University of Nairobi

School of Mathematics

ISSN: 2410-1397

Masters Project in Actuarial Science

Modelling Time to Default on Kenyan Bank Loans using Non-Parametric Models

Research Report in Mathematics, Number 54, 2020

Jonah Mudogo Masai

November 2020

Submitted to the School of Mathematics in partial fulfilment for a degree in Master of Science in Actuarial Science

Masters Project in Actuarial Science **University of Nairobi**

November 2020

Modelling Time to Default on Kenyan Bank Loans using Non-Parametric Models

Research Report in Mathematics, Number 54, 2020

Jonah Mudogo Masai

School of Mathematics
College of Biological and Physical sciences
Chiromo, off Riverside Drive
30197-00100 Nairobi, Kenya

Masters Thesis

Submitted to the School of Mathematics in partial fulfilment for a degree in Master of Science in Actuarial Science

Submitted to: The Graduate School, University of Nairobi, Kenya

Abstract

Financial institutions in past decades have been facing many risks that must be dealt with sensitively and in accordance with the instructions of the Central Bank of Kenya (CBK). In the forefront of these risks is credit risk which in case is ignored would likely plunge the banks into myriads of problems or even to bankruptcy. Papers on statistical models detailing on how to model credit risks have been published and have enabled banks to differentiate ’good’ and ’bad’ clients contingent on repayment performance during loan term. Credit granting is one of the main ingredients required for an economic spur in any given country. However, the technicalities attached to it poses a dilemma to the lending institutions on the appropriate approach to adopt when lending to minimize losses resulting from default.

The objective of this research is to identify credit scoring factors and to select non-parametric models of survival analysis which is most effective to model time to default. Variables considered based on FICO include income of the company, age of the company and account. It was evident that oldest companies whose accounts were opened more than 8 years before loan application have lower tendency of default. Also study show that Nelson Aalen is a better estimator of time to default to Kaplan-Meier. The study recommends more studies to incorporate macroeconomic variables to establish their impacts on client’s loan repayment performance and further estimate time to second default. It will also be interesting to extend this studies to the mixture curse model and study the performance of the resulting model in comparison with Cox proportional hazard model with penalized splines as our study involved univariate method.

Keywords Time to Default, Survival analysis, Censoring, Credit Scoring, Non Parametric techniques.

Master Thesis in Mathematics at the University of Nairobi, Kenya.
ISSN 2410-1397: Research Report in Mathematics
©Jonah Mudogo Masai, 2020
DISTRIBUTOR: School of Mathematics, University of Nairobi, Kenya

Declaration and Approval

I the undersigned declare that this dissertation is my original work and to the best of my knowledge, it has not been submitted in support of an award of a degree in any other university or institution of learning.

Signature   Date

JONAH MUDOGO MASAI
Reg No. I56/11036/2018

In my capacity as a supervisor of the candidate's dissertation, I certify that this dissertation has my approval for submission.

Signature   Date

Prof. Ivivi Mwaniki
School of Mathematics,
University of Nairobi,
Box 30197, 00100 Nairobi, Kenya.
E-mail: ivivi.mwaniki@gmail.com

Dedication

This project is dedicated to my parents,Mwanga's family,wife Emmah and our son Edens for their incredible support during research period.

Contents

List of Tables

Acknowledgments

First, I thank our Heavenly Father for His grace that enabled me come this far. Secondly,I thank my Supervisor Pof. Ivivi Mwaniki for his patience, inspiration and constant encouragement during the research. His invaluable knowledge, guidance and critique sailed me through valleys of giving up. Finally, my classmates with whom we constantly encouraged one another. Not forgetting the school of mathematics, University of Nairobi, for offering the opportunity to advance in Actuarial Science.
I thank you all.

Jonah Mudogo Masai

Nairobi, 2020.

List of abbreviations

CBK - Central Bank of Kenya

SACCOS - Savings and Credit Co-operatives

SASRA - Saccos Societies Regulatory Authority

DFI - Development Financial Report

FSSR - Financial Sector Stability Report

FICO - Fair Isaac Corporation

CRB - Credit Reference Bureau

KBA - Kenya Bankers Association

NPL - Non Performing Loans

CSMI - Credit Score Model for Individuals

DA - Discriminant Analysis

PD - Probability of Default

MAD - Mean Absolute Deviation

SE - Standard Error

CI - Confidence Interval

MLE - Maximum Likelihood Estimators

LLR - Log-likelihood Ratio

WLL - Weighted Local Linear

WNW - Weighted Nadaraya-Watson

VKA - Van Keilogom Akritas

INTRODUCTION

1.1 Background

Financial institutions such as banks and insurance companies in the past decades have been facing many risks that must be dealt with sensitively and in accordance with the instructions of the decisions maker in the banking sector.The institution regulating these financial institutions in Kenya is Central Bank of Kenya (CBK) which acts as the governor of the machinery of credit.In the forefront of these risks is credit risk that is considered as one of the banks main activities,which in case it is ignored would likely plunge the banks into myriads of problems or rather to bankruptcy.

The assessment of credit risk is crucial for the financial institutions. Basel committee supervision in June 2006 published the Basel II capital structure which requires that the financial institutions hold a minimum capital to cover the exposures of market,credit and operational risks. Thus,all banks are required to assess their portfolio risks,including credit risk.

As a result,research papers have been published on the use of statistical methods to model consumer credit risk and banks have developed credit risk scoring model and other sophisticated systems in an attempt to model the credit risk arising from important aspects of their business lines and these models are intended to aid banks in differentiating 'good' clients from 'bad' clients depending on their repayment performance (probability of default) over a given period of time,quantifying, aggregating and managing risk across geographical and product lines. The outputs of their models,also play increasing important roles in bank's risk management and performance measurement process.

Particularly,logistic regression has become a standard method for this undertaking(Thomas et al 2002) since the introduction of the classical Z-score model by (Altman 2002)which is applied to verify the grant of credit of an applicant. As an alternative to logistic regression, Narain (1992) first introduced the idea of using survival analysis in the credit risk context. Survival analysis has been employed in examining the possibility of the customer defaulting as well as early early repayment since it allows incorporation of censored data into the model.

Survival analysis approach has been employed also to model credit risk in the pricing of bonds and other financial investment. Lando(1994) in his PhD used Survival analysis to estimate time to default by introducing a proportional hazard model for bonds and he anticipated to used various economic variables as covariates.

Being a mathematical model used in estimating time to default,credit scoring model was first built by Narain(1992) and later was refined by Thomas et al(1999)using Survival analysis technique. Narain (1992) applied accelerated life Exponential model to 24 month loan data and authenticated that the proposed model estimated the number of failures at each failure time. After building scorecard using multiple regression,it was established that a better credit-granting decision could be made had the score supported by the estimated survival times and the author showed that this method can be applied to other areas of credit operations in which there are predictor variables and the time to some event of interest.

Thomas et al (1999) did performance comparison of Weibull,Exponential and Cox proportional hazards semi-parametric model with Logistic regression and was found Survival Analysis methods more competitive to the traditional logistic regression method. He also noted that several ways of improving the performance of the simplest Survival analysis models such as Weibull,Exponential and Cox proportional models exist.

The advantage of using survival analysis in this context is that the time to default can be modeled, and not just whether an applicant will default or not (Thomas et al., 2002).
(Thomas et al 1999)highlighted the merits of studying time to default and these are;

(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

1.2 Statement of the Problem

Kenya's financial sector comprises of deposits taking institutions (commercial banks and mortgage finance companies, micro finance banks and deposit taking Savings and Credit Co-operatives (Saccos)), non deposit taking institutions (insurance industry, pensions industry, capital markets industry, and Development Finance Institutions) and financial markets infrastructure providers. The sector is regulated and supervised by;Capital Markets Authority (CMA); Central Bank of Kenya(CBK); Insurance Regulatory Authority (IRA); Retirement Benefits Authority (RBA); and the Sacco Societies Regulatory Authority (SASRA) and Government Ministries for DFIs. The banking sector most specifically is a very important sector to the Kenyan economy as far as the Big four agendas advocated by the president are concern.

Based on the Financial sector stability report(FSSR,2017),the banking sector resilience saw its assets grew by 8.1% despite credit risk extended to private sector decelerating from 5.5% in the year 2016 to 2.2% in the year 2017. Assets growth were therefore driven by increased lending to the government, considered to be less risky. The banks also shortened loan maturities to less than five (5) years to reflect short-term funding dominated by demand deposits, and increased loan sizes amid reduced number of loan approvals.This was informed by increased in credit risk as reflected in growth of Non–Performing Loans (NPLs).

The gross NPLs as a ratio of gross loans increased from 9.3 percent in December 2016 to 11.0 percent in December 2017. In terms of growth rate however, NPLs decelerated from 43 percent in the year to 2016 to 25 percent in year December 2017,an indication of easing pressure.As a result,there has been 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.

Therefore having appreciated financial economic contribution,capability to foresee the bank failures as a result of default by having the appropriate statistical models to predict time to default is very essential.Due to the sensitivity of the data used,most suitable data is not readily available leading to the use of historical data in predicting failure times.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.Carrying due diligence when awarding loans to customers and having a robust would help in effective risk management.

As a result,estimating time to default is crucial since budget estimates have not been sufficient as it is not known exactly how long will the client issued with a loan will take before starting to miss repaying the loan. It is not clear which kind of distribution do their defaulting pattern follow thus prompting us to use semi-parametric distribution which is distribution-free and this research aims at identifying the better semi-parametric distribution that can best estimate probability of default since it incorporates the stochastic nature of default as it evolves over time in a random manner and it's based on incomplete information.

1.3 Objectives

The broad objective of this research is to use non-parametric models of survival analysis and selecting the most effective model to estimate probability of default which can be used for evaluating the performance of a sample of credit risk portfolio.
The specific objectives of this research are;
1. identifying credit factors that affects time to default
2. estimating probability of default using non parametric models and conducting estimates comparison
3. To test the statistical significance of the differences in the survival curves for distribution based on log-rank tests.

1.4 Justification of the Study

Credit granting is one of the main ingredients required for an economic spur in given country. However,the technicalities attached to it poses a dilemma to the lending institutions on the appropriate approach to adopt when lending to minimize losses resulting from default.

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

More specifically is the application of non-parametric methods which allow statistical inference without making the assumption that the sample has been taken from a particular distribution. The results of the research is anticipated to shed more light on the reliability and consistency of the Survival analysis methods on data analysis.Also,the credit risk analysts to be able to select the best method to adopt when estimating time to default.

Why survival analysis approach?

1. Survival analysis is able to account for censoring, unlike the other techniques.
2. Unlike linear regression, survival analysis has a binary outcome, which more realistic.
3. It analyses time to default rather than mere probability of defaulting.
4. Survival models naturally match the loan default process.
5. It gives a clearer approach to assessing the likely profitability of an applicant.
6. Survival estimates will provide a forecast as a function of time. Banasik et al (1999)

LITERATURE REVIEW

2.1 Introduction

This chapter provides previous researches that have been done and our study will explore those that are related to the area of the study. Author's name,topic of the study,year of publication and the journal used have been provided.

1. **Mingxin(2014): Residential mortgage probability of default models and methods. Financial Institution Commission**.

This paper provides an insight on alternative methods that can be used by lending institutions to assess the to to default on a pool of mortgage loans.

Methodology and results
Six models are studied beginning with the model that was earlier applied in the studies of residential mortgage default
While two models are for corporate loan portfolio,four model are for personal client's loan portfolio

Model 1 utilizes linear probability function to model probability of default. Here,a relationship between dependent variable and a number of variables that may have impact on the default behavior were considered
Upon fitting this model,its established that the estimate coefficients have an impact of the each default risk variable. Despite loan status measurement, model 1 assumes either zero or one only which is not always the case.
Model 2 is logistic model which is appropriate for empirical studies with qualitative data and formulates the chances of loan being non-performing as a logistic function of some combination of explanatory variables. Estimating this model using maximum likelihood techniques and goodness-of-fit tests conducted,it is found that its coefficients approximate the effects of the unit change in variables on the natural logarithm of the odds.However,it was found that its function may not be it a particular dataset.
Model 3 is a modeling technique for time-to event data or duration data.This is a model considers the duration the loan takes before it's defaulted or prepayed and thus confirms that default and prepayment are 'competing' risks as erlier stated by Deng,Quigley and Van Order ([12]). Likelihood methods of model estimation was employed and when estimated coefficients and empirical baseline hazard are utilized, conditional probabilities for some specific mortgages with given values of explanatory variables can be calculated.

Model 4 is an optimization model that tries to explore the key structure of economy around the default process and believes a borrower's decisions on prepayment are geared either towards maximizing wealth or minimizing house-related costs.

The model was based on decisions that borrowers take at each time interval of choosing less costly options which are either defaulting,refinancing or to continue with the current mortgage. At every interval, prices of houses as well as rates of interest are provided at that time and was found that from this distribution,computing probability of default for that time interval is possible results that are consistent with those of Capozza,Kazarian and Thomson ([7]).

Model 5 and 6 are based on the loan portfolio which is viewed as one subject.

Model 5 is a linear regression of default and establishes a relationship between default risk and an a number of factors. Unlike individual loan, default rate is computed as the quotient of loan numbers and the total number of the loans in the portfolio which serves as a measure of default risk for the loan portfolio.

Two approaches adopted are loan-to-value(LTV) ratio which has one independent variable of default and every mortgage in the portfolio as another LTV and average or median measures for the explanatory variables used in the analysis. periodically,one observes the default rate and explanatory variables for the entire portfolio over time.

Another way is to group the entire mortgage portfolio into sub-portfolios and view each part under investigation. This grouping method converts loan-by-loan data into a cohort-by-cohort sample which is then handled as a subject and observed in each period.

Model 6 considers linear regression analysis of log odds whereby a linear correlation between the dependent and predictor factor is deemed inappropriate when the first factor is a probability. In this case,natural logarithm odds are specified as ratio of linear function of the predictor variables.

Conclusions

As a result of applying 4 models for personal loan portfolio and 2 for corporate data,it's crucial to comprehend industry's features for the ideal model to be employed while recognizing every model shortcomings.

2. **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 objective of this research was to approximate probability of defaults at different time points by applying product limit estimator and conducting the statistical significance of the variation in the in the survival curves for both gender using log-rank tests.

Methodology

The total number of applicants which were randomly selected from one of the bank's loan portfolio was 500 each gender constituting half of the loan applicants and whose maturity was two and half years. The applicants loans were collected from the month of January of the year 2007 and monitored for a period of 30 months thus observation ending on June 2010

Upon missing loan instalment repayment for a period of three consecutive months,the loanee is considered to have defaulted. On the other hand,those accounts that were either closed by the individual clients or survived beyond the duration at which observation was being conducted were categorized as having censored. The incidences where clients offset their loan before the agreed duration were also considered to having censored. The lifespan of the account was counted as from the date the account was opened until the time that the account either slide into bad status or got censored.
The creditworthiness of the two groups were then monitored and the time which both genders either offset (censored) or failed to repay the loan for the three consecutive months was ordered in an ascending order. Then,using the product limit estimator as non parametric estimator, the survival probabilities for both genders were estimated. Parameters computed for the two groups also was mean and median survival times after which a comparison the the curves for both groups was conducted using the log-rank test.

Results There were defaults from both groups with male individual applicants leading with 11 number of defaults and 4 comprising those who opted to offset their loans.on the other hand,only 7 number of females defaulted with 6 applicants settling for early repayment. When their mean survival times were computed,there was a slight difference in their values with males having 15 and females 16 which stipulated their average number of months each group took before defaulting.A semblance in their survival curves substantiated by the test statistic log rank of 0.17 with a significance value of 0.678 and therefore it's clear that the at 95 percent confidence interval,the 2 survival curves were not statistically different.

Conclusions

It can therefore be concluded that when both genders are closely monitored and their mean survival estimate plotted against time,it's established that there is no tangible difference between the two groups thus rendering meaningless to classify applicants according to gender. This information will be of great value to underwriters in determining the average time taken by the loan applicants before defaulting in order to realize higher returns emanating from high profit returns.

3.**Jamil, J Jaber;Noriszura,Ismail(2017).Credit risk assessment for progressive right-censored data using survival analysis.Journal of internet banking and commerce,vol.22 no.1**

Purpose

The broad objective of this paper is to employ survival models in order to assess the the impact of the variables on the survival curves for a duration of 30 months on insurance retention and attrition.

Methodology

Insurance policies which were 158 in number were randomly selected from a pool of client policies from one of top performing insurance companies for a period of 18 months. The behavior of policies collected were then closely observed for a period of 30 months in order to obtain information therein that will be vital to the underwriters. Categorizations were conducted arranging the policies based on the time of their admission into study. There were 4 policies groups of policies with A and C admitted into the study on March 5,2013 and C on June 10th same year respectively.Similarly,there were group policies C and D which were admitted into the observation on February 10,2014 and 1st of January same year. Difference in the maximum follow up period was occasioned by variation in the times the policies were studies. While policies A failed to renew the term of the contract when it collapsed,policies C canceled their contract term with with insurance company. They were considered censored. However,both policies B and D were in force throughout all the period of three and half years.

Eventually a graph which estimates policies survivorship function was plotted for the purpose of stating the actual values representation. The commonly used Kaplan Meier estimator was then used with the intention of finding out the retention and attrition patterns for every subject that was being studied. In addition to that,it's by making assumptions that the other covariates were at their mean values that derivation of the cumulative survival probabilities was realized. In order to conduct comparison on survival probabilities among the groups, the Wilcoxon test statistic making use of variation in the group mean was applied. both gender survival curves were then plotted.

Results and conclusion

It was evident that between the two genders, male policy holders have a higher tendency of renewing their policies contrary to their female counterparts which have a lower tendency of renewing their insurance policies. On conducting test statistic,results exhibited p-value that is greater than 0.05 implying both males and females survival curves are similar. When standard error which is a goodness-of-fit was applied on means and medians for survival time at 95 percent confidence interval,there was more overlapping in the confidence intervals clearly showing that variation on their 'average' survival times were not statistically significant

When 4 covariates which include age,gender, mode of payment and policy type of Cox proportional hazard model are applied,it was established that the coefficients of age and gender were consistent with the insurance attrition

4.**Asia,Samreen;Farheen,Batul Zaidi (2012).Design and Development of Credit Scoring Model for the Commercial banks of Pakistan. Forecasting Creditworthiness of Individual Borrowers. International Journal of Business and Social Science**.

Purpose

The aim of the research was to design a credit scoring model for clients in order to determine their credit behavior in terms of repayment performance and to compare the authenticity of the proposed credit scoring model for individual with the already existing statistical credit scoring model.

Methodology

Data collection was done through interviewing the credit managers of the specific institutions as well as preparing the questionnaires. The study managed to use a total of 250 applicants collected from one of the top performing bank in Pakistan. The composition of the data showed that male applicants were more constituting 158 of the applicants and representing 63.2 percent of the entire set of the applicants. Females on other hand constitute 36.8 percent which represent 92 applicants. The variables of the applicant's considered were level of education,location of the clients,gender,their proximity to the bank,marital status,age,number of dependents,occupation,loan period,net monthly income,credit history working period with the last and current employer.

To compute the prospective client credit worthiness,different financial techniques were applied and among them are Descriptive statistics(Frequency Distribution and Cross tabulation),the Discriminant Analysis(DA) and Logistic Regression analysis on SPSS 17.0.

Discussion and Results

Results showed that 17.2 percent females borrowers and 21.2 percent of males borrowers defaulted affirming that male borrowers have the higher tendency of breaching earlier agreed terms of the loans in comparison to the female borrowers.

Furthermore,it was established that individuals who are home owners have higher credit score in terms of default chances since their default percentage is significantly less than those who do not have homes. Also,it was found that Type I error is more costly compared to type II error as it considers bad loanees as good which is highly risky. Banks loose the potential applicants in type II error and hence reduce their revenues. Married loan applicants are taken by banks to be less risky and more creditworthy because they have responsibility of their spouses and families as compared to single applicants.Apart from young,salaried employees,those who also haven't defaulted before have less probability of default. Credit Scoring Model for Individual accuracy was found to be 100 percent more than the other models used.

Conclusions and Recommendations

Out of 250 applicants,there are 96 applicants who are predicted to be bad or defaulters when credit scoring for individual while 154 applicants are predicted to be good customers. The accuracy rate of Discriminant Analysis is 95.2 percent,Logistic Regression 98.8 percent and credit scoring model for Individual(CSMI) is 100 percent and eventually concluded that CSMI have the highest accuracy rate and also the most effective model compared to the two models.

For banks to reduce than performing loans,it is recommended that SCMI is adopted for evaluation process. The current research used the accepted applicants in the sample and it is highly advisable to collect the data of rejected applicants by banks so that more versatile results could be obtained.

5. **Bellotti, T. and Crook JN(2009):Credit Scoring With Macroeconomic Variables Using Survival Analysis.J Opl Res Soc 60:1699-1707**

Purpose

The objective of this research is to use survival analysis techniques to model probability of default using a large data set of accounts pooled from the individual cred card accounts and to confirm that it's competitive for forecasting of default in compared to conventional approach which is Logistic Regression(LR) as well as testing the hypothesis that probability of default is affected by general conditions in the economy over time and how time-varying variables can be incorporated into survival analysis.

Data and Methodology

Data

A large sample of accounts totaling to 100,000 were collected from one of the best UK bank and it constituted data for monthly performance. Four variables were considered and they were income,unemployment,age and housing. The data were for a period of 8.5 years as from 1997 to mid 2005. there were training data set which included those accounts that were opened between 1997 1997 and 2001 as well as testing data set which included those accounts that were opened between 2002 and 2005.

An account was categorized said to have breached the loan terms (default) if t fails to submit monthly instalments for 3 consecutive months or exhibit more default signs within 12 months and was referred to as bad account while non-defaulting accounts were referred as good accounts.

Time-varying covariates

Macro economic variables that were considered mostly to affect default were selected and they are interest(IR),earnings,unemployment index,production,housing and consumer confidence index.Arise in the macroeconomic variables which correlates with an increase in default is reflected in the positive value. Conversely,an increase in value of those covariates with negative value implores a decrease in default risk.

Methods

Here there's description of model training, selection and model assessment. Under the first model which is training,data was modeled using Cox Survival model and contrasted with Logistic Regression.As a result of data skewness occasioned by number of bad cases,more percentage was awarded to bad case for training in comparison to numbers of bad to good cases in training data and it was possible for both Cox PH and LR as both use Maximum Likelihood Estimation(MLE) for which bad cases could be included in the likelihood function multiple cases.

Under second model which is selection, embracing interactions between application and macroeconomic variables lead to more expectation of better models. When each macroeconomic covariate was interacted with an application variable and included to the basic model,there was an inclusion of interaction giving the lowest p-value for its LLR in the optimal macroeconomic Cox PH model.

The final model was assessment which was considered optimal model was assessed due to its explanatory capability on the training data and its approximate power on the predictor test set. The Cox model confirmed explanatory model by reporting its viability to the training data with and without macroeconomic variables using LLR. Each model significance was conducted by Wald statistic obtained from the MLE. This test statistics uses chi-square statistic ,so a p-value can be computed for the null hypothesis whose value is zero. Also,upon testing Cox PH model as a predictor of default in order to determine its viability as a Credit Scoring

Discussion,Results and Conclusion

The three models were found to fit the training data well and significantly. However,inclusion of macroeconomic covariates into the Cox PH model is highly significant in model fit. There was positivity in the coefficients of interest rate and unemployment showing an increase in hazard with increase in bank interest rates and the level of unemployment which is contrary to hazard which decreases with increase in the FTSE index and the levels of real earnings which is what was expected since these are indicators of ability to repay.Interest rate was found to be a leading variable in influencing default risk as expected. Survival analysis was established to significantly improve performance evident by reduction in mean cost.This can be largely attributed to inclusion of macroeconomic variables.

LR was outperformed by cox model in all periods except quarter 3 in 2002. The model was an ideal for stress testing by including macroeconomic conditions that simulate a depressed or booming economy.

METHODOLOGY

Non-parametric technique is one of the three broad categories of estimating survival functions in Survival analysis. The other two are parametric and semi-parametric techniques.

Survival analysis

Survival analysis is also known as Event history analysis(Sociology), Duration models(Political Science and Economics), Hazard models (Bio statistics) and failure-time models (Engineering and reliability analysis). It is an umbrella term for a collection of statistical methods that focus on questions related to timing and duration until an occurrence of an event of interest. The primary aim of survival analysis is to find and interpret the survival function of survival data and it is meant to circumvent the issues arising out of the incomplete information regarding the time until a desired event occurs.

Since most of the survival data tend to be positively skewed and unsymmetrical as well as containing censored observations. However, its interpretation using parametric techniques is more complicated. The models examine the hazard rate which is the conditional probability that an event at a particular time interval (t). It examines how long it takes until the event of interest occurs and it is useful to note that Survival models are actually just regression models with somewhat different likelihood estimators than Ordinary Least-Squares regression (OLS).

An event may take many forms such as an organ transplant, marriage, birth, death, political revolution, time to default or bank merger.The mathematical expressions and relations of statistical functions are then presented in a manner that requires basic background in mathematics and statistics. Survival analysis has wide application in Social Science and more generally useful for any issue in which:

i. The phenomenon of interest is a duration. And/or,

ii. The response is the occurrence of a discrete event in time.

The characteristics of Time-To-Event Data when modeling survival data are:

a. discrete events,

b. take place over time

c. may not (even never) experience the event(i.e possibility of censoring)

Modeling duration data however presents several tricky issues;

(i) Like count data,duration data are strictly non-negative.

(ii) The data are conditional, i.e to survive to some time t, one must necessarily have survived up to t-1 as well.

(iii) Additionally, we regularly encounter observations which have not failed yet(i.e censored data)

Censoring and Truncation

A distinguishing factor of Survival and event history models is that they take censoring into account. Censoring simply means, we have information about an individual's survival time but do not know the exact survival time (Kleinbaum and Klein, 2005). Various types of censoring can occur, with the most common type being right-censoring, which will also be the primary focus in this research.

Truncation refers to the complete lack of information about an occurrence of the event. A confusion has often been arising as to whether observations are censored or truncated. However, while truncation refers to the cases where subjects do not appear in the data because they are observed, censoring refer to the cases when subjects are known to fail within a particular episode but the exact failure time is unknown (Allison, 1984).

Types of censoring and truncation
Right-censoring
This occurs when the event under study is not experienced by the last observation and it commonly occurs in the Social Sciences when survey data is used. For instance, individuals are often questioned about their retrospective life histories, such as the birth dates of their children or start and end dates of jobs or education. If we were modeling the transition to second childbirth for example using this type of retrospective data, then all individuals who had a first child but no second child at the time of the observation would be right censored by the survey date.

Interval censoring
This refers to the case where we only have information that the event occurred between two known time points, but not the exact timing of the event.

For example, consider a case where employment status is being studied and only the employment status categories were asked every two years and not the timing of changes.

If someone was unemployed at the first data collection wave but employed at the second wave, we would know that there was a change in employment status, but not exactly when the event occurred during this period.

Left censoring

This is a situation where the event of interest has occurred but we do not know when it happened. For example, when a patient goes to the hospital and is found infected with Covid-19. We do not know when the virus was contracted but what we know is that the patient tested positive for the disease.

Random censoring

This refers to a situation where it is not known exactly when the event of interest will occur. The censoring time is $C_i \leq (T_i)$ where T_i is a time of the event and C_i is the censoring time.

Informative censoring

This occur when a phenomenon that would possibly trigger an event of interest is known in advance. For example, if a worker changes job from one institution to another because of ill health then we have prior information that he was more at risk in the event of death.

Non-informative censoring

This occurs when no phenomenon that would possibly trigger an event of interest is known in advance.

For example, a worker changes job from one company to another. In the event of death, there is no indication that changing job has more or less risk.

Type I censoring

This refers to censoring where the duration for the study is given in advance. e.g if the duration of study is from 10^{th} Feb, 2019 to 15^{th} April, 2020.

Type II censoring

This occurs where the stopping criterion is when a given number of events occur. e.g an estate is declared a Covid-19 hot spot when 15 cases have been confirmed for the disease.

Left truncation

This is the most common type of truncation and is when subjects enter the study at a random age. Here, we do not have information from the onset of the risk to some time after the onset of risk.

Interval truncation

This is also known as gap truncation. This occurs when subject under study drops and the rejoins again during the period of study. For example, in a clinical study, a patient is under observation for the first 3 months of the study, drops out for 2 months and then rejoins the study again for the last 7 months. Dropping out of the study for 2 months creates an interval or gap in the period of observation.

Right truncation

Though less frequent, can also occur, e.g during an examination of an episode from HIV infection until the development of AIDS. If the sample only includes those who have developed AIDS prior to the end of the study, those HIV-infected individuals who have not yet progressed to AIDS are excluded from the sample thus right truncation.

Mathematical expression of Survival analysis functions

Let T be a random variable of Survival time ($T \geq 0$) and t be specific value for T. The values of T have a particular probability distribution, denoted by a probability density function represented by $f(t)$ and a cumulative density function $F(t)$. The distribution function of random variable T is given by;

$$F(t) = \int_0^t f(u)du = Pr(T \leq t)$$

where $Pr(T \leq t)$ is the probability that a survival time T is less than or equal to some value t. For all points at $F(t)$, the probability density function $f(t)$ is given by;

$$f(t) = \frac{\partial F(t)}{d(t)} = \acute{F}(t)$$

This implies that:

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

The density function $f(t)$ expresses the unconditional instantaneous probability that an event occurs in the time interval $(t, \Delta t)$ and is specified as;

$$f(t) = \lim_{\Delta t \to 0} Pr \frac{(t \leq T \leq (t+\Delta t))}{\Delta t}$$

Therefore it is clear that the density function is an unconditional failure rate. It describes the unconditional (i.e not conditional on covariates) instantaneous(at any given instant t) probability of the event (i.e failure rate).

Another core concept that is very instrumental when ESTIMATING TIME TO DEFAULT in Survival and event history model is survival function given by;

$\hat{S}_t = 1 - F(t) = Pr(T \geq t)$

which expresses the probability that survival time T is equal to or greater that some value t. $\hat{S}_t$ denotes the proportion of subjects surviving beyond t.

Properties of Survival function

i) At origin time $t = 0$, $S(0) = 1$ which simply means that all the subjects in the study are surviving at $t = 0$.

ii) $\hat{S}(t)$ is strictly decreasing function.

iii) At time $t = \infty$, $S(\infty) = 0$

The occurrence of an event (e.g failure) and survival are related to each other and is encapsulated by the hazard rate/instantaneous transition/hazard as;

$$ht = \frac{f(t)}{\hat{S}(t)}$$

Relationship among probability density, survival and hazard functions

It is evident that;

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} Pr\frac{(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} Pr\frac{(T \geq t | t < T \leq t + \Delta t)\, Pr\,(t < T \leq t + \Delta t)}{Pr(T \geq t)\Delta t} \\ &= \lim_{\Delta t \to 0} Pr\frac{(T \geq t | t < T \leq t + \Delta t)\, f(t)}{s(t)\Delta t} \end{aligned}$$

but

$$Pr\frac{(T \geq t | t < T \leq t + \Delta t)\, f(t)}{\Delta t} = 1$$

Therefore $h(t) = \frac{f(t)}{1-F(t)}$, where $h(t)$ is the derivative of $-logs(t)$.
Thus

$$h(t) = \frac{f(t)}{s(t)} = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{1 - \int_0^t f(s)ds}$$

Integrate both sides,

$$\begin{aligned}\int_0^t f(s)ds &= \int_0^t \frac{f(s)}{1-\int_0^t f(s)ds}ds\\ &= -\ln[1-\int_0^t f(s)ds]_0^t + C\\ &= 1-\int_0^t f(s)ds = \exp[-\int_0^t h(s)ds]\end{aligned}$$

Differentiate both sides

$$\begin{aligned}-f(t) &= -h(t)\exp - \int_0^t h(s)ds\\ f(t) &= h(t)\exp[-\int_0^t h(s)ds]\end{aligned}$$

And since

$$\begin{aligned}h(t) &= \frac{f(t)}{S(\hat{t})}\\ S(t) &= \frac{f(t)}{h(t)}\end{aligned}$$

Replacing $f(t)$ by $h(t)\exp[-\int_0^t h(s)ds]$

$$\begin{aligned}S(t) &= \frac{h(t)\exp[-\int_0^t h(s)ds]}{h(t)}\\ \therefore S(t) &= \exp[-\int_0^t h(s)ds]\end{aligned}$$

Semi-parametric models

Semi parametric model is a statistical model that has both parametric and non parametric components. A parametric component is characterized by distribution $P_\theta : \theta\varepsilon\ominus$ indexed by a parameter θ_c while in non parametric component, the set of possible values of the parameter θ is a subset of some space V which is not necessarily finite-dimensional.

Why semi parametric?

1. It is easy to understand and to work with and

2. It allows you to have the best of both worlds.

A model that is understandable and can be manipulated while offering a fair representation of the messiness that is involved in real life. However, it also fail to give fair representation of what is happening in the real world.

Cox Proportional Hazard model

Survival analysis methods can also be extended to assess several risk factors simultaneously similar to multiple linear and multiple logistic regression analysis. One of the most popular regression technique for survival analysis is Cox proportional hazards regression, which is used to relate several risk factors or exposures considered simultaneously to survival time.

In Cox proportional hazard regression model, the measure of effect is the **hazard rate** which is the risk of failure (i.e the risk or probability of suffering the event of interest), given that the participant has survived upto a specific time. Though probability must lie in the range 0 and 1, hazard which represents the expected number of events per one unit of time in a group can exceed 1. For example, if the hazard is 0.2 at time t and the time units are months,then on average, 0.2 events are expected per person at risk per month. Another interpretation is based on the reciprocal of the hazard. For instance, $\frac{1}{0.2} = 5$, which is the expected event-free time(5 months) per person at risk.

Important assumptions for appropriate use of the Cox proportional hazards regression model include;

i. Independence of survival times between distinct individuals in the sample,

ii. A multiplicative relationship between the predictors and the hazard (as opposed to a linear one as was the case with multiple linear regression analysis),

iii. A constant hazard ratio over time.

Extension on Cox proportional hazard models.

There are 3 important extensions of the Cox proportional hazard model approach.

1.**Time dependent covariates**

In the previous Cox, we have considered the effect of risk factors measured at the beginning of the study period or at the baseline, but there are many applications where the risk factors or predictors change over time. Suppose we wish to assess the impact of exposure to marijuana and alcohol during pregnancy on time to preterm delivery. Smoking and alcohol consumption may change during the course of pregnancy. These predictors are called time-dependent covariates and they can be incorporated into survival analysis models. The Cox proportional hazards regression model with time dependent covariates taking the form;

$$\ln \frac{h(t)}{h_o(t)} = b_1X_1(t) + b_2X_2(t) + ... + b_pX_p(t)$$

Here, each of the predictors $X_1, X_2, ..., X_p$ now has a time component. Though survival analysis models can include both time independent and time independent predictors simultaneously, a difficult aspect of the analysis of time-dependent covariates is the appropriate measurement and management of these data for inclusion in the model.

2. **Proportionality Assumption**

This is a very important assumption for the appropriate use of the log rank test and the Cox proportional hazards regression model. Specifically, we assume that the hazards are proportional over time which implies that the effect of risk factor is constant over time. There are several approaches to assess the proportionality assumption, some are based on statistical tests and others involve graphical assessment.

In the statistical testing approach, predictor by time interaction effects are included in the model and tested for statistical significance. If one (or more) of the predictor by time interactions reaches statistical significance (e.g $p < 0.05$), then the assumption of proportionality is violated. In graphical analysis, there are several graphical displays that can be used to assess whether the proportional hazards assumption is reasonable. These are often based on residuals and examine trend (or lack thereof) over time (Hosmer and Lemeshow, 2000).

If either a statistical test or a graphical analysis suggest that the hazards are not proportional over time, then the Cox proportional hazards model is not appropriate and the adjustments must be made to account for non-proportionality. One approach is to stratify the data into groups such that within groups, the hazards are proportional and different baseline hazards are estimated in each stratum.

3. **Competing risks**

The competing risks issue is one in which there are several possible outcome events of interest. For example, a prospective study maybe conducted to assess risk factors for time to incident Cardiovascular disease. Cardiovascular disease includes myocardial infarction, Coronary insufficiently and many other conditions. The investigator measures whether each of the component outcomes occur during the study observation period as well as the time to each distinct event. The goal of the analysis is to determine the risk factors for each specific outcome when the outcomes are correlated.

Cox proportional hazard models is give by;

$$h(t) = h_o(t)e^{\beta' \underline{\mathrm{x}}}$$

where,
$h_o(t)$ is the baseline function,
h(t) is the hazard function at time t,

β' is the regression coefficient vector and
$\underset{\sim}{x}$ is the covariate vector.

The two assumptions on this model are;

1. All individuals have the same shape.
2. The hazard function is proportional to baseline function.

$h(t) \propto h_o(t)$

which implies it can be expressed as

$h(t) = e^{\beta'\underset{\sim}{x}h_o(t)}$

Where $e^{\beta'\underset{\sim}{x}}$ is the constant of proportionality.

Generally, the proportional hazard model can be given by;

$\psi\underset{\sim}{x}h_o(t)$ where $\psi\underset{\sim}{x} = \psi(x_1,x_2,x_3,...,x_p)$

which is the general proportional hazard model.

The contribution of Cox was to come up with an estimation method technique for estimating the regression coefficient β without considering the baseline function $h_o(t)$ and this parameter estimation technique is called partial likelihood function technique.

let,
1. $t_1,t_2,t_3,...,t_k$ where t_i is time of an event under investigation.

2. $R(t_i)$ be the risk set.

3. probability of individual "i" falling at time $t_i/R(t_i)$ be given by;

$\frac{\psi_i(\underset{\sim}{x})}{\sum_{j\varepsilon R(t_i)} \psi_j(\underset{\sim}{x})}$

Then the likelihood function

$L = \prod_{i=1}^{k}\{\frac{\psi_i(\underset{\sim}{x})}{\sum_{j\varepsilon R(t_i)} \psi_j(\underset{\sim}{x})}\}$

Now, to estimate β the regression coefficient we take $logL$ and solve

$\frac{\partial logL}{\partial \beta_k} = 0$, for $k = 1, 2, 3, ...$ (Number of parameters)

$$\frac{\partial^2 logL}{\partial \beta_k^2} < 0$$

For Cox proportional hazard model,

$$\begin{aligned} L &= \prod_{i=1}^{k} \{ \frac{h_i(t_i)}{\sum_{j\varepsilon R(t_i)} h_j(t_i)} \} \\ &= \prod_{i=1}^{k} \{ \frac{h_o(t_i) e^{\beta' \underline{x}_j}}{\sum_{j\varepsilon R(t_i)} h_o(t_i) e^{\beta' \underline{x}_j}} \} \end{aligned}$$

Where

$$\beta' \underline{x}_j = \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi}$$

Estimating β for one covariate

$$h(t) = e^{\beta x} L \quad = \prod_{i=1}^{k} \{ \frac{e^{\beta x_i}}{\sum_{j\varepsilon R(t_i)} e^{\beta x_j}} \}$$

Taking log

$$\begin{aligned} LogL &= log \prod_{i=1}^{k} \{ \frac{e^{\beta x_i}}{\sum_{j\varepsilon R(t_i)} e^{\beta x_j}} \} \\ &= \sum_{i=1}^{k} \{ log \frac{e^{\beta x_i}}{\sum_{j\varepsilon R(t_i)} e^{\beta x_j}} \} \\ &= \sum_{i=1}^{k} \{ log e^{\beta x_i} - log \sum_{j\varepsilon R(t_i)} e^{\beta x_j} \} \\ &= \sum_{i=1}^{k} \{ \beta x_i - log \sum_{j\varepsilon R(t_i)} e^{\beta x_j} \} \\ \frac{\partial logL}{\partial \beta} &= \sum_{i=1}^{k} \{ x_i - \frac{\frac{\partial}{\partial \beta} \sum_{j\varepsilon R(t_i)} e^{\beta x_j}}{\sum_{j\varepsilon R(t_i)} e^{\beta x_j}} \} \quad (Qoutient \;\; rule) \\ &= \sum_{i=1}^{k} x_i - \sum_{i=1}^{k} \{ \frac{\sum_{j\varepsilon R(t_i)} x_j e^{\beta x_j}}{\sum_{j\varepsilon R(t_i)} e^{\beta x_j}} \} \\ \frac{\partial^2}{\partial \beta^2} &= -\frac{\partial}{\partial \beta} \sum_{i=1}^{k} \{ \frac{\sum_{j\varepsilon R(t_i)} x_j e^{\beta x_j}}{\sum_{j\varepsilon R(t_i)} e^{\beta x_j}} \} \\ &= -\sum_{i=1}^{k} \{ \frac{(\sum_{j\varepsilon R(t_i)} e^{\beta x_j}) \sum_{j\varepsilon R(t_i)} x_j^2 e^{\beta x_j} - (\sum_{j\varepsilon R(t_i)} x_j e^{\beta x_j}) \sum_{j\varepsilon R(t_i)} x_j e^{\beta x_j}}{[\sum_{j\varepsilon R(t_i)} e^{\beta x_j}]^2} \} < 0 \end{aligned}$$

Solving the equation $\frac{\partial logL}{\partial \beta} = 0$, we get $\hat{\beta}$ as the maximum likelihood estimator of β.

Parametric survival analysis models
Parametric models for survival data don't work with the normal distributions since normal distributions can have any value, even negative values. Since parametric survival model require non negative distribution, then the distributions that work well for survival data include Exponential, Weibull, Gamma and Log normal distributions.

Why parametric model

1. This model provides greater efficiency due to estimation of fewer parameters.
2. Offers room for extrapolation beyond the range of the data.
3. When this model matches some underlying mechanism associated with your data, you end up with more relevant interpretations of your model.

Non Parametric

The most common non-parametric technique for modeling the survival function is the Nelson Aalen and Kaplan Meier estimates. One way to imagine about survival analysis is non-negative regression and density estimation for a single random variable (first event time)in the presence of censoring. In line with this, the Nelson Aalen and Kaplan Meier are a non-parametric density estimates (empirical survival functions) in the presence of censoring

However, in these estimates, it's not easy to incorporate covariates meaning that it's difficult to describe how individuals differ in their survival functions.

1.**Kaplan-Meier Product-Limit**
Kaplan-Meier is one of the non-parametric technique for estimating survival function S(t). Note that it is not continuous, but this estimator is a step function with discontinuities at the failure times. i.e only piece-wise continuous (actually, piece-constant, or "step function").This non-parametric technique estimates the survival function from the incomplete or uncensored data especially right censoring. In the absence of censoring in Kaplan-Meier, the estimate of the survival function is the empirical survival function or proportion alive at time t.

Mathematically, this can be written as;
$\hat{S}(t) = \frac{1}{n}\sum_{i=1}^{n} l(t_i > t)$
If the largest observation time is censored then the curve will not drop to zero but rather

undefined after the last censoring. This method is ideal when no assumptions are made about the functional distribution of hazard rate with time.

Model derivation
Notation

- j^{th} time interval $[t_{j-1}, t_j]$
- t_j-the time an event occurs, $j = 1, 2, ..., k$
- N -the population or sample size under study
- c_j the number of censoring in the j^{th} interval.
- d_j the number of events at time t_j
- n_j number of loans at risk just before time t_j

Let $t_1, t_2, t_3, ...$ denote the actual times of default of the n individuals in the credit risk portfolio. Also let $d_1, d_2, d_3, ...$ denote the number of defaults that occur at each of these times, and let $n_1, n_2, n_3, ...$ be the corresponding number of borrowers remaining in the credit portfolio.
Note that $n_2 = n_1 - d_1, n_3 = n_2 - d_2$ etc.
Then,

$S(t_2) = P(T > t_2) =$ "Probability of surviving beyond time t_2" depends conditionally on $S(t_1) = P(T > t_1) =$"Probability of surviving beyond time t_1."

Likewise,$S(t_3) = P(T > t_3) =$ "Probability of surviving beyond time t_3" depends conditionally on $S(t_2) = P(T > t_2) =$ "Probability of surviving beyond time t_2" etc. By using this recursive idea,we can iteratively build a numerical estimate $\hat{S}(t)$ of the true survival function S(t). Specifically,

$*$ For any time $t \in [0, t_1]$, we have $S(t) = P(T > t) =$ Probability of surviving beyond time time $t = 1$,because no defaults have yet occurred. Therefore,for all t,in this interval,let $\hat{S}(t) = 1$.

Using Bayesian approach,for any two events A and B, P(A and B)=P(A)*P(B/A)

Let;
A = survive to time t_1 and B = survive from time t_1 to beyond some time t before t_2. Having both events occur is therefore equivalent to the event
$(A \;\; and \;\; B)$ = survive to beyond time t before t_2, i.e $T > t$ and hence the following holds,

$*$ For any time $t \in [t_1, t_2]$, we have

$S(t) = P(T > t) = P($survive in $[0, t_1) * P($survive in $[t_1, t] \mid$ survive in $[0, t_1)$

i.e $\hat{S}(t) = 1 * \frac{n_1 - d_1}{n_1}$

$= 1 - \frac{d_1}{n_1}$

Similarly,
$\therefore$ any time $t \in [t_2, t_3)$, we have

$S(t) = P(T > t) = P(\text{survive in } [t_1, t_2)) * P(\text{survive in } [t_2, t] | \text{survive in } [t_1, t_2))$
$\hat{S}(t) = \left((1 - \frac{d_1}{n_1}\right) * \left(\frac{n_2 - d_2}{n_2}\right)$

$= (1 - \frac{d_1}{n_1})(1 - \frac{d_2}{n_2})$

In general, for $t\varepsilon[t_j, t_{j+1})$, $j = 1, 2, 3, ...$ we have

$\hat{S}(t) = (1 - \frac{d_1}{n_1})(1 - \frac{d_2}{n_2})...(1 - \frac{d_j}{n_j})$

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

2. **Nelson-Aalen(Fleming-Harrington) estimator/Alt-Schuler's estimate**

This method which is based on individual event times provides a consistent estimate of the cause-specific hazards. It's closely related to the theory of counting processes representing the expected number of events in $(0, t)$ for a unit permanently at risk and this interpretation is ideal for recurrent events. Also, cumulative incidence functions can specify the joint distribution which represent the probability of failing from a given cause before a specific time. Here, all causes of failure are involved to estimate the cumulative incidence functions of a given cause, and thus other failures cannot be treated as censored observations and it depends on the individual event times. This estimation method estimates the cumulative hazard which in turn is used to estimate the survival function using the relationship,

$\hat{S}(t) = e^{-\hat{H}(t)}$

Nelson Aalen performs better than Kaplan-Meier when the sample size is small but almost similar to the latter when the sample size is very large.

To derive its survival and cumulative hazard function,
we have, Geometric series given by;

$\frac{1}{1-x} = 1 + x + x^2 + x^3 + ...$

Integrating both sides,we have

$$\int \frac{dx}{1-x} = \int [1+x+x^2+x^3+...]dx$$
$$-log(1-x) = x+\frac{x^2}{2}+\frac{x^3}{3}+...$$

and for small x, its powers are ignored.

$$[-log(1-x) \simeq x$$

when we put $x = \frac{d_j}{n_j}$

from K-M, we know that

$$\hat{S}(t) = \prod_{t_j \leq t} \left(1-\frac{d_j}{n_j}\right)$$
$$log\hat{S(t)} = log\prod_{t_j \leq t} \left(1-\frac{d_j}{n_j}\right)$$
$$= \sum_{t_j \leq t} log\left(1-\frac{d_j}{n_j}\right)$$
$$\simeq \sum_{t_j \leq t} \left(\frac{d_j}{n_j}\right)$$

but $S(t) = \exp\left[-\int_0^t h(u)du\right]$

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

H(t)= Cumulative Integrated hazard function

$\hat{H}(t) \simeq \sum_{t_j \leq t} \left(\frac{d_j}{n_j}\right)$ which is Nelson-Aalen estimator. The estimation of this estimator is done by making use of the survival function $\hat{S}(t)$ via Cox regression model without covariates

To estimate the survival function using this technique, we use

$S(t) = e^{-\hat{H}(t)}$ which implies it can be expressed as

Standard error of the estimated survival function

Precision of the estimate which is obtained from the interpretation of an estimate of any quantity is normally captured by the standard error of the estimate. It's the square root of the estimated variance of the estimate and its essential in finding interval estimate for a quantity of interest. Therefore, we estimate the variance of the Kaplan-Meier survival function estimate using Greenwood's formula for variance.

Note that the estimator for Kaplan-Meier is given by,

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

Taking log in both sides,we have

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

and thus let $\left(1 - \frac{d_j}{n_j}\right)$ be p_j

Therefore,

$$\begin{aligned}\log(S\hat{(t)}) &= \sum_{j=1}^{k}\log(p_j)\\ var\left(\log(S\hat{(t)})\right) &= var\left(\sum_{j=1}^{k}\log(p_j)\right)\\ &= \sum_{j=1}^{k} var(\log(p_j))\end{aligned}$$

Since $\frac{d_j}{n_j}$ are asymptotically independent and $S\hat{(t)}$ a function of the $\frac{d_j}{n_j}$, we can estimate its variance using the delta method.

This method states that if Y_n is (asymptotically) normal with mean μ and variance σ^2, g is differentiable and $g\prime(\mu) \neq 0$,then $g(Y_n)$ is approximately normally distributed with mean $g(\mu)$ and variance $[g\prime(\mu)]^2\sigma^2$

Now,applying delta estimation technique, we have

$$\begin{aligned}var\left(\log(p_j)\right) &\approx \left(\frac{1}{\Pi_j}\right)^2\frac{\Pi_j(1-\Pi_j)}{n_j}\\ &= \left(\frac{1}{\Pi_j}\right)\frac{(1-\Pi_j)}{n_j}\end{aligned}$$

Hence,

$$\begin{aligned} var\left(\log(p_j)\right) &\approx \sum_{j=1}^{k}\left(\frac{1}{\Pi_j}\right)\frac{(1-\Pi_j)}{n_j} \\ var\left(\log(S\hat{(t)})\right) &\approx \sum_{j=1}^{k}\left(\frac{1}{\Pi_j}\right)\frac{(1-\Pi_j)}{n_j} \\ var\left(\hat{S}(t))\right) &\approx [S\hat{(t)}]^2 var\left(\log(\hat{S}(t))\right) \\ &= [\hat{S}(t)]^2\sum_{j=1}^{k}\left(\frac{1}{\Pi_j}\right)\frac{(1-\Pi_j)}{n_j} \end{aligned}$$

Substituting $p_j = \frac{n_j - d_j}{n_j}$ for Π_j, we have variance equals to the

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

and the standard error of the estimate given by,

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

A more consistent measure of goodness-of-fit is obtained when the total difference of 99% confidence interval is applied as a criteria for selecting the best non-parametric model.

Now,if 99% confidence interval is to be obtained,the we use

$\hat{S}(t) \underline{+} z_{1-\frac{\alpha}{2}} se[\hat{S}(t)]$

Where $se[\hat{S}(t)]$ is calculated using the Greenwood's formulae. Here, t is fixed and is referred to as point-wise confidence interval.

TEST OF STATISTICS

Consider the grouping in some study.

Let us denote the distinct times observed failures as $t_1 < t_2 < ...t_k$ and define

n_{ij}-the population sizes of the i^{th} group at time $t_j(i = 0, 1; j = 0, 1, 2, 3, ..., k)$

d_{ij}-those in group i who fail (uncensored) at time t_j $(i = 0, 1; j = 1, 2, 3, ..., k)$

n_j- the total population size from the two groups at time t_j

d_j=$d_{oj} + d_{1j}$-the total number of failures at time t_j

Tabularly, this can be represented as

Group	No. of failure	No. of success	Total
1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
2	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
Total	d_j	$n_j - d_j$	n_j

Table 1. Table 1

DATA ANALYSIS AND FINDINGS

4.1 Data Description

The data used in this study have been obtained from one of the tier I banks in Kenya. The monthly corporate data were randomly sampled from all the 159 branches of their loan portfolios for a period of five years starting from March 2013 until march 2019. The size of the data sample is 1,038 and the total number of defaults is 143.

It contains confidential information for different companies which have been concealed as per the agreement stricken before allowed to access the data. In this research, a borrower(company) is declared having defaulted when its monthly installment is not received for a period of three consecutive months. The number censored in the context of this study refer to the number of companies who opt to offset their loans before the maturity time.

4.2 Credit scoring variables

The process of assigning measurable and comparable numbers of likelihood of default risk is termed as credit risk quantification and the concept is a major frontier in credit risk management. Financial institutions attempt to mitigate the risk of lending to borrowers by performing a credit risk analysis on individuals and companies applying for loans.

Being one of the most well-known and exceptional credit score in United States comprising 90 percent of lending decisions, Fair Isaac Corporation (FICO) which is a method of quantifying and evaluating companies creditworthiness list 3 key variables that strongly predict the probability of a borrower defaulting on its loans.The variables that affect credit risk varies from borrower-specific to market-specific factors. Our research was based on corporate data and there were limited number of variables (borrower-specific)that influenced the behaviour of companies towards loan repayment after loan disbursement and are used to determine credit granting criteria. Therefore, the three variables adopted from FICO whose effects on credit risk were studied are;

1. Financial health of the company.
2. Age of the company.
3. Age of the account.

4.3 Data analysis

1. **Financial health of the company**

Income is the flow of money that comes into the company from owning a business, state benefits, rent on properties etc. It's a better measure of financial health than wealth as it is usually a better indicator of company's resource masculinity. Therefore, financial health of a company is one of the commonly used variable by credit risk analysts to determine whether a company will be granted a loan or not.

Our sample data was categorized according to the company's net worth and there were 10 bands with the first band being a cohort of companies whose income ranges from 1 to 1,000,000 and the last band and highest being a band of companies whose income is above 10,000,000. From the table, it's evident that the highest default percentage was seen on the 2^{nd} band while the lowest was on the 10^{th} band. It is in the 8^{th} band that there were more qualified companies for loans than any other bands.The average default percentage on this variable is 15.76% which is nearly equal to the default percentage of the 4^{th} band. Thus, companies with large income have low tendency of defaulting contrary to those which have small income whose probability of default is high. For proper visual clarity and focus on the main points,a histogram is drawn

Amount in Ksh ('000')	Number of accounts	Defaults	Default percentage
0001-1,000	61	21	34.4
1,001-2,000	83	35	42.2
2,001-3,000	97	27	27.8
3,001-4,000	108	16	14.8
4,001-5,000	139	13	9.4
5,001-6,000	124	8	6.5
6,001-7,000	195	11	5.6
7,001-8,000	89	6	6.7
8,001-9,000	67	5	7.5
Above 9,000	75	2	2.7

Table 2. Table 2

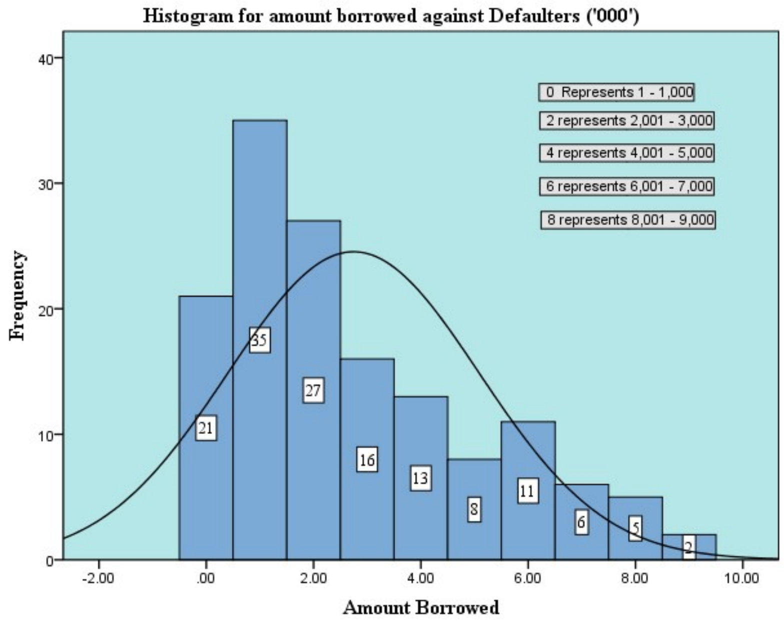

Figure 1. Figure 1

2.**Age of the company account**

This refers to the duration from the time the account was opened to the time of loan vetting. Kenya Bankers Association(KBA) requires prospective clients to have had an account with the lending bank for at least six months prior to loan application time and this serves as a preliminary risk mitigave measure. The duration is normally used to assess the account's credit flow which provides the basis for projecting company's credit worthiness.

Fair Isaac Corporation(FICO) credit scoring model classifies age of the account as a strong predictor of future credit risk and makes about 15% of credit score. Results of our research on the variable shows that those companies whose accounts have been operational in the bank for a period of 2 to 2.99 years have highest default percentage while those whose accounts have been operational for a duration of above 9 years have the lowest default percentage. Average default based on this predictor is 13.0%. Based on the results, it can be seen that there's a decreasing default default trend in conjunction with account longevity except for those companies whose account ages are less than 3 years.

Duration in years	Number of accounts	Defaults	Default percentage
0.50-1.00	58	14	24.1
1.01-2.00	74	20	27.0
2.01-3.00	187	48	25.7
3.01-4.00	88	19	21.6
4.01-5.00	61	10	16.4
5.01-6.00	149	15	10.0
6.01-7.00	125	6	4.8
7.01-8.00	102	5	4.9
8.01-9.00	64	3	4.6
9.01-10.00	77	2	2.6
Above 10 years	53	1	1.9

Table 3. Table 3

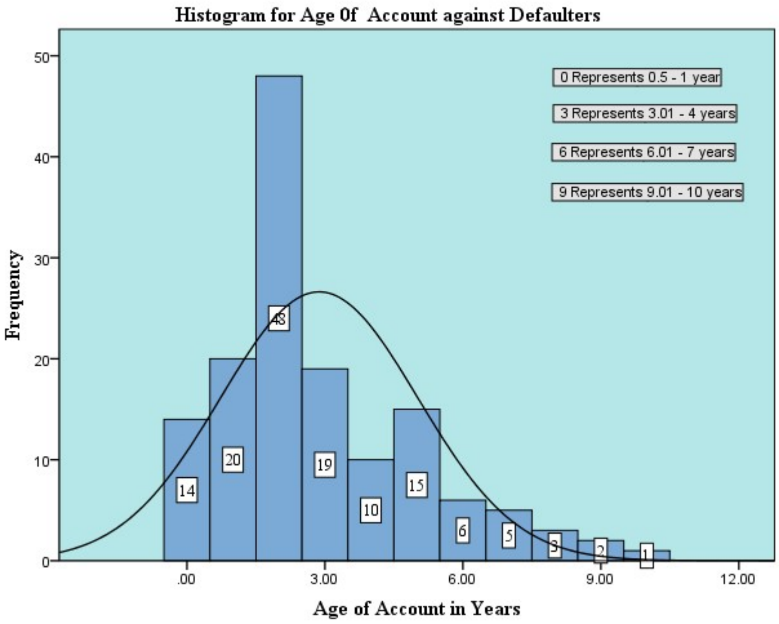

Figure 2. Figure 2

3. **Age of the account**

There are various firm-specific variables influencing the lender's credit granting decision and among them is the age of the company.This refers to the duration since the company was registered. These particulars were obtained from the company search conducted on the registry of companies by the independent legal team outsourced by the bank.The impacts of the firm's age in its repayment performance is very crucial. Therefore, it's one of the variables in our research.

Client's companies were grouped in a band of 10 categories with each band having a duration of 1 year. Band 1 composed of companies which were 'youngest' while the last band was for 'oldest' companies. Many of the companies which the lender awarded loans were on the 8^{th} band and it's 199. Analysis shown that those companies between ages 2.00 and 2.99 have the highest tendency of defaulting while those above 9 years least default. Also the default percentage based on the company's age is 15.8. The highest default in 3^{rd} band can be attributed to low professionalism among the employees,lack of proper business plan,marketing mishaps and lack of enough resources to seek outside professional advice. Below are results in tabular form.

Duration in years	**Number of accounts**	**Defaults**	**Default percentage**
0.50 − 1.00	58	14	24.1
1.01 − 2.00	74	20	27.0
2.01 − 3.00	187	48	25.7
3.01 − 4.00	88	19	21.6
4.01 − 5.00	61	10	16.4
5.01 − 6.00	149	15	10.0
6.01 − 7.00	125	6	4.8
7.01 − 8.00	102	5	4.9
8.01 − 9.00	64	3	4.6
9.01 − 10.00	77	2	2.6
Above 10 years	53	1	1.9

Table 4. Table 4

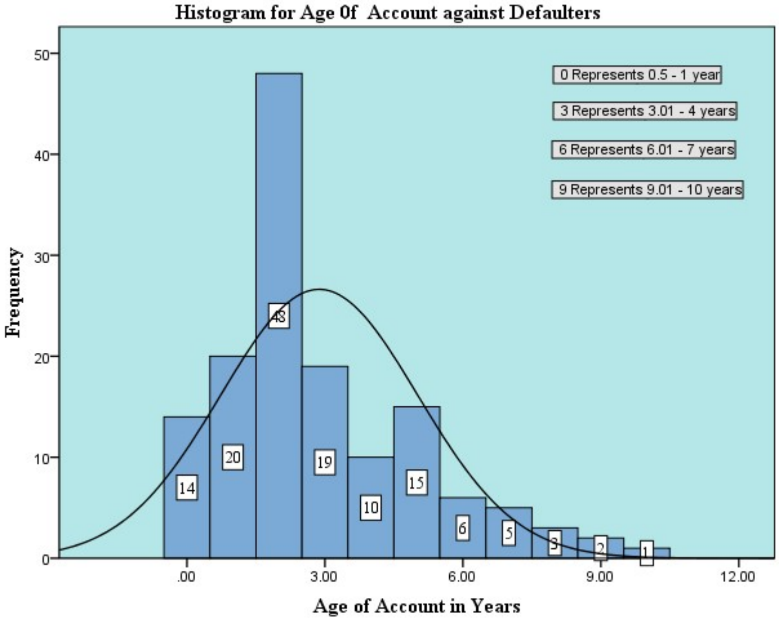

Figure 3. Figure 3

Kaplan Meier and Nelson-Aalen estimators

In the process of estimating the probability of loan defaulting, the non-parametric method via survival analysis was adopted. Kaplan Meier and Nelson-Aalen estimators were computed in R software using the survival and flexsurv packages. 1,038 loaned clients were followed with varying amount of loans for a period of 5 years in which they were expected to have been servicing their loans. Table 5 indicates the number of clients at risk of defaulting, those who actually defaulted and the Kaplan Meir survival estimate at each time period.

At time zero, none of the clients had defaulted on their loans and at time 1, 58 clients had defaulted. The probability of time-to-defaulting at time 1 was 0.9441. This probability gradually decline until at time 60, not all of the clients followed had defaulted, giving 0.0119 defaulting probability. The standard error gradually increased from 0.00713 during the first month to 0.01601 at the 28^{th} month. Thereafter, it gradually declined to 0.00439 at the 60^{th} month. The cumulative hazards estimates indicated a gradual increase from 0.05588 in the first month to 3.60381 at the 60^{th} month.

Time	Number at risk	No. of defaulted	Kaplan Meier estimates	Cumulative hazards	Standard error	Lower 95 percentage CI	Upper 95 percentage CI
1	1038	58	0.9441	0.05587669	0.00713	0.93025	0.9582
2	976	33	0.9122	0.08968816	0.00879	0.89513	0.9296
3	940	40	0.8734	0.13224135	0.01034	0.85335	0.8939
4	897	15	0.8588	0.14896376	0.01083	0.83781	0.8803
5	878	21	0.8382	0.17288176	0.01146	0.81607	0.8610
6	854	9	0.8294	0.18342040	0.01171	0.80676	0.8527
7	843	23	0.8068	0.21070391	0.01231	0.78301	0.8313
8	818	13	0.7940	0.22659633	0.01262	0.76961	0.8191
9	803	9	0.7851	0.23780430	0.01282	0.76033	0.8106
10	792	13	0.7722	0.25421844	0.01310	0.74692	0.7983
11	775	12	0.7602	0.26970231	0.01334	0.73451	0.7868
12	761	13	0.7472	0.28678510	0.01359	0.72106	0.7743
13	745	15	0.7322	0.30691933	0.01386	0.70551	0.7599
14	727	15	0.7171	0.32755206	0.01411	0.68994	0.7453
15	708	20	0.6968	0.35580065	0.01442	0.66912	0.7257
16	682	19	0.6774	0.38365989	0.01469	0.64921	0.7068
17	658	23	0.6537	0.41861430	0.01499	0.62501	0.6838
18	632	27	0.6258	0.4613358	0.01528	0.59656	0.6565
19	600	26	0.5987	0.50466915	0.01552	0.56903	0.6299
20	573	22	0.5757	0.54306356	0.01567	0.54578	0.6073
21	548	19	0.5557	0.57773510	0.01579	0.52564	0.5876
22	521	17	0.5376	0.61036466	0.01587	0.50738	0.5696
23	502	13	0.5237	0.63626107	0.01592	0.49338	0.5558
24	486	9	0.5140	0.65477959	0.01595	0.48365	0.5462
25	473	13	0.4999	0.68226373	0.01599	0.46948	0.5322
26	457	16	0.4824	0.71727467	0.01602	0.45196	0.5148
27	439	18	0.4626	0.75827695	0.01602	0.43221	0.4951
28	418	15	0.4460	0.79416212	0.01601	0.41567	0.4785
29	403	16	0.4283	0.83386435	0.01598	0.39808	0.4608
30	386	18	0.4083	0.88049648	0.01591	0.37828	0.4407
31	366	17	0.3893	0.92694456	0.01582	0.35953	0.4216
32	348	11	0.3770	0.95855376	0.01575	0.34739	0.4092
33	335	14	0.3613	1.00034480	0.01564	0.33187	0.3933
34	320	12	0.3477	1.03784480	0.01554	0.31856	0.3796
35	307	8	0.3387	1.06390344	0.01546	0.30968	0.3704
36	298	9	0.3284	1.09410478	0.01537	0.29966	0.3600
37	286	10	0.3170	1.12906981	0.01525	0.28843	0.3483
38	275	13	0.3020	1.17634254	0.01509	0.27380	0.3330
39	261	13	0.2869	1.22615097	0.01490	0.25916	0.3177
40	248	10	0.2754	1.26647355	0.01474	0.24793	0.3058
41	237	14	0.2591	1.32554528	0.01450	0.23218	0.2891
42	222	12	0.2451	1.37959933	0.01427	0.21866	0.2747
43	208	14	0.2286	1.44690703	0.01397	0.20279	0.2577
44	191	14	0.2118	1.52020546	0.01365	0.18671	0.2403
45	176	13	0.1962	1.59406909	0.01331	0.17176	0.2241
46	162	9	0.1853	1.64962465	0.01306	0.16139	0.2127
47	150	8	0.1754	1.70295798	0.01282	0.15200	0.2024
48	142	9	0.1643	1.76633826	0.01253	0.14148	0.1908
49	132	7	0.1556	1.81936856	0.01229	0.13326	0.1816
50	120	11	0.1413	1.91103523	0.01189	0.11983	0.1667
51	109	6	0.1335	1.96608110	0.01165	0.11254	0.1585
52	102	6	0.1257	2.02490463	0.01140	0.10521	0.1501
53	94	5	0.1190	2.07809612	0.01118	0.09898	0.1431
54	88	9	0.1068	2.18036885	0.01075	0.08771	0.1301
55	77	8	0.0957	2.28426495	0.01032	0.07749	0.1183
56	67	14	0.0757	2.49322018	0.00945	0.05930	0.0967
57	51	8	0.0638	2.65008292	0.00885	0.04866	0.0838

On the other hand, Table 6 shows the number of clients at risk of defaulting, those who actually defaulted and the Nelson Aalen survival estimate for the 60 months period. At time zero, none of the clients had defaulted on their loans and at time 1, 58 clients had defaulted. The probability of time-to-defaulting at time 1 was 0.9457 which is slightly higher than 0.9441 obtained for Kaplan Meier estimate. This probability gradually decline until at time 60, not all of the clients followed had defaulted, giving 0.0272 defaulting probability that is also higher than in Kaplan Meier estimation. The standard error gradually increased from 0.00694 during the first month to 0.01596 at the 28^{th} month. Thereafter, it gradually declined to 0.00588 at the 60^{th} month. The standard error values were also higher for Nelson Aalen estimation. The cumulative hazards estimates indicated a gradual increase from 0.05588 in the first month to 3.60381 at the 60^{th} month.

Time	Number at risk	No. of defaulted	Nelson Aalen	Cumulative hazards	Standard error	Lower 95 percentage CI	Upper 95 percentage CI
1	1038	58	0.9457	0.05587669	0.00694	0.9322	0.9594
2	976	33	0.9142	0.08968816	0.00860	0.8975	0.9312
3	940	40	0.8761	0.13224135	0.01013	0.8565	0.8962
4	897	15	0.8616	0.14896376	0.01064	0.8410	0.8827
5	878	21	0.8412	0.17288176	0.01127	0.8194	0.8636
6	854	9	0.8324	0.18342040	0.01153	0.8101	0.8553
7	843	23	0.8100	0.21070391	0.01213	0.7866	0.8341
8	818	13	0.7972	0.22659633	0.01245	0.7732	0.8220
9	803	9	0.7884	0.23780430	0.01266	0.7639	0.8136
10	792	13	0.7755	0.25421844	0.01294	0.7506	0.8013
11	775	12	0.7636	0.26970231	0.01319	0.7382	0.7899
12	761	13	0.7507	0.28678510	0.01345	0.7248	0.7775
13	745	15	0.7357	0.30691933	0.01372	0.7093	0.7631
14	727	15	0.7207	0.32755206	0.01398	0.6938	0.7486
15	708	20	0.7006	0.35580065	0.01429	0.6732	0.7292
16	682	19	0.6814	0.38365989	0.01457	0.6534	0.7105
17	658	23	0.6580	0.41861430	0.01486	0.6295	0.6877
18	632	27	0.6304	0.46133581	0.01515	0.6014	0.6609
19	600	26	0.6037	0.50466915	0.01539	0.5743	0.6346
20	573	22	0.5810	0.54306356	0.01556	0.5513	0.6123
21	548	19	0.5612	0.57773510	0.01568	0.5313	0.5927
22	521	17	0.5432	0.61036466	0.01577	0.5131	0.5750
23	502	13	0.5293	0.63626107	0.01583	0.4991	0.5612
24	486	9	0.5196	0.65477959	0.01587	0.4894	0.5516
25	473	13	0.5055	0.68226373	0.01591	0.4752	0.5376
26	457	16	0.4881	0.71727467	0.01595	0.4578	0.5204
27	439	18	0.4685	0.75827695	0.01596	0.4382	0.5008
28	418	15	0.4520	0.79416212	0.01596	0.4217	0.4843
29	403	16	0.4344	0.83386435	0.01593	0.4042	0.4667
30	386	18	0.4146	0.88049648	0.01587	0.3846	0.4469
31	366	17	0.3958	0.92694456	0.01579	0.3660	0.4280
32	348	11	0.3834	0.95855376	0.01573	0.3538	0.4156
33	335	14	0.3678	1.00034480	0.01564	0.3383	0.3997
34	320	12	0.3542	1.03784480	0.01554	0.3250	0.3860
35	307	8	0.3451	1.06390344	0.01547	0.3161	0.3768
36	298	9	0.3348	1.09410478	0.01539	0.3060	0.3664
37	286	10	0.3233	1.12906981	0.01528	0.2947	0.3547
38	275	13	0.3084	1.17634254	0.01513	0.2801	0.3395
39	261	13	0.2934	1.22615097	0.01495	0.2655	0.3242
40	248	10	0.2818	1.26647355	0.01480	0.2543	0.3124
41	237	14	0.2657	1.32554528	0.01457	0.2386	0.2958
42	222	12	0.2517	1.37959933	0.01435	0.2251	0.2814
43	208	14	0.2353	1.44690703	0.01407	0.2093	0.2646
44	191	14	0.2187	1.52020546	0.01376	0.1933	0.2474
45	176	13	0.2031	1.59406909	0.01344	0.1784	0.2312
46	162	9	0.1921	1.64962465	0.01320	0.1679	0.2198
47	150	8	0.1821	1.70295798	0.01298	0.1584	0.2094
48	142	9	0.1710	1.76633826	0.01271	0.1478	0.1978
49	132	7	0.1621	1.81936856	0.01248	0.1394	0.1885
50	120	11	0.1479	1.91103523	0.01210	0.1260	0.1736
51	109	6	0.1400	1.96608110	0.01188	0.1186	0.1653
52	102	6	0.1320	2.02490463	0.01164	0.1111	0.1569
53	94	5	0.1252	2.07809612	0.01143	0.1047	0.1497
54	88	9	0.1130	2.18036885	0.01101	0.0934	0.1368
55	77	8	0.1018	2.28426495	0.01061	0.0830	0.1249

Kaplan Meier and Nelson Aalen survival estimators

From the Kaplan Meier (KM) and Nelson Aalen (NA) survival estimators computed, KM estimates were slightly lower than the NA estimates for the 60 period recorded. Likewise in their standard errors, NA estimates had a slightly higher standard error as compared to KM estimates standard errors. Conversely, their cumulative hazards from the first to the last month were similar as tabulated in Table 7 below.

Time	Number at risk	No. of defaulted	Kaplan Meier			Nelson aalen		
			Survival probability	Cumulative hazard	standard error	Survival probability	Cumulative hazard	standard error
1	1038	58	0.94412331	0.05587669	0.007129039	0.94565574	0.05587669	0.006938246
2	976	33	0.91220111	0.08968816	0.008790905	0.91421623	0.08968816	0.008599175
3	940	40	0.87338404	0.13224135	0.010339686	0.87612951	0.13224135	0.010132216
4	897	15	0.85877896	0.14896376	0.010832645	0.86160034	0.14896376	0.010636000
5	878	21	0.83823869	0.17288176	0.011463422	0.84123708	0.17288176	0.011274693
6	854	9	0.82940479	0.18342040	0.011714707	0.83241814	0.18342040	0.011533355
7	843	23	0.80677572	0.21070391	0.012308726	0.81001387	0.21070391	0.012132172
8	818	13	0.79395410	0.22659633	0.012616345	0.79724254	0.22659633	0.012447224
9	803	9	0.78505548	0.23780430	0.012818889	0.78835696	0.23780430	0.012655978
10	792	13	0.77216947	0.25421844	0.013097216	0.77552238	0.25421844	0.012940851
11	775	12	0.76021330	0.26970231	0.013341445	0.76360678	0.26970231	0.013191241
12	761	13	0.74722674	0.28678510	0.013591039	0.75067303	0.28678510	0.013446699
13	745	15	0.73218190	0.30691933	0.013861420	0.73570995	0.30691933	0.013722446
14	727	15	0.71707499	0.32755206	0.014113565	0.72068577	0.32755206	0.013979761
15	708	20	0.69681863	0.35580065	0.014423394	0.70061227	0.35580065	0.014292766
16	682	19	0.67740580	0.38365989	0.014693076	0.68136313	0.38365989	0.014566284
17	658	23	0.65372748	0.41861430	0.014986075	0.65795792	0.41861430	0.014860932
18	632	27	0.62579925	0.46133581	0.015279318	0.63044093	0.46133581	0.015153483
19	600	26	0.59868128	0.50466915	0.015515196	0.60370528	0.50466915	0.015391136
20	573	22	0.57569526	0.54306356	0.015674361	0.58096570	0.54306356	0.015556143
21	548	19	0.55573502	0.57773510	0.015785636	0.56116792	0.57773510	0.015675001
22	521	17	0.53760164	0.61036466	0.015871394	0.54315277	0.61036466	0.015768942
23	502	13	0.52367968	0.63626107	0.015923147	0.52926762	0.63626107	0.015829062
24	486	9	0.51398191	0.65477959	0.015953027	0.51955657	0.65477959	0.015866151
25	473	13	0.49985556	0.68226373	0.015988444	0.50547144	0.68226373	0.015909648
26	457	16	0.48235514	0.71727467	0.016016096	0.48808062	0.71727467	0.015945213
27	439	18	0.46257748	0.75827695	0.016023452	0.46847293	0.75827695	0.015960269
28	418	15	0.44597781	0.79416212	0.016011411	0.45195977	0.79416212	0.015956976
29	403	16	0.42827150	0.83386435	0.015975903	0.43436749	0.83386435	0.015930356
30	386	18	0.40830029	0.88049648	0.015909296	0.41457703	0.88049648	0.015872682
31	366	17	0.38933552	0.92694456	0.015821287	0.39576109	0.92694456	0.015794584
32	348	11	0.37702894	0.95855376	0.015750301	0.38344704	0.95855376	0.015733438
33	335	14	0.36127251	1.00034480	0.015644904	0.36775262	1.00034480	0.015638525
34	320	12	0.34772479	1.03784480	0.015539353	0.35421727	1.03784480	0.015543348
35	307	8	0.33866355	1.06390344	0.015461126	0.34510608	1.06390344	0.015473723
36	298	9	0.32843546	1.09410478	0.015365485	0.33483923	1.09410478	0.015387149
37	286	10	0.31695170	1.12906981	0.015251327	0.32333388	1.12906981	0.015282481
38	275	13	0.30196853	1.17634254	0.015085879	0.30840466	1.17634254	0.015127278
39	261	13	0.28692795	1.22615097	0.014900064	0.29341979	1.22615097	0.014952175
40	248	10	0.27535828	1.26647355	0.014741597	0.28182371	1.26647355	0.014804036
41	237	14	0.25909239	1.32554528	0.014497614	0.26565806	1.32554528	0.014571502
42	222	12	0.24508739	1.37959933	0.014266540	0.25167937	1.37959933	0.014352510
43	208	14	0.22859113	1.44690703	0.013970925	0.23529693	1.44690703	0.014070027
44	191	14	0.21183576	1.52020546	0.013645689	0.21866696	1.52020546	0.013759401
45	176	13	0.19618880	1.59406909	0.013309959	0.20309751	1.59406909	0.013439946
46	162	9	0.18528942	1.64962465	0.013056959	0.19212201	1.64962465	0.013202076
47	150	8	0.17540732	1.70295798	0.012819520	0.18214395	1.70295798	0.012979084
48	142	9	0.16428995	1.76633826	0.012531196	0.17095785	1.76633826	0.012706138
49	132	7	0.15557761	1.81936856	0.012291714	0.16212809	1.81936856	0.012480374
50	120	11	0.14131633	1.91103523	0.011893327	0.14792717	1.91103523	0.012098937
51	109	6	0.13353745	1.96608110	0.011654923	0.14000445	1.96608110	0.011875301
52	102	6	0.12568230	2.02490463	0.011401989	0.13200643	2.02490463	0.011637012
53	94	5	0.11899707	2.07809612	0.011180601	0.12516829	2.07809612	0.011428866
54	88	9	0.10682692	2.18036885	0.010747923	0.11299984	2.18036885	0.011013480
55	77	8	0.09572802	2.28426495	0.010322765	0.10184890	2.28426495	0.010608256

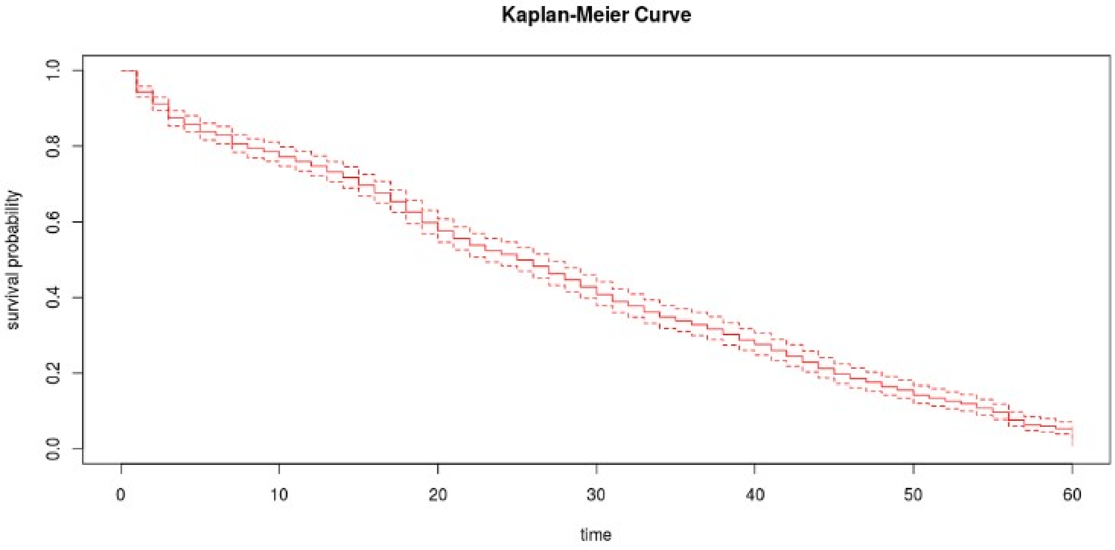

Figure 4. Figure 4

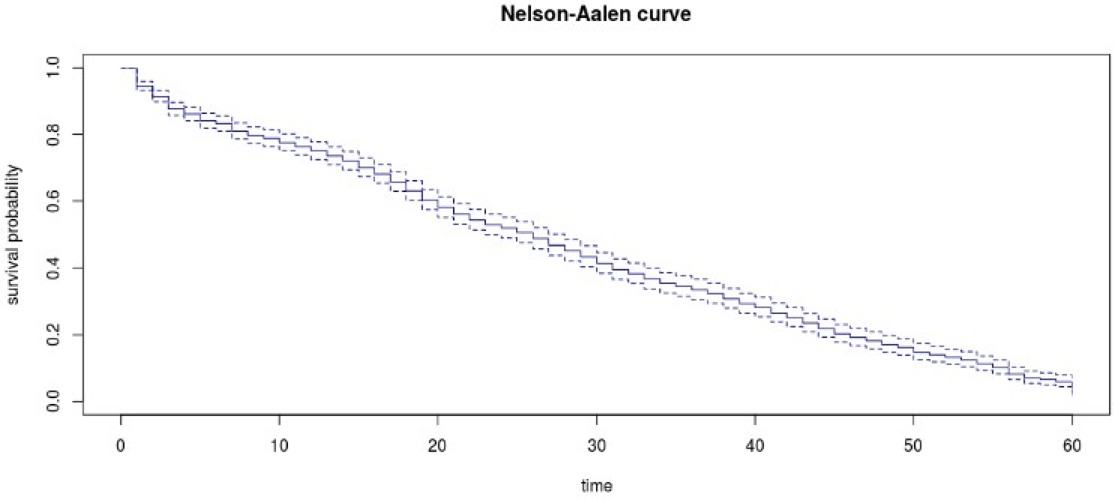

Figure 5. Figure 5

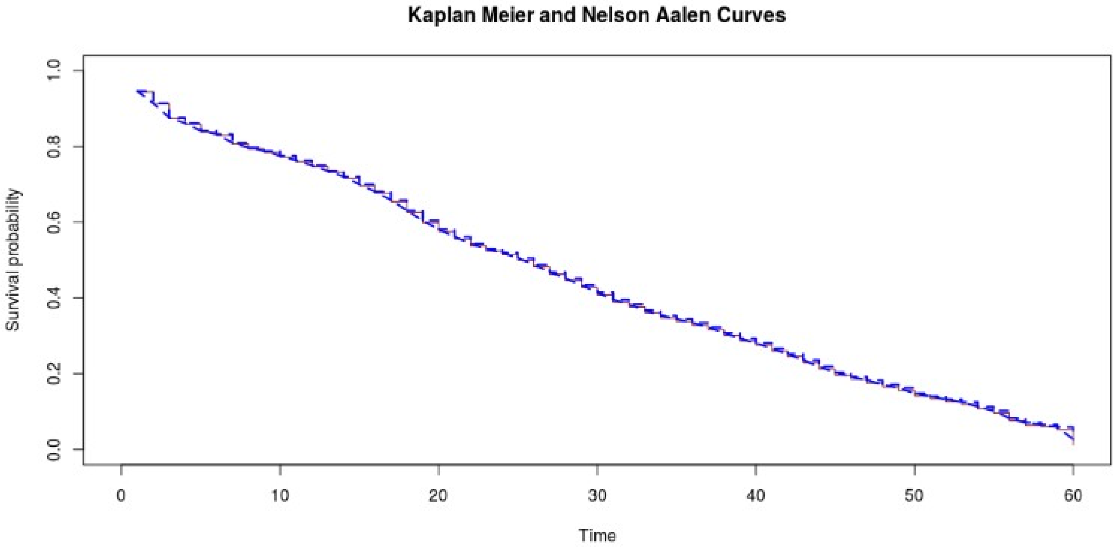

Figure 6. Figure 6

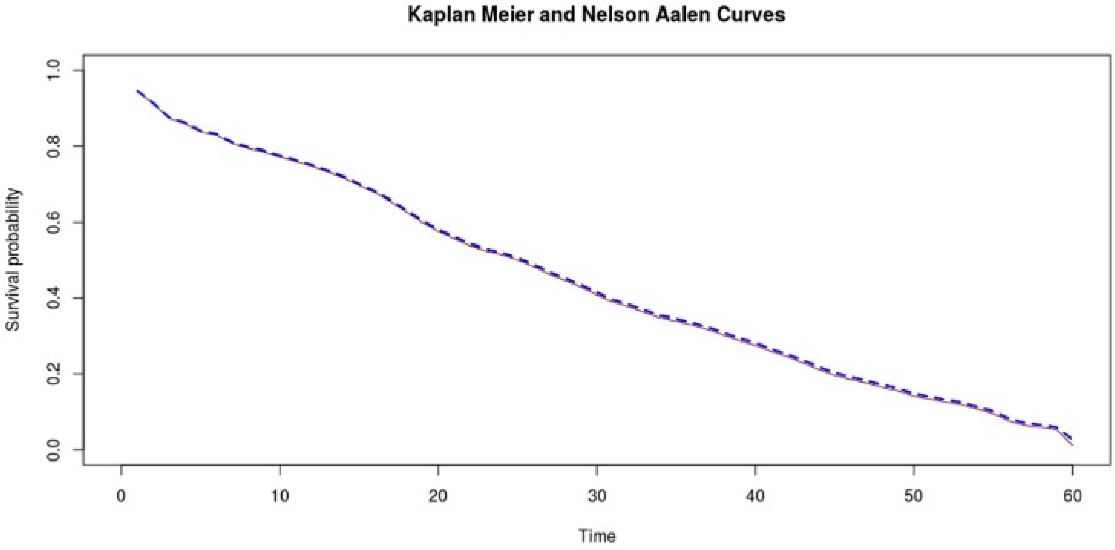

Figure 7. Figure 7

From Kaplan Meier and Nelson Aalen plots, they are almost tracing the same line from the beginning at month one to the end of the 60^{th} month. Figure 4 and Figure 5 illustrates the individual plots while Figure 6 and Figure 7 shows Nelson Aalen curve superimposed on Kaplan Meier curve that produces almost the same lines.
However, Nelson Aalen curve is slightly above Kaplan Meier curve.

4.4 Log rank test

With regards to the differences in survival curves for those clients who had loans below and above 2 million Kenyan shillings, log rank p_{value} of 1 indicated that the two groups did not differ significantly from each other and may not influence the rate of defaulting or repayment.

Chisq=0 on 1 degree of freedom, $p_{value} = 1$.

	N	Observed (O)	Expected (E)	$\frac{(O-E)^2}{E}$	$\frac{(O-E)^2}{V}$
Below 2 Million	416	317	372	0.00147	0.00268
Above 2 Million	622	524	523	0.00104	0.00268
Total	1038	841	895		

Table 5. Log rank table

Standard error is the goodness-of-fit that indicates the reliability of the estimate average and the smaller the SE,the clear the indication that the estimate average is a more accurate reflection of the actual default. Based on our analysis, the SE of the Kaplan-meier is relatively higher compared to the SE of Nelson-Aalen for the first 33 months. The two estimates have almost equal SE for the 4 months from 34 to 37 then Nelson Aalen's SE was higher than Kaplan Meier for the remaining period loan period.

From these results, it's evident that Nelson Aalen estimate is better than Kaplan meier estimate for determining time to default within a loan duration of close to 3 years. It's also established that the two estimates produces same results on the probability of default on $34^{th} - 37^{th}$ month. However,Kaplan-Meier produces more accurate probability of default from 38^{th} month to 60^{th} month reflected in smaller SE in relation to Nelson Aalen.

Thus, a more consistent measure of goodness-of-fit was obtained by making use of a total difference of 99 percent confidence interval (CI) and the mean absolute deviations a selection criteria for choosing the best non-parametric estimator. Analysis shows that the Nelson Aalen estimator is better than the Kaplan Meier estimator exhibited by the smaller value of 99 percent CI difference and the MAD for the estimates of the standard error,survival function and cumulative hazard functions. Results are shown in the table 9.

	Selection criteria	Kaplan Meier	Nelson Aalen
Standard error of $S\hat{(}t)$	99 percent CI diff	0.0287373	0.0245017
	MAD	0.0339623	0.0295201
Survival function,$S\hat{(}t)$	99 percent CI diff	0.1792680	0.1782078
	MAD	0.2333197	0.2319854
Cumulative hazard function,$H\hat{(}t)$	99 percent CI diff	0.5701079	0.5473633
	MAD	0.6648901	0.6608129

Table 6. Goodness-of-fit

Conclusion and Recommendation

5.1 Conclusion

The primary aim of survival analysis models is to find and interpret the survival functions of the survival data and circumvent the issues arising of incomplete information regarding time until the event of interest occur. Thus, it's an instrumental tool for credit risk analysts in their quest to find an ideal risk management and mitigative methods. This research delved on time to default for loans obtained from one of the tier 1 bank loan portfolio in Kenya. Credit scoring variables applicable for corporate data were considered and number of client defaults were fitted on non-parametric models to establish which was a better estimator.

Our study found that there are limited number of variables that were borrower-specific that influences the company's repayment performance. The three variables considered based on FICO included financial health of the company (income of the company) ,age of the company and age of the account. It was evident that oldest companies whose accounts were opened more than 8 years before loan application have lower tendency of default.Log rank test showed that financial health of a company may not influence the rate of defaulting or repayment.

The study shown also that between the two commonly used non parametric models, Nelson Aalen is a better estimator of time to default compared to Kaplan-Meier. This was confirmed by smaller values of $0.0245017, 0.1782078$ and 0.5473633 for standard error,survival function and cumulative hazard function respectively at 99 percent CI difference for Nelson Aalen in comparison to values $0.0287373, 0.1792780$ and 0.5701079 for standard error,survival function and cumulative hazard function respectively for Kaplan Meier. Therefore, it's clear that Nelson Aalen provides a more reliable estimate that reflect the true estimate of time to default.

5.2 Recommendation

This study was confined to microeconomics variables that influence loan repayment performance and it recommends more study to incorporate macroeconomic variables in order to establish their impacts on client loan repayment performance.

In addition,having estimated time to first default on corporate data using non parametric estimation models,this research also recommends further study on estimating time to second default which can be conducted using parametric, non parametric and semi parametric models.It will also be interesting to further extend this studies to the mixture curse model and study the performance of the resulting model in comparison with Cox proportional hazard model with penalized splines as our study involved univariate method.

REFERENCES

[1] Altman E. I. (2002). Revisiting Credit Scoring Models in a Basel II Environment, Prepared for "Credit Rating: Methodologies, Rationale, and Default Risk".

[2] Asia S. and Zaidi, F. B. (2012). Design and development of credit scoring model for the commercial banks of Pakistan: Forecasting creditworthiness of individual borrowers. *International Journal of Business and Social Science*, 3(17), pp 112-116.

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

[4] Bellotti T. and Crook J. (2009) Credit scoring with macroeconomic variables using survival analysis', *Journal of the Operational Research Society*, vol. 60, no. 12, pp. 1699-1707.

[5] Dirick L., Claeskens G., and Baesens B. (2017). Time to default in credit scoring using survival analysis: a benchmark study. *Journal of the Operational Research Society*, 68(6), pp 652-665.

[6] Jaber J., Ismail N., and Ramli S. N. M. (2017). Credit Risk Assessment Using Survival Analysis For Progressive Right-Censored Data: A Case Study in Jordan. *The Journal of Internet Banking and Commerce*, 22(1), pp 1-18.

[7] Kleinbaum, D. G., and Klein M. (2012). Kaplan-Meier survival curves and the log-rank test. In Survival analysis, Springer, New York, pp 55-96.

[8] Kleinbaum D. G., and Klein M. (2005). Competing risks survival analysis. Survival Analysis: A self-learning text, pp 391-461.

[9] Mingxin Li (2014). Residential mortgage probability of default models and methods. *The Journal of Financial Institutions Commission*, British Columbia, 4(7), pp 233-240.

[10] Pelea'ez S. R., Abad, R. C., and Fernández, J. M. V. (2019). Probability of default estimation in credit risk using a non parametric approach, pp 1-23.

[11] Wekesa A. and Okumu W. (2012). Modelling credit risk for personal loans using product-limit estimator (Doctoral dissertation, University of NAIROBI).