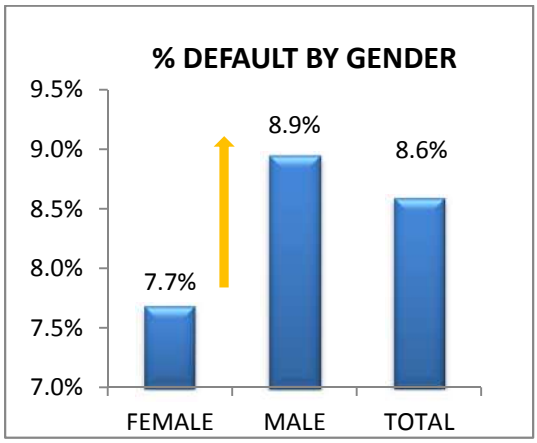


Figure 5. 2

- Figure 5.1 shows that bookings concentrated more on male than female, with male comprising 72% of all loans booked.
- Figure 5.2 shows that male seem to have higher default rate than female. The difference is to be established by confirmatory analysis.

5.1.2 Analysis by Income

Income is a continous variable and various categories created to compare performance.

Where low ≤ 100k, middle (100-300k), high >300k (All in KES Currency)

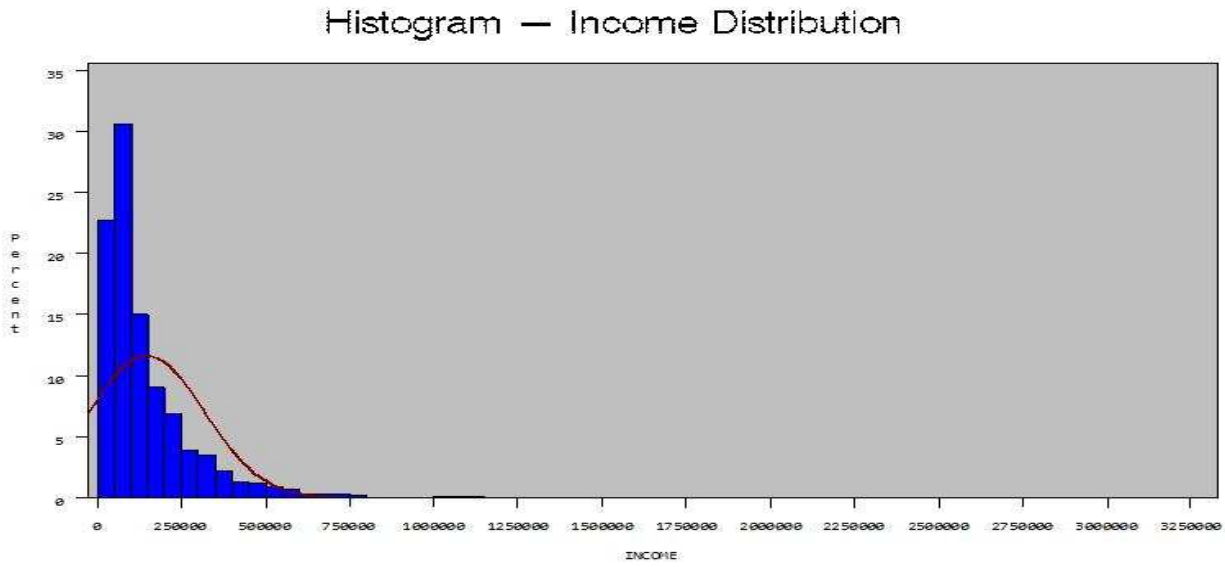


Figure 5. 3

INCOME SEGMENT (KES)				
STATUS	LOW	MIDDLE	HIGH	Grand Total
CENSORED	834	545	186	1565
DEFAULTED	89	48	10	147
Grand Total	**923**	**593**	**196**	**1712**

Table 5. 2

N	1712
Mean	147066.9
Median	90949.5
Mode	60000.0

Extreme Observations

-----Lowest----		-----Highest-----	
Value	Obs	Value	Obs
9580	1051	1212210	1048
11000	1052	1299200	1711
12617	325	1414304	1049
13276	1053	1625000	1050
13483	1054	3262995	1712

Quantile	Estimate
100% Max	3262995.0
99%	735648.0
95%	437500.0
90%	323453.0
75% Q3	179186.5
50% Median	90949.5
25% Q1	52508.0
10%	34703.0
5%	26307.0
1%	19117.0
0% Min	9580.0

Table 5. 3

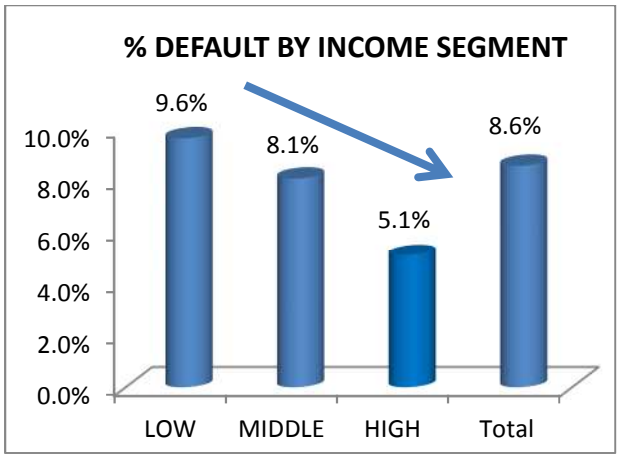


Figure 5. 4

Since the histogram is highly skewed because of few individuals earning very high salaries, a more robust stem & leaf and histogram for salaries ≤ 250,000 was fitted

STEM AND LEAF FOR FEMALE & MALE BASED ON INCOME ≤ 250000

Gender = F

```
Stem Leaf                                                    #            Boxplot
25 001                                                       3               0
24 0367                                                      4               0
23 01359                                                     5               0
22 005                                                       3               |
21 578                                                       3               |
20 0000001557789                                             13              |
19 166                                                       3               |
18 223678                                                    6               |
17 11779                                                     5               |
16 0001235689                                                10              |
15 00003445666778                                            14              |
14 1223777                                                   7               |
13 0022455577889                                             13              |
12 0000000012223344589                                       19              |
11 00112234557888889                                         17             +-----+
10 00001345555566779                                         17             |     |
9 000011123444556666667788888999                             29             |     |
8 000000000222366666677888899                                27             |  +  |
7 000000222222344444455667778899                             31             *-----*
6 000000000001111222222333445555667777889999                 42             |     |
5 000000000111111222233334444455555666666788888899           48             |     |
4 00000011112222233345555555666788899999                     38             +-----+
3 000011222222223334445555555666666777889999999             44                |
2 00000222223455555666678889                                 25                |
1 13455688899                                                11                |
----+----+----+----+----+----+----+----+----+---
Multiply Stem.Leaf by 10**+4
```

Gender = M

```
Histogram                                                    #            Boxplot
```

```
255000+**                                   5                |
        .*****                             14               |
        .******                            18               |
        .******                            18               |
        .******                            16               |
205000+********                            23              |
        .*****                             15               |
        .******                            16               |
        .******                            18               |
        .**********                        28               |
155000+**************                      40              |
        .**********                        30               |
        .*************                     39         +-----+
        .*************                     39         |     |
        .************                      35         |     |
105000+**************                      41         |     |
        .***************                   44         |  +  |
        .********************              58         |     |
        .*****************************     87         *-----*
        .*****************************     87         |     |
55000+************************             73         |     |
        .***********************************  103     +-----+
        .*******************************     91          |
        .**********************             64          |
        .****                              10          |
5000+*                                      1          |
        ----+----+----+----+----+----+----+
* may represent up to 3 counts
```

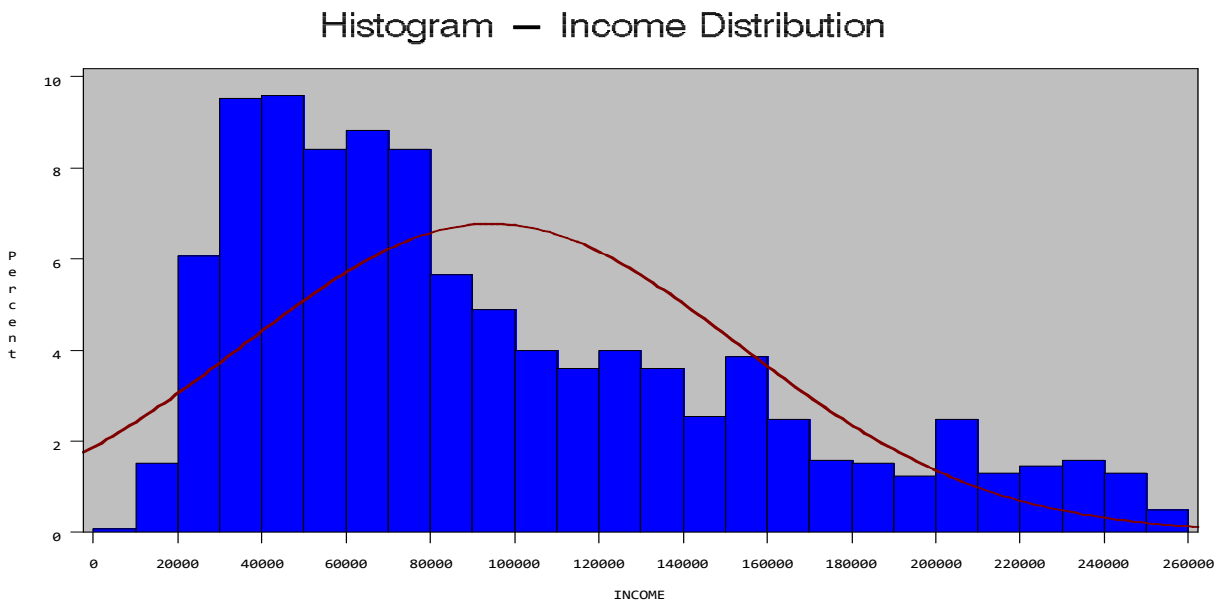


Figure 5. 5

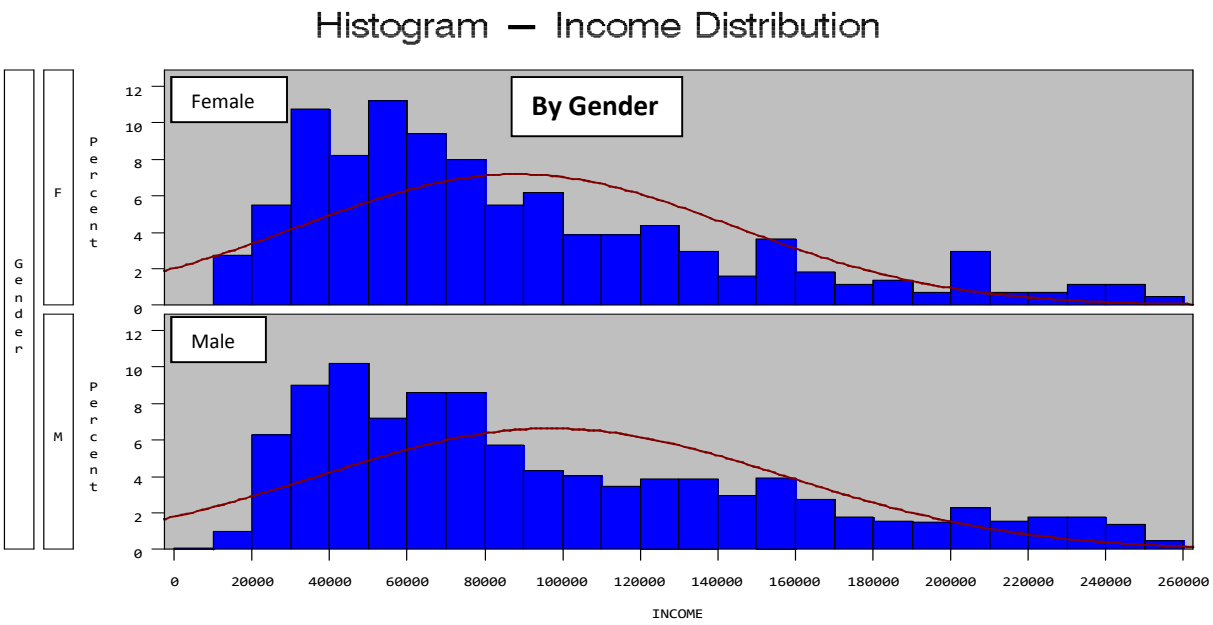


Figure 5. 6

Location	
N	1450
Mean	94451.00
Median	77000.00
Mode	60000.00

Figure 5.5 indicates that most of the customers given loans earn between KES. 30,000 and KES. 80,000. Income of KES 60,000 is the most common salary for the customers. This income distribution is also the same across gender as depicted by Figure 5.6. Mean for men is KES 97,000 with a mode of KES 60,000 and for women is KES 88,000 with a mode of KES 50,000. This is also the same group (<KES 100,000) whose default rate is higher compared to other groups as shown in Figure 5.4

5.1.3 Analysis by Commitments

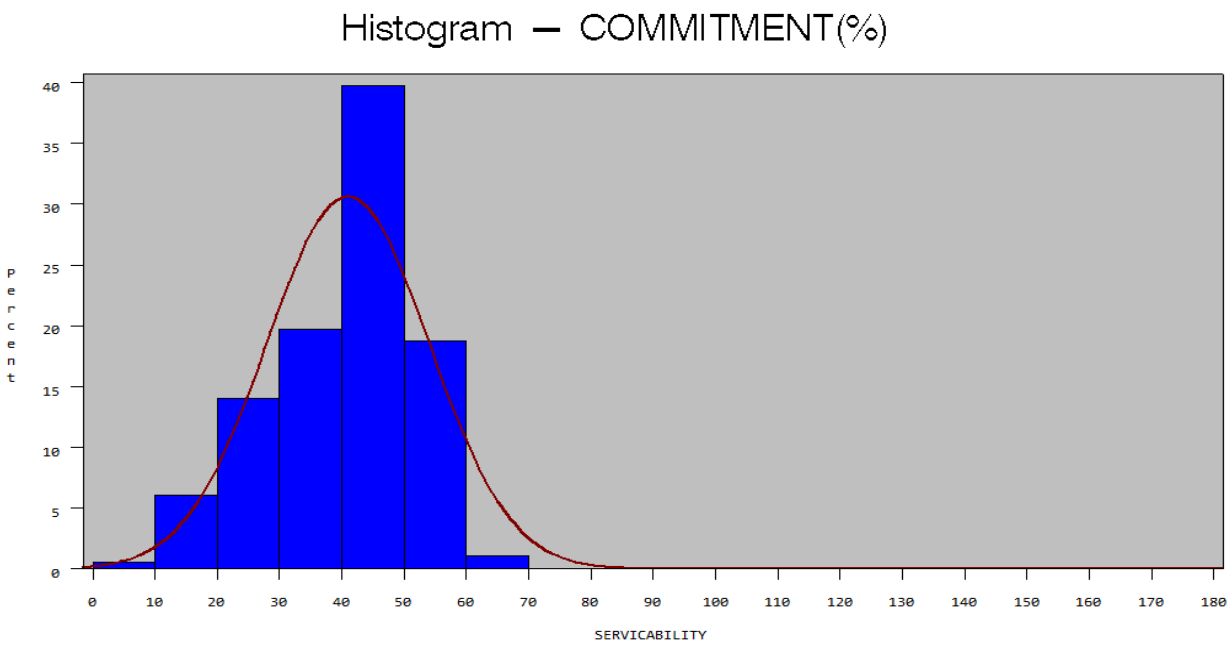


Figure 5. 7

N	1712
Mean	40.99065
Median	44.00000
Mode	49.00000

Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
0	1680	60	1459
0	1430	60	1511
0	474	60	1619
3	482	60	1646
3	169	172	309

Quantile	Estimate
100% Max	172
99%	60
95%	59
90%	58
75% Q3	49
50% Median	44
25% Q1	32
10%	23
5%	17
1%	11
0% Min	0

Table 5. 4

COMMITMENTS (%)					
STATUS	**≤ 30**	**31-40**	**41-50**	**>50**	**Grand Total**
CENSORED	349	332	606	278	1565
DEFAULTED	31	18	66	32	147
Grand Total	**380**	**350**	**672**	**310**	**1712**

Table 5. 5

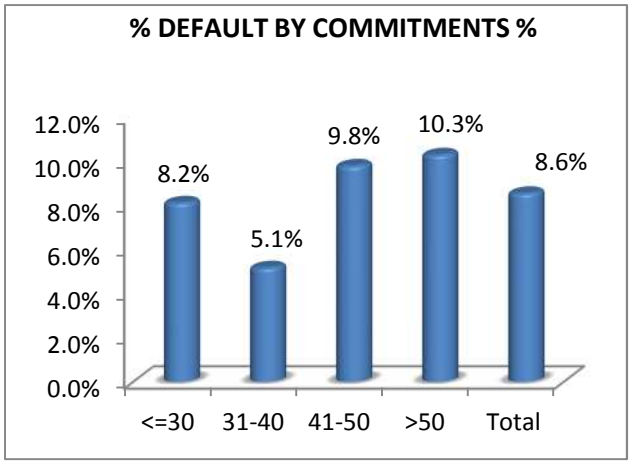


Figure 5. 8

Figure 5.7 and Table 5.4 indicate that most of the customers have committed between 40% and 50% of their salaries, the mode being 49% while default is higher in commitment >50% followed by 41%-50% category as shown in Figure 5.8

5.1.4 Analysis by Age

Age was analyzed as a continuous variable as well as categorical variable to compare differences.

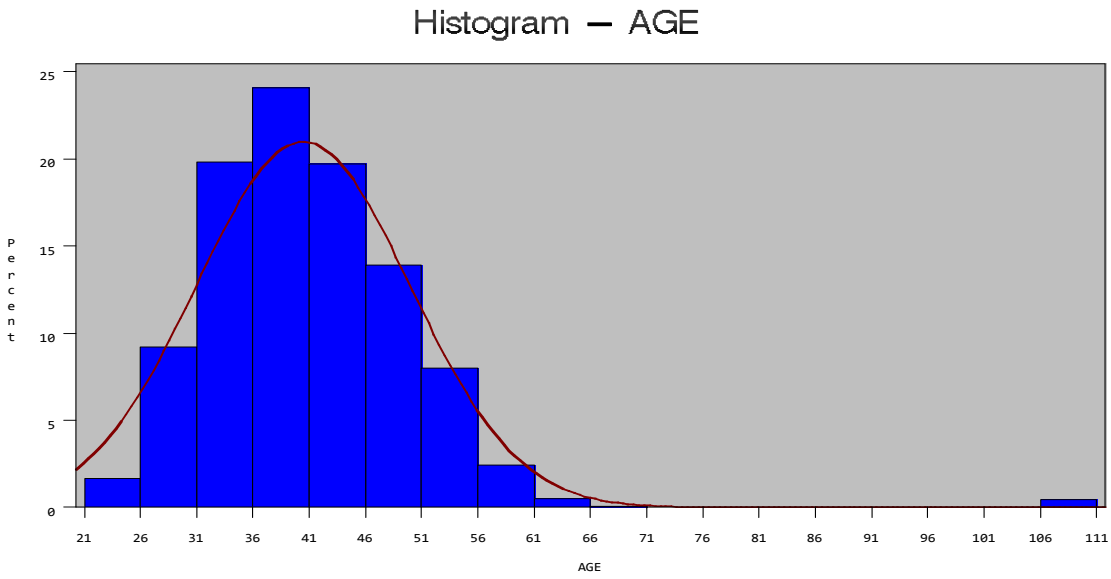


Figure 5. 9

LOCATION	
N	1712
Mean	40.46028
Median	40.00000
Mode	35.00000

Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
21	516	110	1506
21	229	110	1639
23	404	110	1675
23	349	111	1205
23	236	111	1659

Quantile	Estimate
100% Max	111
99%	62
95%	54
90%	51
75% Q3	46
50% Median	40
25% Q1	34
10%	30
5%	28
1%	25
0% Min	21

Table 5. 6

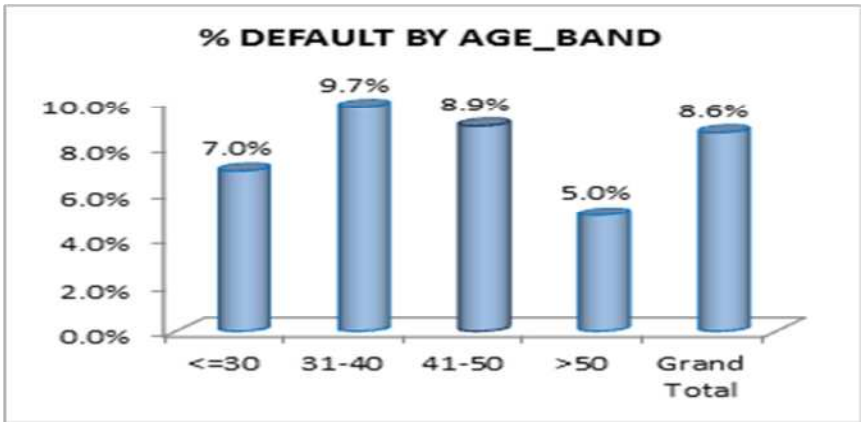


Figure 5. 10

	AGE_BAND				
STATUS	**≤ 30**	**31-40**	**41-50**	**>50**	**Grand Total**
CENSORED	174	678	524	189	1565
DEFAULTED	13	73	51	10	147
Grand Total	**187**	**751**	**575**	**199**	**1712**

Table 5. 7

Figure 5.9 and **Table 5.6** indicate that most of the customers are between the age group 31- 45 with 40 being the mode and the mean. Default is also higher in the age band 31-40 as shown in **Figure 5.10**

5.1.5 Analysis by Term

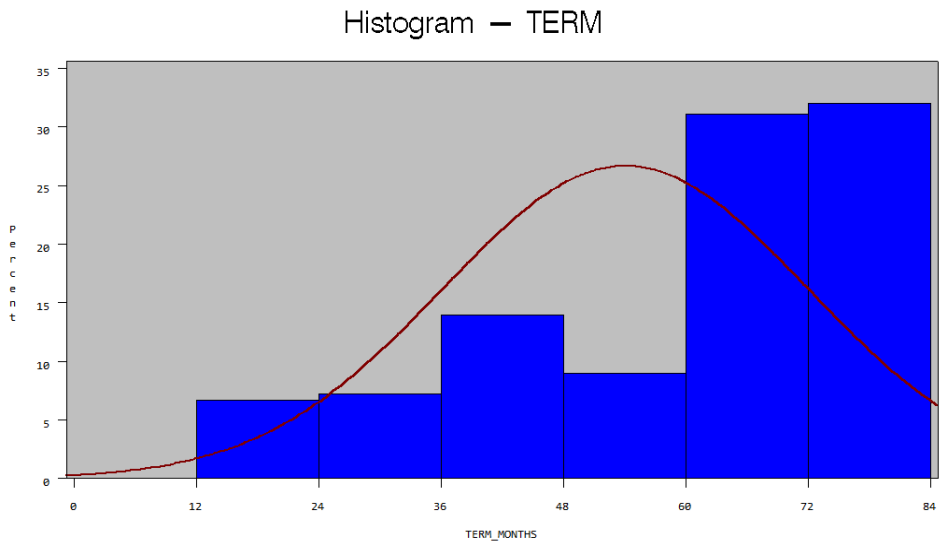
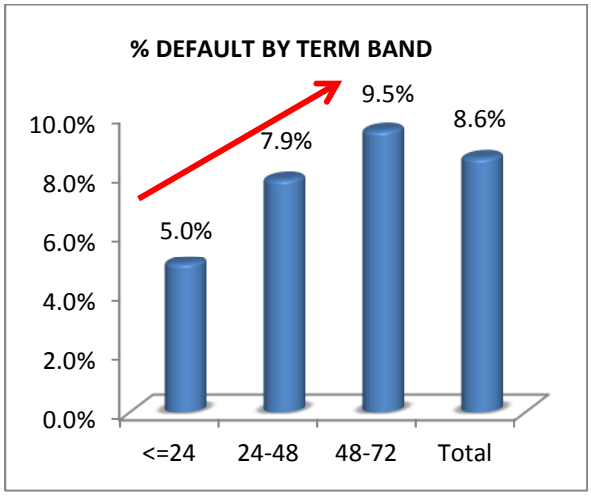


Figure 5. 11

LOCATION		Extreme Observations				Quantile	Estimate
N	1712	----Lowest----		----Highest---		100% Max	72
Mean	54.09229					99%	72
Median	60.00000					95%	72
Mode	72.00000	Value	Obs	Value	Obs	90%	72
						75% Q3	72
		11	1091	72	1679	50% Median	60
		12	1699	72	1692	25% Q1	36
		12	1698	72	1695	10%	24
		12	1686	72	1697	5%	18
		12	1684	72	1700	1%	12
						0% Min	11

Table 5. 8



TERM				
STATUS	**≤ 24**	**24-48**	**48-72**	**Total**
CENSORED	190	375	1000	1565
DEFAULTED	10	32	105	147
Total	**200**	**407**	**1105**	**1712**

Table 5. 9

Figure 5. 12

Figure 5.11 indicates that the highest term of loan in the data is 72 months. Most of the customers are given loans with term of 72 followed closely by 60 months. Default seems to increase as term of loan increases as shown in Figure 5.12

5.1.6 Analysis by Banking History

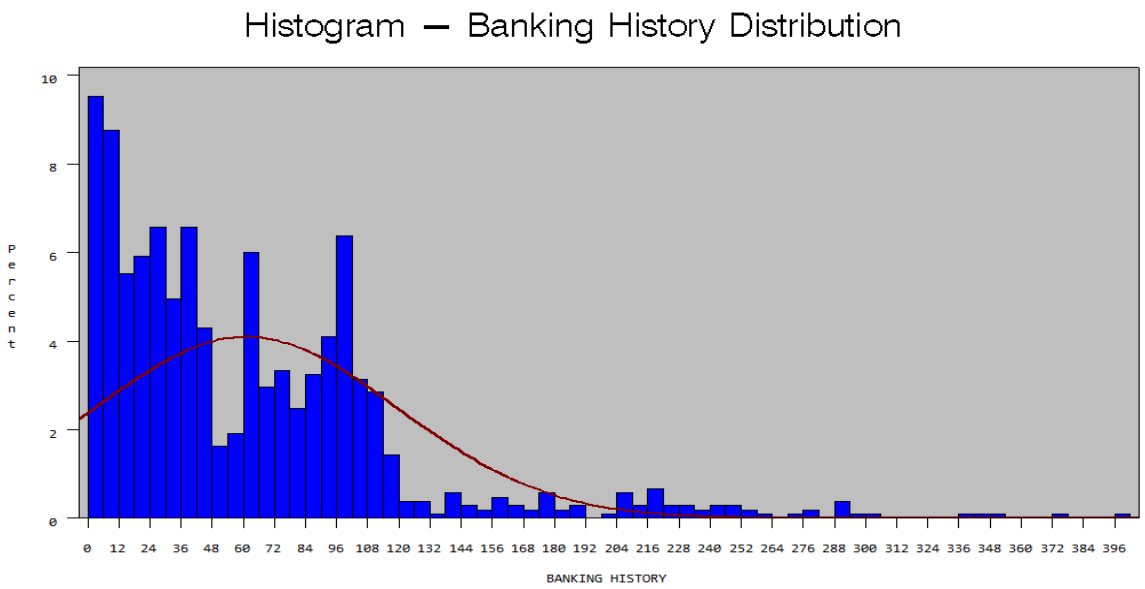


Figure 5. 13

LOCATION				Quantile	Estimate
N	1050			100% Max	399
Mean	61.27714			99%	290
Median	44.00000			95%	180
Mode	1.00000			90%	113
		Extreme Observations		75% Q3	91
		----Lowest----	----Highest---	50% Median	44
		Value Obs	Value Obs	25% Q1	19
		0 98	341 1044	10%	6
		0 96	347 677	5%	1
		0 90	351 936	1%	0
		0 85	373 876	0% Min	0
		0 80	399 779		

Table 5. 10

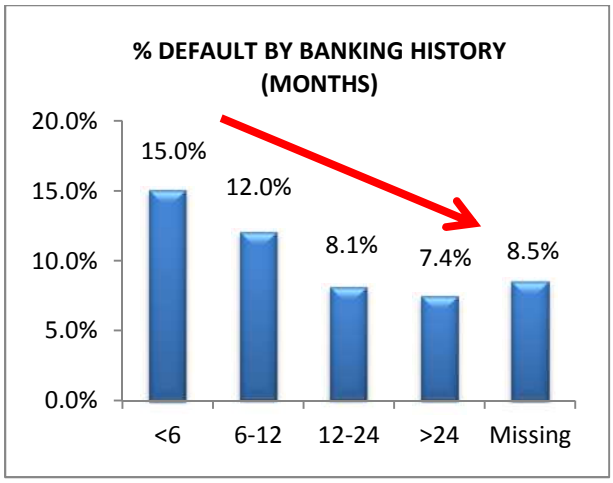


Figure 5. 14

BANKING HISTORY (MONTHS)						
STATUS	**<6**	**6-12**	**12-24**	**>24**	**Missing**	**Grand Total**
CENSORED	85	88	114	672	606	1565
DEFAULTED	15	12	10	54	56	147
Grand Total	**100**	**100**	**124**	**726**	**662**	**1712**

Table 5.1 1

Figure 5.13 indicate that most of the customers have banking history less than 12 months with the majority in this group having banked for just 1 month. Default increases as banking history declines as shown in Figure 5.14

5.2 CONFIRMATORY RESULTS

5.2.1 Analysis by Gender

LEGEND: 1=Female 2=Male

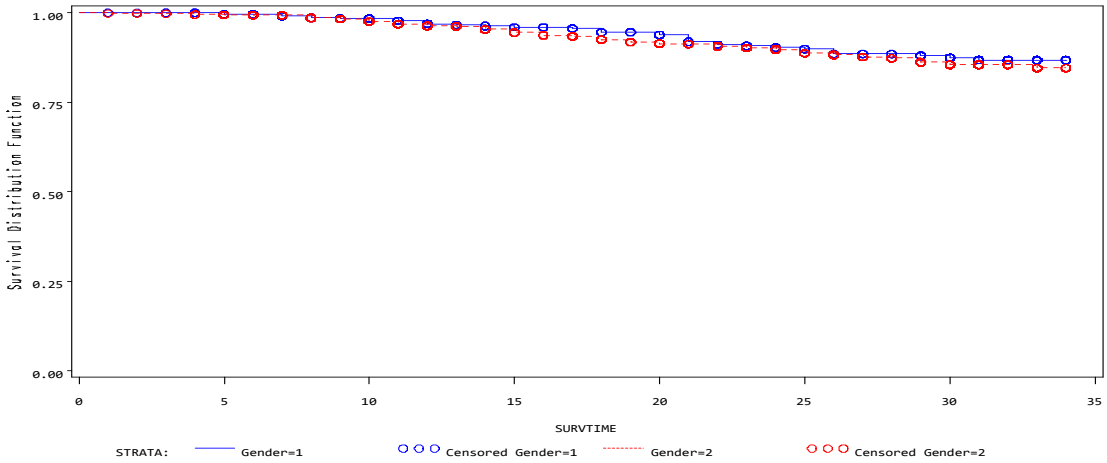


Figure 5. 15

Stratum 1: Gender = 1

Product-Limit Survival Estimates

SURVTIME	Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.0000	1.0000	0	0	0	482
1.0000*	.	.	.	0	481
1.0000*	.	.	.	0	480
2.0000*	.	.	.	0	479
3.0000*	.	.	.	0	478
3.0000*	.	.	.	0	477
3.0000*	.	.	.	0	476
4.0000*	.	.	.	0	475
4.0000*	.	.	.	0	474
4.0000*	.	.	.	0	473
4.0000*	.	.	.	0	472
5.0000	.	.	.	1	471
.					
31.0000	0.8683	0.1317	0.0211	37	126

.					
33.0000*	.	.	.	37	11
34.0000*	.	.	.	37	10
34.0000*	.	.	.	37	9
34.0000*	.	.	.	37	8
34.0000*	.	.	.	37	7
34.0000*	.	.	.	37	6
34.0000*	.	.	.	37	5
34.0000*	.	.	.	37	4
34.0000*	.	.	.	37	3
34.0000*	.	.	.	37	2
34.0000*	.	.	.	37	1
34.0000*	.	.	.	37	0

The LIFETEST Procedure

NOTE: The marked survival times are censored observations.

Summary Statistics for Time Variable SURVTIME

Quartile Estimates

Percent	Point Estimate	95% Confidence Interval [Lower	Upper)
75	.	.	.
50	.	.	.
25	.	.	.

Mean	Standard Error
29.4708	0.2589

NOTE: The mean survival time and its standard error were underestimated because the largest observation was censored and the estimation was restricted to the largest event time.

Stratum 2: Gender = 2

Product-Limit Survival Estimates

SURVTIME	Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.0000	1.0000	0	0	0	1230
1.0000	.	.	.	1	1229
1.0000	0.9984	0.00163	0.00115	2	1228
1.0000*	.	.	.	2	1227
2.0000*	.	.	.	2	1226
2.0000*	.	.	.	2	1225
3.0000	0.9976	0.00244	0.00141	3	1224

3.0000*	.	.	.	3	1223
3.0000*	.	.	.	3	1222
4.0000	.	.	.	4	1221
.					
33.0000	0.8462	0.1538	0.0160	110	94
.					
.					
34.0000*	.	.	.	110	5
34.0000*	.	.	.	110	4
34.0000*	.	.	.	110	3
34.0000*	.	.	.	110	2
34.0000*	.	.	.	110	1
34.0000*	.	.	.	110	0

NOTE: The marked survival times are censored observations.

Summary Statistics for Time Variable SURVTIME

Quartile Estimates

Percent	Point Estimate	95% Confidence Interval [Lower	Upper)
75	.	.	.
50	.	.	.
25	.	.	.

Mean	Standard Error
30.9033	0.1980

NOTE: The mean survival time and its standard error were underestimated because the largest observation was censored and the estimation was restricted to the largest event time.

Summary of the Number of Censored and Uncensored Values

Stratum	Gender	Total	Failed	Censored	Percent Censored
1	1	482	37	445	92.32
2	2	1230	110	1120	91.06
Total		1712	147	1565	91.41

Testing Homogeneity of Survival Curves for SURVTIME over Strata

Test of Equality over Strata			
Test	Chi-Square	DF	Pr>Chi-Square
Log-Rank	0.5970	1	0.4397
Wilcoxon	0.6964	1	0.4040
-2Log (LR)	0.6421	1	0.4230

Table 5.12

5.2.2 Analysis by Income

Income is a continuous variable and various categories were created after grouping to compare performance.

Legend: **1: LOW (≤ 100K);** **2: MIDDLE (100-300K);** **3: HIGH (>300K)**

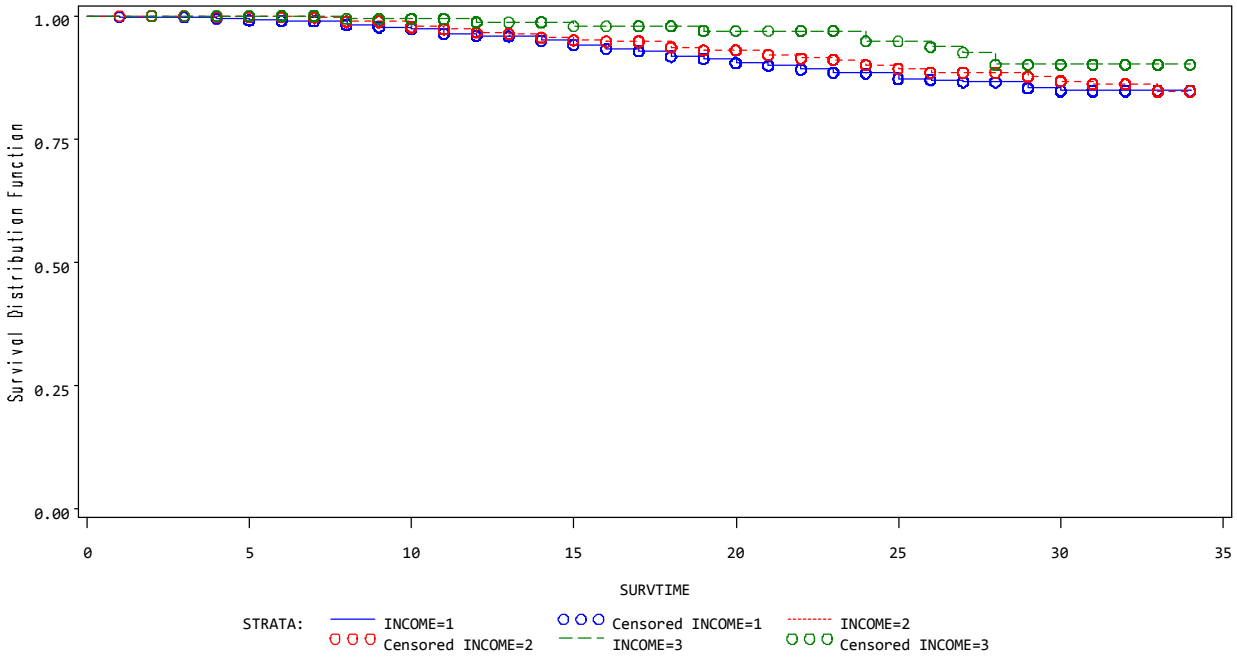


Figure 5. 16

The LIFETEST Procedure

Summary of the Number of Censored and Uncensored Values

Stratum	INCOME	Total	Failed	Censored	Percent Censored
1	1	923	89	834	90.36
2	2	593	48	545	91.91
3	3	196	10	186	94.90
Total		1712	147	1565	91.41

Test of Equality over Strata			
Test	Chi-Square	DF	Pr>Chi-Square
Log-Rank	4.3086	2	0.1160
Wilcoxon	6.2107	2	0.0448
-2Log (LR)	4.8686	2	0.0877

Table 5.13

Analysis of Income as a continous variable

The PHREG Procedure

Model Information

Data Set	WORK.MASTERS
Dependent Variable	SURVTIME
Censoring Variable	STATUS
Censoring Value(s)	0
Ties Handling	BRESLOW

Number of Observations Read	1712
Number of Observations Used	1712

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored

1712	147	1565	91.41

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	2027.990	2020.979
AIC	2027.990	2022.979
SBC	2027.990	2025.970

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	7.0110	1	0.0081
Score	4.6398	1	0.0312
Wald	5.3168	1	0.0211

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Income	1	-1.6756E-6	7.26665E-7	5.3168	0.0211	1.000

5.2.3 Analysis by Commitments

Legend: 1 :≤ 30%; 2: (31-40)% 3: (41-50)% 4: >50%

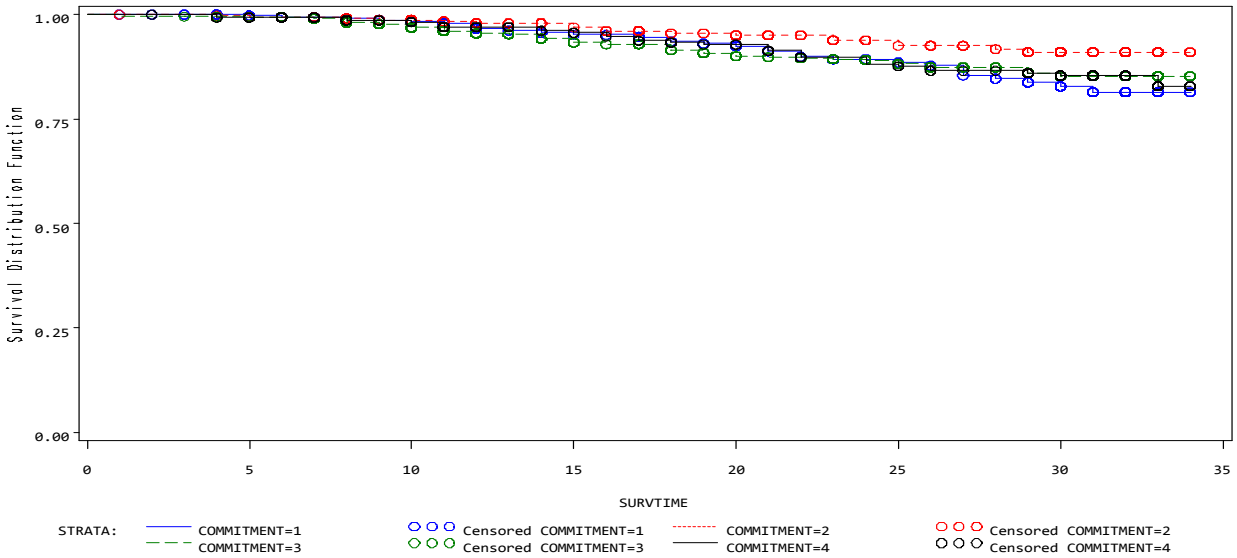


Figure 5. 17

Summary of the Number of Censored and Uncensored Values

Stratum	COMMITMENT	Total	Failed	Censored	Percent Censored
1	1	380	31	349	91.84
2	2	350	18	332	94.86
3	3	672	66	606	90.18
4	4	310	32	278	89.68
Total		1712	147	1565	91.41

Testing Homogeneity of Survival Curves for SURVTIME over Strata

Test of Equality over Strata			
Test	Chi-Square	DF	Pr>Chi-Square
Log-Rank	5.5376	3	0.1364
Wilcoxon	5.0223	3	0.1702
-2Log (LR)	6.2694	3	0.0992

Table 5.14

5.2.4 Analysis by Age

Age was analyzed as a continuous variable as well as categorical variable to compare differences

Legend: 1 :≤ 30; 2: 31-40 3: 41-50 4: >50

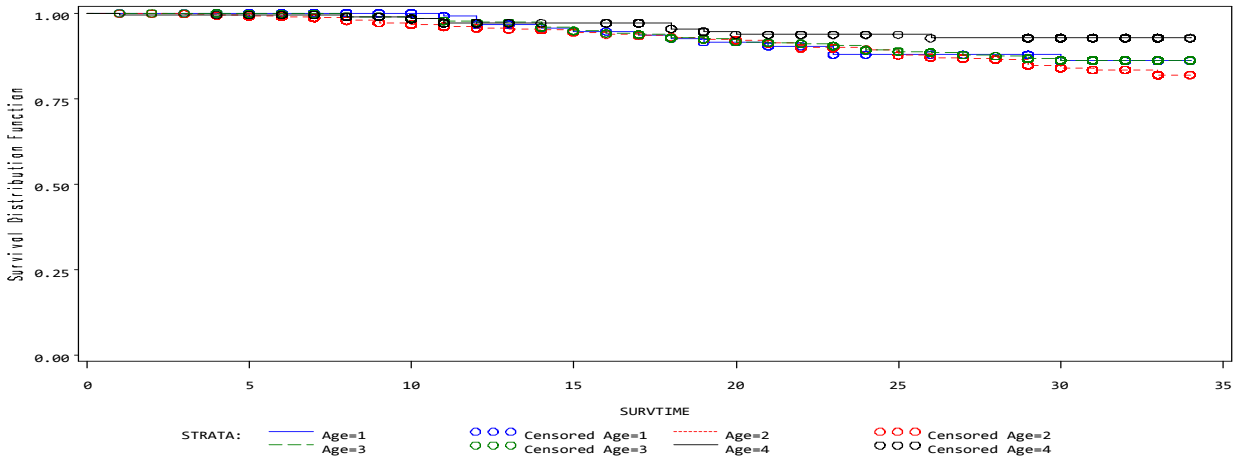


Figure 5. 18

The LIFETEST Procedure

Summary of the Number of Censored and Uncensored Values

Stratum	Age	Total	Failed	Censored	Percent Censored
1	1	187	13	174	93.05
2	2	751	73	678	90.28
3	3	575	51	524	91.13
4	4	199	10	189	94.97
Total		1712	147	1565	91.41

Test of Equality over Strata			
Test	Chi-Square	DF	Pr>Chi-Square
Log-Rank	5.9598	3	0.1136
Wilcoxon	5.0467	3	0.1684
-2Log (LR)	6.1131	3	0.1062

Table 5.15

ANALYSING OF AGE AS A CONTINOUS VARIABLE

Model Information

Data Set	WORK.MASTERS
Dependent Variable	SURVTIME
Censoring Variable	STATUS
Censoring Value(s)	0
Ties Handling	BRESLOW

Number of Observations Read	1712
Number of Observations Used	1712

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
1712	147	1565	91.41

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	2027.990	2025.373
AIC	2027.990	2027.373
SBC	2027.990	2030.364

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2.6168	1	0.1057
Score	2.3587	1	0.1246
Wald	2.4115	1	0.1204

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Age	1	-0.01522	0.00980	2.4115	0.1204	0.985

5.2.5 Analysis by Term

Legend: 1: ≤ 24 2: 24-48 3:48-72

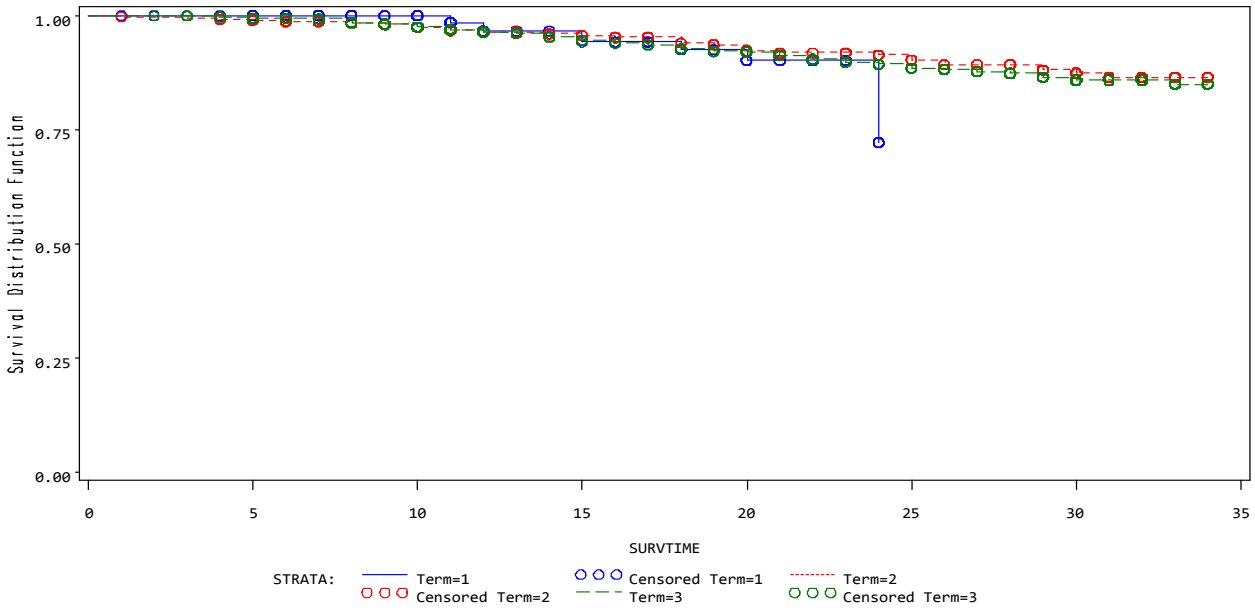


Figure 5. 19

Stratum	Term	Total	Failed	Censored	Percent Censored
1	1	200	10	190	95.00
2	2	407	32	375	92.14
3	3	1105	105	1000	90.50
Total		1712	147	1565	91.41

Test of Equality over Strata			
Test	Chi-Square	DF	Pr>Chi-Square
Log-Rank	0.3424	2	0.8427
Wilcoxon	0.1622	2	0.9221
-2Log (LR)	0.5346	2	0.7655

Table 5.16

5.2.6 Analysis by Banking History

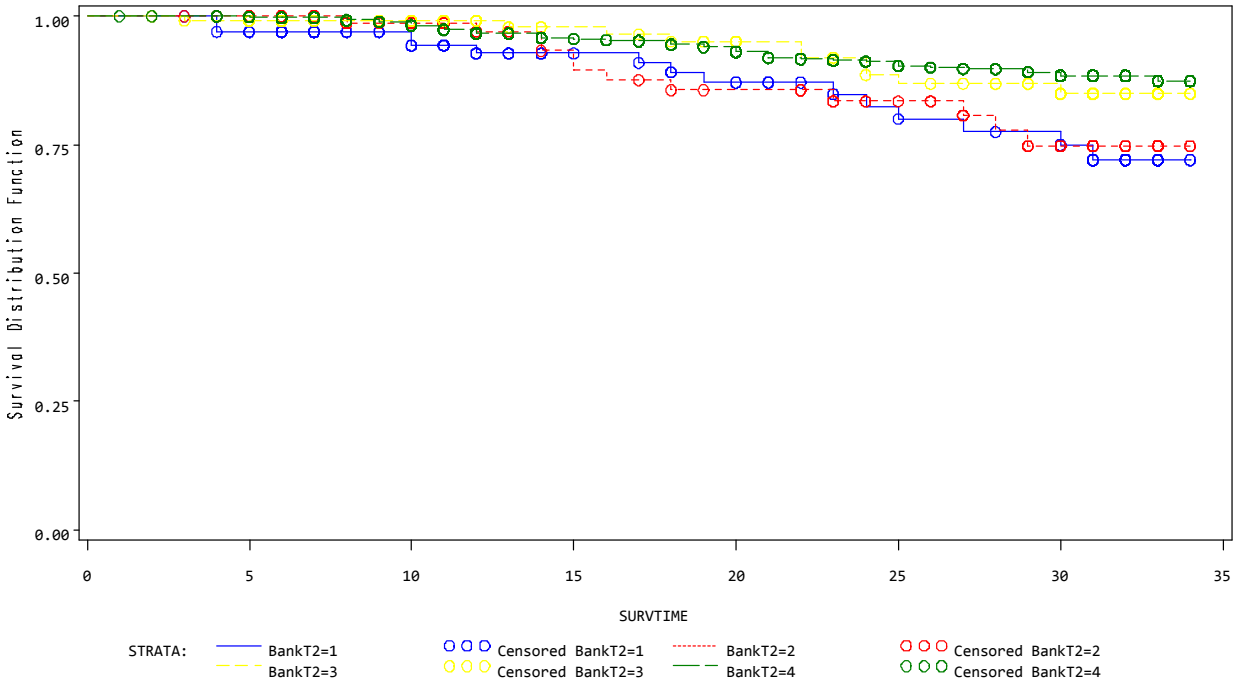


Figure 5. 20

Summary of the Number of Censored and Uncensored Values

Stratum	Banking History	Total	Failed	Censored	Percent Censored
1	1	100	15	85	85.00
2	2	100	12	88	88.00
3	3	124	10	114	91.94
4	4	726	54	672	92.56
Total		1050	91	959	91.33

NOTE: There were 662 observations with missing values, negative time values or frequency values less than 1.

Test of Equality over Strata			
Test	Chi-Square	DF	Pr>Chi-Square
Log-Rank	12.9076	3	0.0048
Wilcoxon	11.4502	3	0.0095
-2Log (LR)	9.3543	3	0.0249

Table 5.17

5.2.7 Final Model

From the above results, only income and banking history have p-values < 0.05 hence were confirmed to be significant in predicting loan default. Therefore the final model only had income and banking history as shown in **Figure 5.21**

Analysis of Maximum Likelihood Estimates

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
MID_INCOME	1	-0.20789	0.18044	1.3274	0.2493	0.812
HIGH_INCOME	1	-0.70509	0.33407	4.4547	0.0348	0.494
BANKINGL_6	1	0.83492	0.27674	9.1020	0.0026	2.305
BANKING6_12	1	0.67735	0.30504	4.9307	0.0264	1.969
BANKING12_24	1	0.08747	0.33079	0.0699	0.7915	1.091

Linear Hypotheses Testing Results

Label	Wald Chi-Square	DF	Pr > ChiSq
INCOME	5.0332	2	0.0807
BANKING_HISTORY	12.7005	3	0.0053

Figure 5. 21

5.2.8 Model Interaction

Income was interacted with banking history to check whether there was significance. No interaction is noted as evidenced in Figure 5.22

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
INCOME	1	-0.24998	0.46152	0.2934	0.5881	0.779
BankT2	1	-0.30666	0.23412	1.7157	0.1902	0.736
INCOME_BANKHISTORY	1	-0.00986	0.13913	0.0050	0.9435	0.990

Figure 5. 22

5.2.9 Proportionality Assumption Test

The PHREG Procedure

Analysis of Maximum Likelihood Estimates

Variable	Parameter DF	Standard Estimate	Error	Chi-Square	Pr > ChiSq	Hazard Ratio
INCOME	1	-1.41910	0.98210	2.0879	0.1485	0.242
BankT2	1	-0.39129	0.48067	0.6627	0.4156	0.676
INCOMET	1	0.40754	0.34288	1.4128	0.2346	1.503
BANKT2T	1	0.02614	0.17270	0.0229	0.8797	1.026

Linear Hypotheses Testing Results

Label	Wald Chi-Square	DF	Pr > ChiSq
Proportionality_test	1.4144	2	0.4930

Figure 5.23

Since the P-value for proportionality test is not significant as shown in **Figure 5.23,** no time dependency was evident. Therefore our final model shown in **Figure 5.21** was adopted. The model adopts the use of Cox PH model with no time dependency covariate.

5.3 DISCUSSION

5.3.1 Gender

Aggregated default rate over 33 months is 8.9% for male and 7.7% for female (Figure 5.2) with the number of males included in the model being almost 3 times that of female (1230 male compared to 483 of female) - Figure 5.1. Over 91% of both genders are censored and this is attributed to early repayments and individuals whose term of loan exceeded 36 months.

The estimated survival function for female and male does not go below 0.8683 and 0.8462 respectively, which is why the 75%, 50% and 25% quantiles (percentiles) cannot be estimated for both genders.

Mean survival time for female was 29.47 months while male is 30.9 months meaning male survived longer. The mean is not simply the average of the observed survival times, since it must take account of censored observation. This is the reason why an indication is given that the mean survival time and its standard error were underestimated because the largest observation was censored and the estimation was restricted to the largest event time.

In SAS the mean is computed as $\sum_{j=1}^{r} \hat{S}(t_j)(t_j - t_{j-1})$, in which the summation is over the *r* ordered death times, $\hat{S}(t_j)$ is the Kaplan-Meier estimate of the survivor function at the *jth* death times, t_j , and t_0 is defined to be Zero.

In general, the location of survival data is better estimated using the median survival time, and so the output on mean survival time and its error can generally be ignored

The graphs of the survival curves by gender (**Figure 5.15**) appear to be very close together throughout the observation period indicating no difference between the groups. This expectation is confirmed by the Wilcoxon test **(Table 5.12)** for comparison of the gender survival curves that is non-significant with p-values of 0.4040 indicating that gender is not a risk ranking variable hence no need to classify borrowers with respect to gender. The log-rank test places more emphasis on the differences in the curves at longer time values.

This finding is also supported by the aggregated analysis where default rate for female is 7.7% while for male is closer at 8.9% even though male are 3 times as many as female (Figure 5.2).

This is also in line with the findings of (Argan, Corresponding, Samuel, & Peter, 2012) who used same sample size of female and male (250 each), compared the survival curves of the 2 groups and found the difference not to be significant and concluded that it was not meaningful to classify borrowers on the basis of gender as this did not affect credit risk.

Gender will thus not be included as a potential candidate for the final model.

5.3.2 Income

Income is a continuous variable and is considered a significant factor in modeling default by field experts. The default graph show that income is risk ranking i.e. default increases as income reduces depicting negative correlation. This theory is supported by analyzing

income as continuous variable and indicates that it is very significant with a p-value of 0.0211.

To compare various groups of income, income was categorized into bands considered similar by field experts.

Figure 5.16 having the graphs of the survival curves indicate that the lower income band ≤ KES 100,000 appear to be separate and lower than other categories thus some survival difference in the lower segment however graphs in the income segment KES (100k-300k) and ≥ KES 300,000 appear to be just slightly close together throughout the observation period indicating no major difference between the groups .This preliminary finding is confirmed by the Wilcoxon test for comparison of the survival curves that shows very significant values with p-values of 0.0448, ***Table 5.13.***

These results are also in line with the findings of (Bellotti & Crook, 2007) who found out that increase in UK Earnings and individual incomes reduced the risk of default as increase in earnings was an indicator of improving economy providing conditions for reduced risk of default. It was thus meaningful to classify borrowers on the basis of income as this did affect credit risk. (M Stepanova & Thomas, 2001) in their work behavioural scoring on credit customers also found out that income was significant in measuring time default of borrowers.

Income as a categorical variable was then considered as a candidate in the final model.

5.3.3 Age

Age is a continuous variable and is considered a significant factor in modelling default by field experts. Analysis of age as a continuous variable indicates that age is not significant in predicting default with a p-value of 0.1204.

Field experts however believe that default in the lower age group <30 is expected to be higher due their spend thrift nature and thus the ages were grouped. The graph of % default by age-band show a slight increase in default rate in the mid age groups of between 31 and 50. However the groups in the age bands less than 30 and above 50 have lower default probabilities. The default curves show that survival functions for borrowers in the age group greater than 50 is higher compared to the rest. Comparison of survival curves (***Table 5.15)*** show non-significance in the default curves with a Wilcoxon test giving a p-value of 0.1684.
Age or age categories will thus not be considered as a potential candidate for the final model.

5.3.4 Term

Default seems to be increasing with increase in term of loan. The graphs of the survival curve are closer together and do not depict any difference. ***Table 5.16*** confirms the Wilcoxon test that shows there is no difference in the survival curves with p-values of 0.9221.
These findings are similar to that of(Maria Stepanova & Thomas, 2002) in which their study concludes that segmentation by term of the loan has less effect in predicting default than early repayment because default is independent of term of the loan and early repayment is not.

However (M Stepanova & Thomas, 2001) found out that that profit increases as risk decreases suggesting that the longer the term of the loan the lower the profit obtained

from customers as risk of default increases suggesting that underwriters must look at both term and behaviour score when ranking loans of similar amount.

Term was therefore not considered a risk factor in the final model.

5.3.5 Commitment (%)

Default seems to be high in the groups with income commitment greater than 40%.

The graphs of the survival curves however seem to be closer to each other indicating no difference. This is confirmed by (***Table 5.14)*** Wilcoxon test that is non-significant with p-value of 0.1702.

Commitment was thus not considered as a potential candidate for the final model.

5.3.6 Banking History

Figure 5.13 and ***Table 5.10*** indicate that most of the customers have banking history less than 12 months with the majority in this group having banked for just 1 month. Default increases as banking history declines. The survival curves in the four groups show some significance with a p-value of 0.0095. This is shown in (***Table 5.17)***
Banking history was therefore considered a potential candidate in the final model.

5.3.7 Final model

Since only income and banking history were found to be significant under the univariate analysis, they were considered in the final model.
The same variables were found to be significant using the multivariate Cox PH technique with a p-value of 0.0807 and 0.0053 for income and banking history respectively ***(Figure 5.21).***

Interaction of the two variables was non-significant (***Figure 5.22)***. ***Figure 5.23*** testing for proportionality assumption with the two variables, that is, income and banking history was non-significant meaning that both income and Banking history are not time dependent covariates hence Cox PH method was used to model time to default.
The final model in ***Figure 5.21*** indicate that as customers move from low income ≤ KES100,000 to high income >KES 300,000, rate of default decreases by 51%. Customers with banking history less than 6 months experienced default that is 2.3 times higher than those who have banked more than 24 months and customers with banking history of between 6-12 months have a default rate that is 96% higher than those who have banked for more than 24 months. These results on income are consistent with those of Stepanova M and Thomas LC (2001) who used Cox's PH Regression model to build behavioural scoring models then compared performance with LR and (Bellotti & Crook, 2007) who used Cox PH model with Time Varying Covariates and contrasted with LR and Cox PH Model with & without macroeconomic variable. They both found income to be highly significant with default with the results demonstrating that survival analysis is competitive in comparison with logistic regression as a credit scoring method for prediction.

CHAPTER 6: CONCLUSION AND RECOMMENDATION

6.1 SUMMARY

Given that the banking sector has recorded double digit growth in profits for most of the past decade while economy has been growing by average of 5%, there have been growing concerns that the banking sector growing faster than the rest of the economy might result in institutions and households that are not able to repay their debts leading to the increase of non-performing loans. Banks would be required to hold higher capital buffers to absorb possible shocks. The buffers impacts on a bank's profits as it is a deductable expense. Effective management of credit portfolio to have buffers at manageable levels is therefore important to banks.

Traditionally, banks have used credit scoring to differentiate good customers from bad customers. Credit scoring looks at the borrowers' status after a fixed period of time. The idea of markov chain where borrowers move from one state to another brings to light that borrower's status is dynamic and not static. Credit scoring puts a static element to this dynamism.

Dynamicity has become a key research question with most studies focusing on not if but when will the borrowers default. Lending institutions would want customers whose profile would maximize their profits and this therefore means that if time to default is long, interest income will compensate or even exceed losses resulting from default. Time to default is modeled by using survival analysis techniques already described above and results of the analysis are also discussed above.

From the data used, gender, term, commitment and age are not associated with the risk of default but higher income brackets and longer banking history are found to be associated with lower default rates levels and therefore policies applying to customers in these segments should be more relaxed so that they are encouraged to borrow from the bank to increase banks profitability.

Since the estimated survival function does not go below 0.8462 due to censoring, we are unable to estimate the correct mean and median survival times hence unable to estimate how long a given loan ought to be granted.

6.2 RECOMMENDATION FOR POLICY

Lenders should not classify borrowers with respect to gender, term, commitment and age as these variables do not affect credit risk individually. However they can be considered in developing credit scoring models which are static models if ranking is found to be significant.

Higher income brackets (>KES 300,000) and longer banking history (>24 months) associated with lower default rates levels and therefore policies applying to customers in these segments should be more relaxed so that they are encouraged to borrow from the bank to increase banks profitability.

If institutions increase their appetite for the risky segments, the interest rates charged should be commensurate to the loss expected since if time to default is short, then interest income should be realized faster so as to compensate or even exceed losses resulting from default.

6.3 RECOMMENDATION FOR FURTHER RESEARCH

Since no assumption is made about hazard rate, no extrapolation of study results beyond the study period is possible. Parametric methods can be used to overcome this disadvantage. Other models like the Accelerated Failure Time models can also be considered in future.

From the data, income is not homogeneous across the population. We have so many people earning low income while just a few in the higher income band. COX PH assumes

homogeneity of data but the reality is that individuals across populations are not homogenous. Similar work using the same data would therefore also be explored in future with models that take care of heterogeneity like the frailty models.

REFERENCES

Andreeva, G. (2005). European generic scoring models using survival analysis. *Journal of the Operational Research Society*, *57*(10), 1180–1187. doi:10.1057/palgrave.jors.2602091

Argan, O., Corresponding, W., Samuel, M., & Peter, M. (2012). Modelling Credit Risk for Personal Loans Using Product-Limit Estimator, *3*(1), 22–32. doi:10.5430/ijfr.v3n1p22

Baesens, B., Van Gestel, T., Stepanova, M., Van den Poel, D., & Vanthienen, J. (2005). Neural network survival analysis for personal loan data. *Journal of the Operational Research Society*, *56*(9), 1089–1098. doi:10.1057/palgrave.jors.2601990

Banasik, J., Crook, J. N., & Thomas, L. C. (1999). Not if but when will borrowers default, 1185–1190.

Bellotti, T., & Crook, J. (2007). Credit Scoring With Macroeconomic Variables Using Survival Analysis, (i), 1–19.

Cao, R., & Vilar, J. M. (2009). Modelling consumer credit risk via survival analysis, *33*(June), 3–30.

Gandy, A. (2012). Performance monitoring of credit portfolios using survival analysis. *International Journal of Forecasting*, *28*(1), 139–144. doi:10.1016/j.ijforecast.2010.08.006

Giambona, F., & Iacono, V. Lo. (n.d.). Survival models and credit scoring : some evidence from Italian Banking System .

Im, J., Ã, D. W. A., Qi, C., & Shan, X. (2012). A time-dependent proportional hazards survival model for credit risk analysis, 306–321. doi:10.1057/jors.2011.34

Kleinbaum, D. G. (n.d.). *Statistics for Biology and Health*.

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

Prinja, S., Gupta, N., & Verma, R. (2010, April). Censoring in clinical trials: review of survival analysis techniques. *Indian journal of community medicine : official publication of Indian Association of Preventive & Social Medicine*. doi:10.4103/0970-0218.66859

Stepanova, M, & Thomas, L. C. (2001). PHAB scores: proportional hazards analysis behavioural scores. *Journal of the Operational Research Society*, *52*(9), 1007–1016. doi:10.1057/palgrave.jors.2601189

Stepanova, Maria, & Thomas, L. (2002). Survival Analysis Methods for Personal Loan Data. *Operations Research*, *50*(2), 277–289. doi:10.1287/opre.50.2.277.426

Zhang, J., & Thomas, L. C. (2012). Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD. International Journal of Forecasting, 28(1), 204–215. doi:10.1016/j.ijforecast.2010.06.002