2019

Comparison of survival analysis approaches to modelling credit risks

Sammy M. Mungasi
Strathmore Institute of Mathematical Sciences (SIMS)
Strathmore University

Comparison of Survival Analysis Approaches to Modelling Credit Risks

Mungasi, Sammy Monyoncho

Submitted in partial fulfillment of the requirements for the Masters of Science in Mathematical Finance at Strathmore University

Strathmore Institute of Mathematical Sciences
Strathmore University
Nairobi, Kenya

June, 2019

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the dissertation contains no material previously published or written by any other person except where due reference is made in the dissertation itself.

Mungasi, Sammy Monyoncho

.....

June 3^{rd}, 2019.

Approval

This dissertation of Mungasi, Sammy Monyoncho was reviewed and approved by

Dr. Collins Odhiambo,
Lecturer - Strathmore Institute of Mathematical Sciences,
Strathmore University.

Mr. Ferdinand Othieno,
Dean Institute of Mathematical Sciences,
Strathmore University.

Professor Ruth Kiraka
Dean, School of Research Graduate Studies
Strathmore University.

Abstract

Credit risk is a critical area in finance and has drawn considerable research attention. As such, survival analysis has widely been used in credit risk, in particular, to model debt's time to default mechanisms. In this study, we revisit different survival analysis approaches as applied in credit risk defaulters' data and assess their performance in light of the Kenyan context. In practice, inconsistency in the validity of credit risk models used by many companies when predicting and analysis of loan default is a common phenomenon that occurs unexpectedly. Loan defaults often cause major loses to creditors' and can be of great benefit if quantified correctly in advance by using correct models. Here, we address the unbiasedness, analysis, and comparison of survival analysis approaches, particularly, the models of credit risk. We carry out data analysis using the Cox proportional hazard model and its extensions as well as the mixture cure and non-cure model. We then compare the results systematically by investigating the most efficient and preferable model that produces best estimates in the Kenyan real data sets. Results show the Cox Proportional Hazard (Cox PH) model is more efficient in the analysis of Kenyan real data set compared to the frailty, the mixture cure, and non-cure model.

Contents

List of Figures

List of Tables

List of Abbreviations

AFT	Accelerated Failure Time
AUC	Area Under the Curve
AIC	Akaike information criterion
CBK	Central Bank of Kenya
IRB	Internal Rating-Based
IFRS	International Financial Reporting Standards
LGD	Loss Given Default
NPLs	Non-Performing Loans
PD	Probability of Default
PH	Proportional Hazard
RFM	Reduced Form Models
SFM	Structural Form Models
VaR	Value at Risk
ILRFM	Individual Level Reduced Form Model
ROC	Receiver Operating Characteristic
PRFM	Portfolio Reduced Form Model
FM	Factor Model

Acknowledgements

First, I would like to thank the Almighty God for the strength, good health, life and seeing me through the study. Secondly, special thanks to my supervisor Dr. Collins O. Odhiambo for his continuous support and guidance. Lastly, to my classmates, friends and my family thank you for always being there for me and for your support

God bless you all.

Dedication

I dedicate this dissertation to the almighty God above for the gift of life, keeping me healthy, the strength and seeing me through my academic studies. I also dedicate this dissertation to my parents for being patient enough and believing in me and my classmates for their support during the studies.

Chapter 1

Introduction

1.1 Background To The Study

1.1.1 Credit Risk

Credit risk is described as the danger of default on a debt that may arise when a borrower fails to make contractual remittance of payments. Furthermore, Credit risk arises in the case that two counter-parties engage in borrowing and lending (Jarrow,2009). "Obligor" a counter-party who has a financial obligation; for example, a debtor who owes us money, a bond issuer who promises interest, or a counter-party in a derivatives transaction. "Default" failure to fulfill part of the bargain that is obliged, for instance, failure to repay the specified loan or interest/coupon on a loan/bond; generally due to lack of liquidity or insolvency; may entail bankruptcy.

Credit risk remains a critical area both in banking and other lending institutions and is of great concern to many stakeholders,i.e, borrowers, institutions, and policy regulators. Since the advent of Value at Risk (VaR) models in the 1990's, VaR has led to the evolution of risk management practices across the globe. This consequently led to the famous, Basel Committee of 1998, which allowed banks to seek mandatory supervisory approval for putting up capital requirements for market risks with respect to their internal models.

In order to mitigate the adverse effects associated with credit risk, the credit institutions need to first quantify the risks then come up with policies pertaining to risk management. Consequently, the purpose of this dissertation is to review existing survival models in literature and systematically assess their performance using real data from the Kenyan setting.

1.1.2 Kenya's Credit Sector Situation

Currently, Kenya has 42 commercial banks, 8 representative offices of foreign banks, 13 microfinance banks, and one mortgage finance company, (CBK, 2016). According to the Central bank of Kenya 2016 annual report, the banking sector's performance was resilient. The resilience of the economy was attributed to the resilience of Micro Small and Medium Enterprises (MSMEs), public investment in infrastructure, increased growth of the digital economy, and strong performance of the tourism sector.

However, despite being resilient, the public national debt and publicly guaranteed debt by Kenya, increased by 14.3 percent during the financial year 2017/18, with both domestic and external debt increasing at 17.4 percent and 11.6 percent, respectively. Public debt portfolio comprised of 49.2 percent and 50.8 percent domestic and external debt respectively by the end of the financial year 2017/18. The ratio of public debt to GDP declined marginally to 57 percent at end-June 2018 as the projected rate of economic expansion surpassed the rate of build-up in public debt.

There has been an increase in non-performing loans as well. In a Credit Survey Report for the Quarter ended March 2018, the ratio of gross non-performing loans to gross loans grew from 10.66 percent in 2017 December, to 11.81 percent in March 2018. This was attributed to a slowdown in business activities. The ratio of core capital to total risk-weighted assets also rose slightly from 16.05 percent in 2017 December, to the tune of 16.15 percent as of March 2018.

Given the above situation, credit providers required a more sophisticated credit risk measurement techniques that can better assess the risks of their clients. The country has currently adopted logistic regressions which model a dichotomous variable before predicting good and bad clients. This, however, does not inform subsidies on when the initial classification

will depreciate over time. Moreover, with the adoption of the International Financial Reporting Standards 9 (IFRS9), credit providers will be required to forecast not only likelihoods for default but also the time to default.

1.1.3 Survival Analysis

The pioneer of using survival analysis in the context of credit risk is documented to be Narain (1992); where he proposed a survival analysis approach as an improvement to logistic regression. Thomas et al, (2002) highlight that the critical reason for using survival analysis in credit risk setting is that default time could be modeled with other determining features. Consequently, several authors have trailed the same approach and have even utilized more advanced methods. The Cox proportional hazard (Cox PH) model was the first substitute to the AFT model (Thomas et al, 1999) due to its flexible non-parametric baseline hazard. When identifying defaulters in the first year, the research suggested that the Cox PH models are relatively competitive compared to the logistic regression approach, and correspondingly superior to that approach for determining significant covariates. Several authors have extended both Cox PH and AFT models. For instance, modeling the probability of default and early payment of loans (Stepanova and Thomas, 2002) suggests that categorizing continuous variables and using survival analysis techniques in ordering discrete ones, are more appropriate than the conventional technique of using risk ratios.

1.2 Problem Definition

Given the emerging of many financial services provider firms and the growth in the financial industry, there is a need for efficient credit risk measurement techniques. Standards such as the Basel II and International Financial Reporting Standards 9 (IFRS9) have furthermore implied the need for effective credit risk measurement models.

Apart from Omoga, A.A. (2017), there has not been a systematic performance contrasting of different models in the Kenyan context. For instance in the Kenyan market the studies carried out on the probability of default by Wekesa, Samuel, and Peter (2012) use the product-life estimator, Wagacha and Othieno (2016) use a Semi-Markov approach, Obuda

(2016) uses the Cox PH model, and Gitahi and Othieno (2017) uses the Cox PH rate model.

Most authors have not tried to fit different models and compare their performance. Thus, in this study, we fit different survival models into the Kenyan credit data and systematically compare the performance of models in order to evaluate the best model that takes into consideration the time-varying aspect and produces effective results.

1.3 Objectives of the study

- Fit different survival models to Kenyan credit data.
- Systematically investigates the most efficient and preferable model that produces best estimates in Kenyan real data sets.

1.4 Research Question

The question that we intend to answer at the end of the study is which model is most effective in modeling credit risk according to Kenyan real data set.

1.5 Significance of the Study

Results obtained from this study will be mainly useful to the commercial banks in computing the probability of default(behavior scoring) as well as the time to default(profit scoring) of their different clients. This is of importance since the adoption of International Financial Reporting Standards 9 (IFRS9) and the IRB framework, banks are required to compute both the probability and time to default for asset loss provisioning as well as capital requirements assessment.

This work attempts to evaluate and give recommendations on the best survival model in credit risk context in relation to the Kenyan data. This will help financial institutions that are credit providers in assessing their clients' risk of default thus lowering their chances of losses arising from default.

Chapter 2

Literature Review

2.1 Introduction

In credit risk management, decisions are made on the basis of the creditworthiness of an individual or institution which is determined through the use of credit scoring models. According to the CBK's annual supervision report (2016), banks and CRBs must work hand in hand to deliver credible credit scores. The consistent credit scored would then be incorporated into pricing models and credit risk appraisal.

In this section, we evaluate the conceptual foundations of credit risk analysis. Section 2.2 reviews the literature on survival analysis while section 2.3 discusses a case for survival analysis in Kenya.

2.2 Survival Analysis

Survival analysis was historically used in the medical and engineering fields where the duration until the occurrence of an event of interest is examined, for instance, the time until death or machine failure (Collett, 2003; Kalbeisch and Prentice, 2002; Cox and Oakes, 1984). Therefore in 1992, Narain pioneered the introduction of using the survival analysis technique in the credit risk sector as an alternative to logit regression due to its importance of modeling the time of default as opposed to only whether or not an applicant would default (Thomas et al, 2002).

Since then improvements have been done in the field, for instance in a Quantitative Finance paper by Jose Angelo Divino, Edna Souza Lima and Jaime Orrillo (2013), they theoretically and experimentally analyzed the possibility of default in the financial market in Brazil considering both the contract and borrowers' specific characteristics and the nation's macroeconomic conditions. A major Brazilian bank availed the data set of 445889 individual contracts of a short run credit operation. They had access to contracts signed between January 2003 and December 2007. However, the year 2007 did not enter the approximation, it was vital in the scrutiny of the forecasting performance of the model. The time-dependent covariates were also monthly in the stated span of time.

The Cox proportional hazard model with time-varying covariates was estimated. The initial experimented outcomes indicated that the possibility of default is sensitive to specic characteristics of both contracts and borrowers as well as macroeconomic conditions. The findings based on theory, on the adverse effects coming from distinct interest rates over the probability of default were affirmed by the data. A decrease in the economy real interest rate, would imply by an expansionist monetary policy, leads banks to assume more credit risks and ease the analysis of borrowers credit history. By expanding credit operations, banks could compensate for financial losses due to a lower real interest rate. This strategy will bring borrowers with a higher probability of default to the financial market. Conversely, higher rates of interest on loans intensify the chances of default because it reduces the borrower's capacity to settle their debt.

Jose Angelo Divino, Edna Souza Lima and Jaime Orrillo (2013), however, warned that the previous results were based on a particular data set and set of variables. They might not hold for other samples or financial assets. The positive relationship between the probability of default and the loan interest rate might also be a result of risk-based pricing when the lenders charge higher rates to those portfolio segments that have historically shown higher default rates.

In recent studies, Dirick et al (2016), analyzes the performance of various survival analysis techniques applied to ten actual credit data sets. The sets of data were acquired from

the UK and Belgian financial institutions consisting of loans of small enterprises and personal loans, with varying loan terms.

In their paper, (Dirick et al, 2016) analyzed ten dierent data sets from ve banks, using dierent classes of models, that is, Cox PH, Parametric/AFT, Nonparametric, AFT/Cox PH + extensions, Multievent mixture cure and Mixture cure, as well as using both statistical (AUC and default time predictions) and economic evaluation measures applicable to all model types considered, the "plain" survival models as well as the mixture cure models.
Since techniques for survival analysis are incapable to cope with missing data, and with several data sets having a significant count of missing inputs they preferred to employ the rule of thumb used in a benchmarking paper by Dejaeger et al (2012).

As a result, for continuous inputs, median imputation was put in use when $\leq 25\%$ of the values were missing, and the inputs were removed if more than 25%was missing. For categorical inputs, a missing value category was created if more than 15% of the values were missing, otherwise, the observations associated with the missing values were removed from the data set.

In their paper, they used two opposing denitions for censoring. First, censored cases are the loans that did not reach their predened end date during the time of data gathering (called "mature" cases) and neither experienced default nor early repayment by this time. According to the second denition, a censored case corresponds to a loan that did not experience default by the moment of data gathering. Early loan settlement and mature cases are marked censored. This kind of censoring is used in models where the default is the only event of interest.

The number of input variables in the resulting data sets did vary from 6 to 31, and the number of observations from 7521 to 80,641. For each observation, an indicator for default, early repayment and maturity were included, taking the value of 1 for the respective event of interest that took place, and 0 for the others (note that only one event type can occur for each observation). For censored observations according to the rst censoring denition, all indicators are zero. According to the second censoring denition, only defaults are considered uncensored.

In terms of our data sets, this means that censoring rates are ranging from around 20 to 85% according to the first definition (used for the multiple event mixture cure model), whereas censoring percentages are not lower than 94.56% up to 98.16% according to the second definition.

A test set consisting of 2/3 and 1/3 of the observations, respectively was reached by splitting each data set randomly. Estimation of the training sets are made on the models, and the corresponding test sets are used for evaluation. For all the models, the software R is used.

In comparison, Cox PH-based models were all proven to work predominantly well, more so a Cox PH model in combination with penalized splines for the continuous covariates. The Cox PH model often outperforms the multiple event mixture cure model. However, the mixture cure model is among the top models using economic evaluation. It does not perform signicantly dierent in most of the cases. This model does not require the survival function to go to zero when time goes to innity as often regarded as appropriate for credit scoring data, making it advantageous. However, the study also notes that finding a suitable evaluation measure to compare survival analysis persisted as an interesting setback, as the AUC did not seem to have the right properties to really differentiate one method from the other.

The fact that, in the existing literature, some questions remain inspired by the researchers. Except for Zhang and Thomas (2012), no attempt has been made primarily to contrast the available methods in one paper. Secondly, in most recent papers conclusions on the type of survival methods to use could not be made explicitly, since only one data set was analyzed. Finally, the assessment remains mostly fixated on classication and the area under the receiver operating characteristics curve (AUC) as presented in most of the papers.

2.2.1 A Case for Survival Analysis in Kenya

Bellotti and Crook (2009), uses time-varying covariates and in their conclusion, they show that when compared to the benchmark survival model and logistic regression, the inclusion of macroeconomic variables progresses the predictive performance of the model. Despite the analysis of the explanatory model giving enlightenment of how each macroeconomic variable contributes to modeling the data, they recommended extensive experimental work to assess the separate effect of each of the macroeconomic variable on the estimation of PD.

In the recent studies, Pauline N., Lucy M. and Collins O. (2018), contributed to the study by analysis different variables that can lead to Higher Education Loans Board (HELB) loan default in Kenya. They perform a quantitative analysis of loan applications by computing the probability of default of students who apply for Kenyan Higher Education Loans Board (HELB) loans to help in financing their studies. They used the information provided in the Kenya Higher Education Loans application forms. Taking into account of different factors leading into student default of the loan as independent, they use multiple logistic regression with the binomial nominal variable defined as a defaulter or a non-defaulter.

They opted for Multiple logistic regression given its ability to predict a nominal dependent variable from one or more independent variables. From their study, they conclude that the amount of loan being reimbursed was the main factor affecting default. However, they were faced with a challenge of lacking time to defaulting variables which are of interest in survival analysis.

Gitahi and Othieno (2017) analyzes the survival probabilities and hazard rates using the Cox PH model on real data sets obtained from the Metropol Credit Reference Bureau then use the probabilities in estimating the probability of default. In their research, they conclude that the Cox PH model can predict with over 60 percent accuracy both in the probability and time of default. However, in the proportionality test, Gitahi and Othieno(2017) use the -log(-log) test which computationally reduces the covariates used to an expression that drops the baseline hazard function and therefore does not involve time. Thus there was a need to do analysis taking into account time-varying covariates. This was addressed by Omago A.(2017)

in his research dissertation *Predictive modeling in credit risk: a survival analysis case.*

In his study Omago A. (2017), fits the Accelerated Failure Time (AFT) Models, Cox proportional hazard (PH) Model and the Mixture Cure Model (MCM) to a data set consisting of 33,238 active credit facilities from a financial institution operating in Kenya. He evaluates the performance of the models using the Area under the Curve (AUC) and financial evaluation using the annuity theory. He concluded that the Cox Proportional Hazard (PH) and the Mixture cure model performed significantly well.

From the study, Omago established that an appropriate valuation measure for relating survival models remained a challenge since the Area under the Curve (AUC) alone does not have the suitable properties show apart from the different survival model. The results from a data set which comprised of a credit facility of the individual unsecured facility from a financial institution based in Kenya, cannot be generalized to another portfolio since only a single sample was analyzed. He, therefore, proposed a survival analysis modeling benchmark on revolving products such as overdrafts and credit cards and the suitability of the mixture cure model to those products where the facility term is long due to its rotating nature. Omago also proposed an extension of the research to a mobile lending scheme such as the Mshwari loans offered by the commercial bank of Africa through Mpesa platform.

Chapter 3

Methodology

The study relies on Cox Proportional Hazard (Cox PH) model and it's extensions, that is, Penalized Splines and Frailty model as well as the mixture cure and non-cure model in developing the probability of default model for a consumer loan portfolio. This chapter contains the research design, data collection, and model framework.

3.1 Study Design

The study contributes to the body of knowledge of credit risk modeling using survival analysis approaches. The study fits the Cox PH model and its extensions, that is, penalized splines and frailty model, as well as the mixture cure and no-cure model to real Kenyan data, set. Assessment is done to ascertain the most effective model in analyzing credit risk.

3.2 Data

The data used for the study was obtained from the Metropol Credit Reference Bureau for the period 2014 to 2017. The data comprised of 20299 individuals and included various covariates namely: the age of the individuals(from 22 years to 75 years),age bracket(18-33, 34-43, 44-53, >54), gender(male and female), marital status(married,single,divorced and widowed),the

type of the account(credit card, loan account and current account),status of the loan (active or defaulted), loan amount(ranged from kshs 1,000.00 to kshs 229,068,599.00) and the loan amount group.

3.3 Survival Analysis Framework

In survival analysis, we are usually concerned with the time variable, T, of an event of interest. The survival function is usually, articulated as the likelihood of not experiencing the incident of concern at some observed time t, hence yielding $S(t) = P(T > t)$. In the setting of credit risk, where the default is the event of interest. See Dirick et al (2016). Given the survival function, the probability density function f(u) is given by

$$f(u) = -\frac{d}{du}S(u) \tag{3.3.1}$$

and the hazard function

$$h(t) = \lim_{\tau \to 0} \frac{P(t \leq T < t + \tau | T > t)}{\tau} = \frac{f(t)}{S(t)}, \tag{3.3.2}$$

where τ is the Δt (change in time).

The hazard function models the instantaneous risk.

When carrying out survival analysis censoring is done, that is, the incident of concern has not been witnessed at the time of assembling data. For instance, Dirick et all (2016) considers two types of censoring, one where some credit applicants had failed to pay, replayed in advance or some loans were completely paid back at the completion of the loan period. Censoring is done to the cases where none of the above events had been observed. In the second scenario, censoring is entirely labeled on the cases that matured or repaid their loans early. Thus only censoring and default states are been considered.

3.3.1 Cox Proportional Hazard Model

The Cox proportional hazard model is more flexible than any accelerated failure time (AFT) model as it contains a non-parametric baseline hazard function, $h_0(t)$, along with a parametric part (Cox, 1972). The Cox model has the advantage of preserving the variable in its original quantitative form, and of using a maximum of information. However, very restrictive conditions of application of this model make its use rather limited (Bugnard F., 1994). The model's hazard function is denoted as;

$$h(t|x) = h_0(t)exp(\beta' x) \tag{3.3.1.3}$$

where the covariate vector is denoted by x and the parameter vector by β'.

The survival function is denorted as;

$$S(t|x) = exp(-exp(\beta' x) \int_0^t h_0(u)du \tag{3.3.1.4}$$

where $\int_0^t h_0(u)du$ can be expressed as $H_0(t)$ which is the cumulative baseline hazard function and can be estimated by Breslow's method as;

$$H_0(t) = \sum_{t_i \leq t} \frac{1}{\sum_{r \epsilon R(t_i)} exp(\beta' \mathbf{x}_r)} \tag{3.3.1.5}$$

with $R(t_i)$ denote the set of people that haven't failed to pay at time t_i

3.3.2 Cox Proportional Hazard Model with Penalized Splines

The hazard function in the Cox PH model assumes a proportional hazards structure with a log-linear model for the covariates. Thus for any continuous variable, e.g., age, the default hazard ratio between 5-10 years is the same as the hazard ratio between 50-55 years. This assumption normally doesn't hold thus splines are used due to their flexible functions defined by piecewise polynomials that are joined in points called "knots." (Therneau and Grambsch, 2000).

When a total sum of knots in a given spline turns out to be sufficiently huge, a fitted function of the spline depicts more variation than justified by the data. The penalized spline can be considered as a variant of smoothing spline with a more flexible choice of knots, bases, and penalties. A smoothness penalty was introduced by O'Sullivan (1986) when he implemented the procedure by incorporating the square of the second derivative of the fitted spline function. Thereafter, Eilers et al. (1996), revealed that this penalty could also be based on higher-order finite differences of adjacent B-splines.

3.3.3 Frailty Model

Vaupel et al. (1979), came up with the term frailty and used it in univariate survival models. Frailty models offer an improved way for integrating random effects in a given model to account for association and heterogeneity that is not observed. Generally, a frailty model can be considered as an unobserved random factor that modifies multiplicatively the hazard function of an individual, group or cluster of individuals. Andreas W. (2010), revealed that frailty proposes an appropriate way of introducing unobserved heterogeneity and associations into models for survival data. In his book Andreas W. (2010), the model is represented by the following hazard given the frailty:

$$\lambda(t|Z, X) = Z\lambda(t|X) \tag{3.3.3.1}$$

Where λ is the hazard function and the frailty Z is an unobservable random variable varying over the sample which increases the individual risk if $Z > 1$ or decreases if $Z < 1$.

The conditional survivor function for the model is presented as:

$$S(t|Z,X) = \exp(Z\int_0^t \lambda(u|X)du) = \exp(Z\Lambda(t|X)), \tag{3.3.3.2}$$

where

$$\Lambda(t|X) = \int_0^t \lambda(u|X)du.S(t|Z,X)$$

represents the portion of individuals surviving until time t given Z and given the vector of observable covariates X.

Until now, the model is described at the individual level, but this individual model is not observable. Hence, it is essential to consider the model at a population level. The survival of the total population is the mean of the individual survival functions

Hougaard (1984) introduced the Laplace transform for these calculations. The Laplace transform of a random variable Z is defined as:

$$L(s) = \int exp(-sz)g(z)dz = E[exp(-sZ)] \tag{3.3.3.3}$$

where g(z) is the density of Z. The integral is over the range of the distribution. The marginal survivor function can be calculated by

$$S(t|X) = \int S(t|Z,X)g(z)dz = E[S(t|Z,X)] = L(\Lambda(t|X)) \tag{3.3.3.4}$$

Univariate frailty models are not identifiable from the survival information alone. However, Elbers and Ridder (1982), proved that a frailty model with finite mean is identifiable with univariate data when covariates are included in the model.

3.3.4 Mixture Cure and Non-Cure Model

Conventionally, mixture cure models have been inspired by the presence of disaggregated long-term survivors (Taylor, 2000; Peng and Dear, 2000).On the other hand, under non-mixture survival models, the incident of concern is anticipated to occur eventually. Both mixture cure and non-cure models are used in the setting where a given fraction of the population under study will not experience the event of interest. Therefore, the mixture cure model can be viewed as a combination of distributions where a logit regression model generates a mixing proportion of non-susceptibility while the survival model, on the other hand, defines the survival function of the cases subject to the event of interest. The models are of particular interest in credit risk modeling as default, which is the main event of interest will not occur for a huge proportion of the cases. This idea was introduced in the credit risk setting for the first time by Tong et al (2012).

The survival function of the mixture cure model is given as;

$$S(t|x) = \pi(x)S(t|Y = 1, x) + 1 - \pi(x)t \tag{3.3.4.1}$$

where Y is the susceptibility indicator ($Y = 1$ if an account is susceptible, and $Y = 0$ if not).

The conditional survival function modeling the cases that are susceptible is given by a Cox proportional hazards model:

$$s(t|Y = 1, x) = exp(-exp(\beta' x) \int_0^t h_0(u|Y = 1)du) \tag{3.3.4.2}$$

In a non-cure mixture context, the Breslow-type estimator is used for estimation of the cumulative baseline hazard similar to the Cox proportional hazards model. Excellent summary on the non-cure mixture model can be found in Tong et al (2012)

3.4 Model Parameters Estimation

3.4.1 The Proportional Hazard Model

The β information is obtained from the orderings of survival times. Let A_i be the incident that the individual i experiences default in $[u.u+\Delta u]$ and $t_i, .., t_u$ define individual default times, then

$$\begin{aligned} P[I(u) = i(u)|F(u) = f(u); \lambda_0(.), \beta] \\ = P[A_{i(u)}|A_1 U ... A_n] \\ = \frac{P[A_{i(u)}]}{\sum_{l=1}^{n} P[A_l]} \\ = \frac{\lambda_0(u) exp(x_{i(u)}^T \beta) \Delta u}{\sum_{i=1}^{n} \lambda_0(u) exp(x_{i(u)}^T \beta) Y_i(u) \Delta u} \\ = \frac{exp(x_{i(u)}^T \beta)}{\sum_{l=1}^{n} exp(x_{i(u)}^T \beta) Y_i(u)} \end{aligned}$$

Where $Y_{i(u)}(u) = 1$ when the individual is at risk at$u$.The Partial Likelihood is expressed as;

$$PL(\beta) = \Pi[\frac{exp(x_{i(u)}^T \beta)}{\sum_{l=1}^{n} exp(x_{i(u)}^T \beta) Y_i(u)}]^{dN(u)}$$

The function is dependent on β, the parameter of interest, and is free of the baseline hazard $\lambda_0(t)$.

We then express the Log partial likelihood function of β as;

$$l(\beta) = \sum dN(u)[x_{I(u)\beta - log(\sum_{l=1}^{n} exp(x_l \beta) Y_l(u))}]$$

The log-likelihood has a novel maximizer and can be gotten by solving the partial likelihood equation

$$U(\beta) = \frac{dl(\beta)}{d\beta} \sum dN(u)[x_{I(u)}\beta - \frac{\sum_{l=1}^{n} exp(x_l\beta)Y_l(u))}{\sum_{l=1}^{n} exp(x_l\beta)Y_l(u))} = 0]$$

3.4.2 The Frailty Model

We use gamma distribution because it is easy to derive the closed form expressions of survival, density and the hazard function. This is due to the simplicity of the Laplace transform. The density function of the gamma distribution gamma $(z; \theta, \beta)$ is given by

$$g(z) = \theta^{\beta} z^{\beta-1} exp(-\theta z | \lceil(\beta))$$

where $\theta > 0, \beta > 0$ and $z > 0.\theta$ is a scale parameter and β is a shape parameter. We define the hazard function

$$\lambda(t_{ij}|Z_i) = Z_i\lambda_0(t_{ij})exp(\beta^t X_{ij}), i = 1, ..., nj = i, ..., k_i$$

as the hazard function of the j^{th} individual of group i given the frailty of group $i(Z_i)$, where $\lambda_0(t_{ij})$ is an arbitrary baseline hazard rate and x_{ij} is the corresponding covariate vector. We denote the joint survival function as

$$S(t_{i1}, ..., t_{iki}) = Pr(T_{i1}, ..., T_{iki} > t_{iki})$$
$$= [1 + \frac{1}{\theta} \sum_{j=1}^{ki} \Lambda_0(t_{ij})exp(\beta^t X_{ij})]^{-\theta}$$

we obtain β, θ and $\Lambda_0(t)$ using the EM (Expectation Maximization) algorithm (Dempster et al., 1977) which provides a means of maximizing complex likelihoods. The likelihood of the frailty is give as $l_{full} = l_1(\theta) + l_2(\Lambda_0)$ where

$$l_1(\theta) = n[\theta log\theta - log \in (\theta)] + \sum_{i=1}^{n} [(D_i + \theta - 1)logZ_i - \theta Z_i]$$

$$l_2(\Lambda_0, \beta) = \sum_{i=1}^{n} \sum_{j=i}^{ki} d_{ij}[\beta^t X_{ij} + log\lambda_0(t_{ij})] - Z_i\lambda_0(t_{ij})exp(\beta^t X_{ij})$$

In the E step we complete the expected value of the full likelihood given the current estimates of the parameters and the observable data. In the M step the estimates of the parameters which maximize the expected value of the full likelihood from the E step are obtained.(Klein and Moeschberger (1997)).

3.4.3 The Mixture Cure Model

Note;

Y_m denotes the matured observation, hence repaid at the maturity date.

Y_d denotes the occurrence of the event of interest, default.

Y_e denotes early repayment.

For a single event we use a semi-parametric regression model where the conditional survival probability at time t is modelled yielding the unconditional survival function and the corresponding observed likelihood;

$$L_{obs}(b, \beta) = \Pi_{i=1}^{n}\{\pi(x_j; b)f(t_i|Y_i = 1, x_i)\}^{\delta_i} * \{(1 - \pi(x_j; b)) + \pi(x_{ji}b)S((t_i|Y_i = 1, x_i; \beta)\}^{1-\delta_i}$$

given full information of Y, the complete likelihood function is given as

$$L_{complete}(b, \beta) = (1 - \pi(x_i; b))^{1-Y_i}(x_i; b)^{Y_i}h(t_i|Y_i = 1, x_i; \beta)^{\delta_i, Y_i}S(t_i|Y_i = 1, x_i; \beta)^{(Y_i)}$$

In a multiple event the three indicators (Y_e, Y_d, Y_m) are used in the formulation of this model.Using the dummy variable '1' denoting a credit event default 'd' and '2' denoting early repayment 'e', the observed likelihood is;

$$L_{obs}(\Theta) = \Pi_{i=1}^{n}\{\pi_{j=1}^{2}\pi_j(x_i; b_j)f_i(t_i|Y_{j,i} = 1, x_{j,i}; \beta_j)^{Y_{j,i}}(1 - \sum_{j=1}^{2}\pi_j(x_i; b_j))^{Y_{m,i}}\}^{\delta_i} *$$

$$\{(1 - \sum_{j=1}^{2}\pi_j(x_i; b_j) + \sum_{j=1}^{2}(x_i; b_j)S_j(t_{ij}|Y_{ij} = 1, x_{ij}; x_{\beta,j})\}^{1-\delta_i}$$

where Θ=$b_e, b_d, \beta_e, \beta_d$. Maximum of the observed likelihood does not exist hence Zeng and Lin (2007), proposed maximization of the Kernel smoothed profile likelihood using an EM Algorithm. The model can then be rewritten starting from the complete likelihood, hence

the likelihood expression under the assumption that the full information on $Y = Y_e, Y_d, Y_m$ is present.

$$L_{complete}(\Theta) = \Pi_{i=1}^{n}\{\Pi_{j=1}^{2}(\pi_j(x_i;b_j))^{Y_{j,i}}(1-\pi_{j=1}^{2}(\pi_j(x_i;b_j))^{Y_{m,i}})*$$

$$\{\Pi_{j=1}^{2}h_j(t|Y_{j,i}=1,x_{j,i};\beta_j)^{\delta_j}S_d(t_{j,i}|Y_{i,j}=1,x_{j,i};\beta_j)^{Y_{j,i}}$$

Using the model density with parameters Θ_1 we compute the expected value by converting the Likelihood to a log likelihood translating to the $Q-function$

$$Q(\Theta_1|\Theta_2) = E_f[logL_{complete}(\Theta_2;\tau_i,\delta_i,\Theta_1)]$$

$$= \sum_{i=1}^{n}\{W_{ji}log(\pi_j(x_i;b_j)) + W_{mi}log(1-\sum_{j=1}^{2}\pi_j(x_i;b_j))+$$

$$\sum_{j=1}^{2}W_{j,i}\delta_i log(h_j(t_i|Y_j=1,x_{j,i};\beta_j)) + W_{ij}log(h_i(t_i|Y_j=1,x_{i,j};\beta_j))\}$$

The conditional expectations of $Y_{i,j}(j=1,2), E_f[Y_i|\tau_i|\delta_i|\Theta_1]$ are calculated with respect to the model density using parameterΘ_1 denoted by W_{ji} with $W_{mi} = 1 - W_{1i} - W_{2i}$ and for j=1,2, $W_{mi} = W_{mi}(\Theta) = P(Y_{i,j}=1|\tau_i=t;\delta_i;\Theta)$

$$= \frac{\pi_j(x_i;b_j)S_j(t_i;\beta_j)}{\sum_{k=1}^{2}\pi_k(x_i;b_k)S_k(t_i;\beta_k)+1-\sum_{k=1}^{2}(x_i;b_k)} \qquad \text{for } \delta_i = 0$$

$$1 \qquad \text{for } Y_{i,j}=1 \text{ and } \delta_i = 1$$

$$0 \qquad \text{for } Y_{i,j}=0 \text{ and } \delta_i = 1$$

Chapter 4

Presentation of Research Findings

4.1 Data Overview

The data used for analysis was from Kenya Metropol and comprised of 20,299 individuals. The data was extracted between December 2014 and December 2017 where different individual accounts were tracked over a certain period of time to obtain the accounts where the individuals went into default as well as those that did go into default within the period of the study. The date captured various covariates, that is, time difference, gender, age, type of the product, marital status, year of data retrieval and the loan amount groups.

The time difference means for active individuals is 24.17 months, with a lower confidence interval of 24.04 months and an upper confidence interval of 24.30 months. For defaulters, the mean time difference is 24.70 months with a lower confidence interval of 24.45 months and an upper confidence interval of 24.94 months. The mean original amount for active individuals is Kshs 300,077.82 with a lower confidence interval of Kshs 247,589.00 and an upper confidence interval of Kshs 352,566.63. For the defaulters, the mean original amount is Kshs 195,210.76 with a lower confidence interval of Kshs 137,670.56 and an upper confidence interval f Kshs 252,750.95. The mean age for active individuals is 41 years with a lower confidence interval of 40.83 years and an upper confidence interval of 41.17 years. For the defaulters, the mean age is 38.11 years with a lower confidence interval of 37.81 years and an upper confidence interval of 38.40 years.

The time (in years) when data was retrieved is December 2014, 2015, 2016 and 2017. The gender for the individuals under study is male and female, we shall denote male as gender 1 and female as gender 2 for our analysis part. The individuals are grouped into the age brackets of 18-33, 34-43, 44-53 and above 54 years. The amounts of the individuals are under different products namely; current account, loan account, and the credit card. The individuals under study are considered to be either married, divorced, single or widowed.

The individuals amount is also grouped in ranges starting from 0-50,000, 50,001-100,000, 100,001-250,000, 250,001-500,000, 500,001-1000,000 and over 1000,000. Note that all the covariates have a p-value of less than 0.001 apart from the marital status which has a p-value of 0.006. The data can be summarized in a table as shown below.

		Active	Defaulter	P-Value
Mean Time Difference (95% CI)		24.17(LCI=24.04, UCI=24.30)	24.70(LCI=24.45, UCI=24.94)	<.001
Mean Original Amount (95% CI)		300,077.82(LCI=247,589.00, UCI=352,566.63)	195,210.76(LCI=137,670.56, UCI=252,750.95)	<.001
Mean Age (95% CI)		41.00(LCI=40.83, UCI=41.17)	38.11(LCI=37.81, UCI=38.40)	<.001
Year of Data Retrieval	2014	68 (71.58%)	27 (28.42%)	<.001
	2015	2827 (81.54%)	640 (18.46%)	<.001
	2016	6486 (77.18%)	1918 (22.82%)	<.001
	2017	6465 (77.58%)	1868 (22.42%)	<.001
Gender	Female	7857 (82.61%)	1654 (17.39%)	<.001
	Male	7989 (74.05%)	2799 (25.95%)	<.001
Age Bracket	18-33	4652 (71.95%)	1814 (28.05%)	<.001
	34-43	4908 (78.73%)	1326 (21.27%)	<.001
	44-53	4136(81.05%)	967 (18.95%)	<.001
	>54	2150(86.14%)	346(13.86%)	<.001
Product Name	Current Account	2419 (78.36%)	668 (21.64%)	<.001
	Loan Account	10877 (83.75%)	2111 (16.25%)	<.001
	Credit Card	2550 (60.37%)	1674 (39.63%)	<.001
Marital Status	Divorced	43 (75.44%)	14 (24.56%)	0.006
	Married	10483 (78.60%)	2854 (21.40%)	0.006
	Single	5311 (77.11%)	1577 (22.89%)	0.006
	Widowed	9 (52.94%)	8 (47.06%)	0.006
Amount Group	0-50,000	9808 (76.55%)	3004 (23.45%)	<.001
	50,001-100,000	1657 (75.66%)	533 (24.34%)	<.001
	100,001-250,000	1473 (78.94%)	393 (21.06%)	<.001
	250,001-500,000	1527 (84.32%)	284 (15.68%)	<.001
	500,001-1,000,000	848 (84.38%)	157 (15.62%)	<.001
	Over 1,000,000	533 (86.67%)	82 (13.33%)	<.001

Table 1: Summary of data covariates with there correspondent counts and p-values

4.2 Performance Evaluation Measure

4.2.1 Aikaike Information Criteria (AIC)

Collett (1994), documents the Aikaike Information Criteria (AIC) for a given model as a function of its maximized log-likelihood (ℓ) and the number of (number of independently adjusted parameters within the model (K)

$$AIC = -2\ell + 2K$$

The criteria used for this study is the Akaike Information Criterion (AIC), as is assigns scores to every single model and provides us which a choice of choosing the model with the best score. The lower the AIC compared to the null deviance, the better the model will be.

Akaike Information Criteria (AIC) provides a versatile procedure for statistical model identification which is free from the ambiguities inherent in the application of conventional hypothesis testing procedure. The fact that the maximum likelihood estimates are under certain regularity conditions, asymptotically efficient shows that the likelihood function tends to be a quantity which is most sensitive to the small variations of the parameters around the true values.

4.2.2 Fisher Scoring

Fisher scoring iteration is concerned with how the model was estimated (Pauline, (2018)). Newton-Raphson iterative algorithm is used by default in R for logistic regression. Based on an approximation of estimates a model if fit and the algorithm explores for an enhanced fit by using alternative approximations. Thus engrosses the same route using higher values for the estimates and fits the model again. The algorithm quits when it notices that searching over can't result in any other additional enhancements. In our model, we had 719 iterations before the process quit and output the results.

4.3 Results

A log rank test was first done using R statistical software before any analysis to check the significance of the variables. Spline graphs, box plots, and the Kaplan-Meier curves were then generated for visualization of variables relationship in determining the probability of default. Thereafter the data was analyzed using survival analysis approaches in the study to determine the most efficient in modeling credit risks.

4.3.1 Log Rank Estimation

Log Rank test estimation was done using R statistical software in an attempt to study the significance of the variables in our data set. The log-rank test is used to test the null hypothesis that there is no difference between the populations in the probability of an event at any time point. The analysis is based on the times of events. The log rank test is based on the same assumptions as of the Kaplan Meier survival curve3—namely, that censoring is unrelated to prognosis, the survival probabilities are the same for subjects recruited early and late in the study, and the events happened at the times specified. Deviations from these assumptions matter most if they are satisfied differently in the groups being compared, for example, if censoring is more likely in one group than another. Because the log rank test is purely a test of significance it cannot provide an estimate of the size of the difference between the groups or a confidence interval.

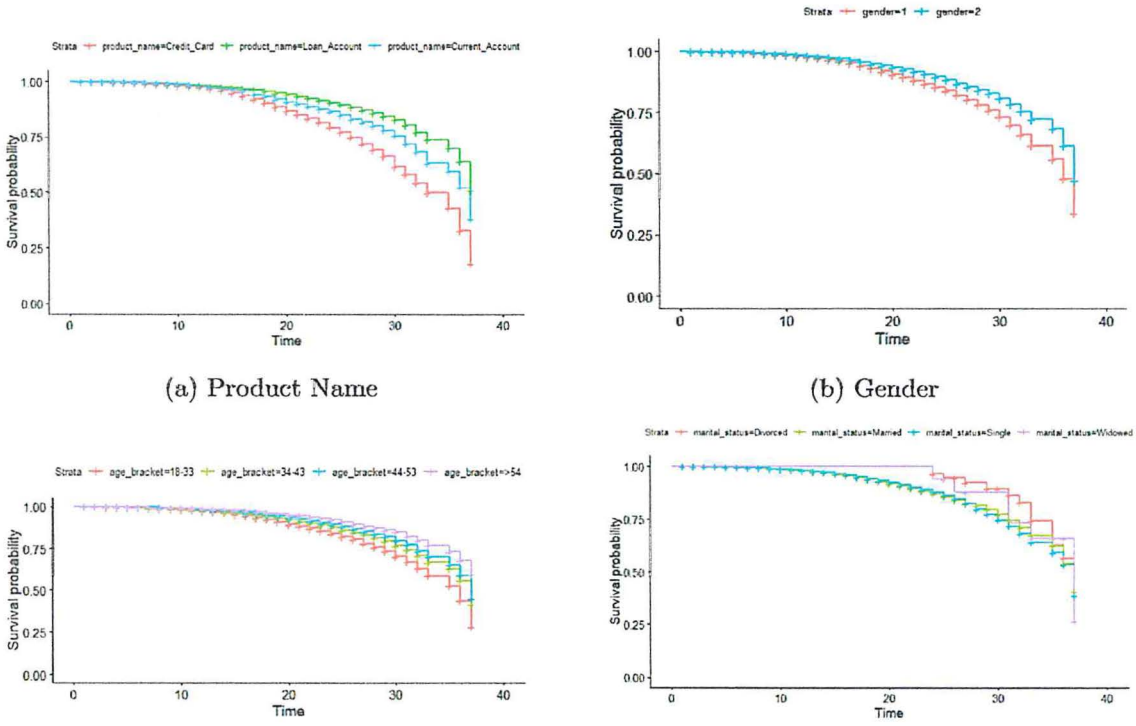
Log Rank Estimate	Chi-Square	df	significance
Gender	159.257	1	0.000
Age Bracket	248.220	3	0.000
Marital Status	4.825	3	0.185

Table 2: Overall Comparison of Log Rank Tests

From Table 2 above, it's clearly evidenced that the variables in our data set are significant to our study since they all had a significant difference value of less than 0.5.

4.3.2 Kaplan-Meier (KM) curves

The Kaplan-Meier was used in generating the variables KM curves. The curves were only obtained for the age bracket, gender, marital status, and product name covariates. Its shown from the KM curves that the young people between the age of 18-33 years, the male, single and individuals with a credit card account are more likely to default a loan as shown below;



(a) Product Name

(b) Gender

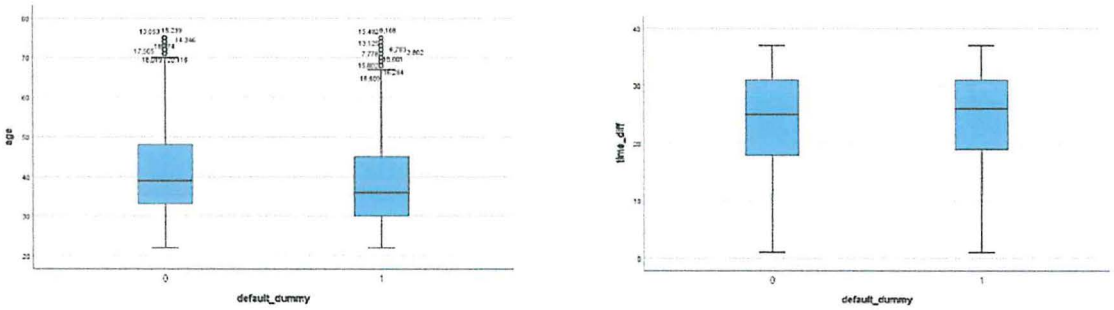
(c) Age Bracket

(d) Marital Status

Figure 4.1: Kaplan Meier Curves

4.3.3 Age and Time Difference Boxplots generated using SPSS version 25.

Age and time difference boxplots were generated using SPSS version 25. From it is evidence that time difference is not a significant variable in determining the probability of default. However, age is a significant variable in determining the probability of default as shown below;



(a) [Age Boxplot (b) Time Boxplot

Figure 4.2: Age and Time Difference Boxplots

4.3.4 Spline graphs for age and loan dispatched predicting default

The Cox spline was also applied to all the covariates in the data set but it only worked with age and the loan dispatched as shown above. From the graphs, it's shown clearly that the age covariate is a sufficient predictor of a loan defaulter as it has a smooth curve. For the amount of loan dispatched graph, it's clear that it's not a sufficient predictor of a loan defaulter.

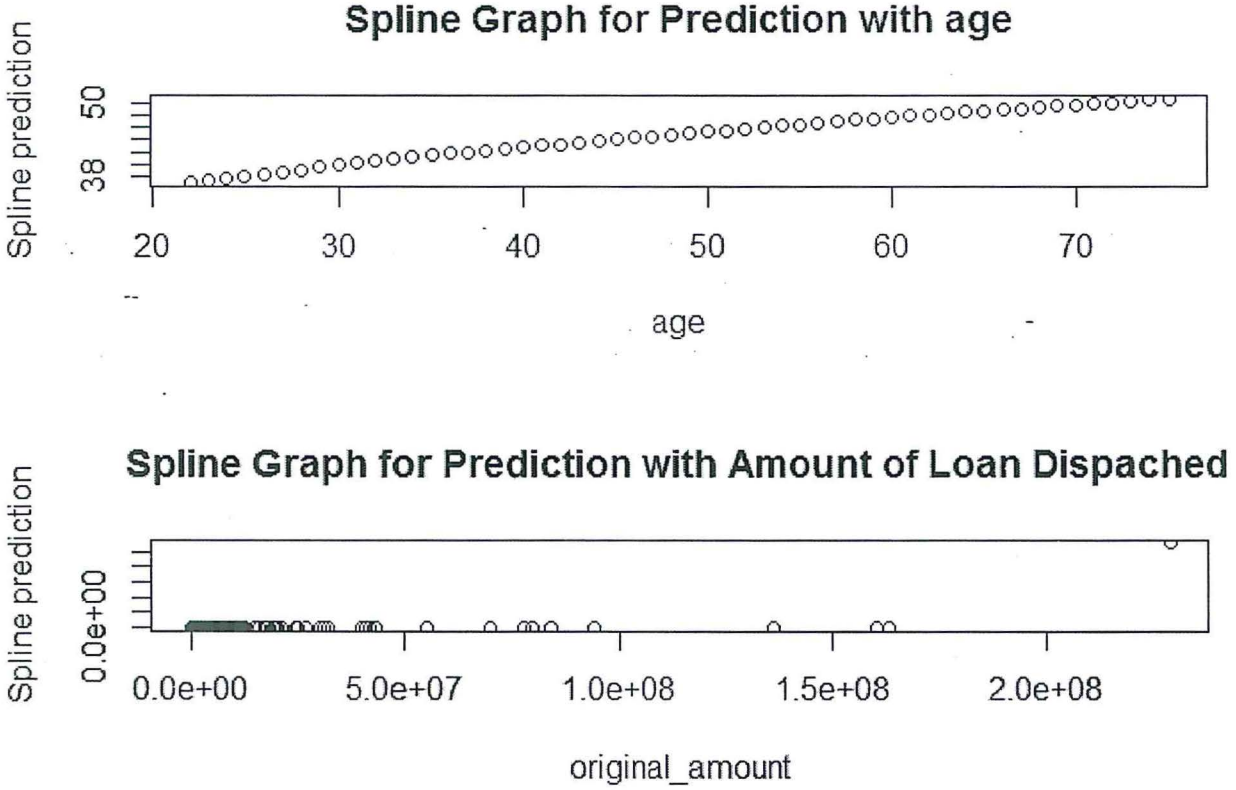


Figure 4.3: Spline graphs for age and loan dispatched predicting default

4.3.5 Survival Analysis Models Estimation

The survival analysis models were analyzed using R statistical software. The Cox Proportional hazard model was significant generally. The model worked well with all the covariates in the estimation of PD. The frailty model as well worked better with all the covariates.

The individual in the age bracket of 18 years and 33 years are more likely to default. This can be attributed to most of them being youths who are still studying and depend on their parent for financial support. Mostly they are unemployed thus lacking financial stability. The males are more likely to default a loan compared to female. This can be attributed to men being breadwinners to their families as well as the extended responsibilities. This can lead to one having many loans and if not financially stable can lead to default. Individuals with credit card account are more likely to default followed by the ones with a current account then the loan account individuals.

Those individuals that are single are also more likely to default a loan. This comes along in that they are not committed with any responsibilities as they have no one to depend on them. Mostly they live a luxurious life which may be unable to maintain thus they will borrow money to spend with no future thought of investments.

Covariates	coef	exp(coef)	se(coef)	z	Pr(>|z|)	exp(coef)	Exp(-coef)	LCI (95%)	UCI (95%)
gender2	-3.190e-01	7.269e-01	3.122e-02	-10.218	< 2e-16	0.7269	1.3757	0.6837	0.7728
age_bracket (34-43)	-2.667e-01	7.659e-	3.622e-02	-7.363	1.8e-13	0.7684	1.3014	0.7157	0.8250
age_bracket (44-53)	-4.569e-01	6.333e-01	3.985e-02	-11.463	< 2e-16	0.6321	1.5821	0.5845	0.6835
age_bracket (>54)	-6.933e-01	4.999e-01	5.879e-02	-11.792	< 2e-16	0.5039	1.9844	0.4491	0.5655
Divorced	-2.756e-01	7.591e-01	2.680e-01	-1.028	0.3038	1.0000	1.0000	1.0000	1.0000
Single	4.053e-02	1.041e+00	3.140e-02	1.291	0.1967	1.3194	0.7579	0.7803	2.2310
Widowed	-1.233e-01	8.840e-01	3.546e-01	-0.348	0.7280	1.3618	0.7343	0.8045	2.3052
Loan Account	-8.604e-01	4.230e-01	3.286e-02	-26.180	< 2e-16	NA	NA	NA	NA
Current Account	-4.777e-01	6.202e-01	4.583e-02	-10.423	< 2e-16	0.4375	2.2858	0.4101	0.4667
Original Amount	-3.098e-08	1.000e+00	1.468e-08	-2.110	0.0349	0.6339	1.5775	0.5793	0.6936

Table 3: Summary of Survival analysis models estimates

Chapter 5

Discussions and Recommendations

5.1 Discussions

In this study, we evaluate the effectiveness of five survival analysis models in credit risk scoring. We used the Akaike Information Criteria (AIC) as the main performance evaluation measure. From the study, it's clearly evidenced that all the models were significant in the analysis of a Kenyan real data set. However, the Cox PH model seemed to have outperformed the other models in comparison though the other models did not perform significantly different in most cases. The mixture cure and non-cure model performed significantly the same however the frailty model performed better.

Comparison between the Cox PH model, frailty model, and mixture and non-mixture models assuming different distributions was assessed using the AIC, where a lower AIC value indicates a better model fit. The study concludes that the Cox PH model is more efficient in the analysis of Kenyan real data set compared to the frailty, penalized spline, and the mixture cure and non-cure model. This was as a result of it having the smallest AIC 0f 39,747, followed by the frailty model which had an AIC of 42,100. The Mixture Non-Cure Model had an AIC of 44,478 and lastly, the Mixture Cure Model emerged as the less efficient survival analysis model with an AIC of 44,503 as shown below;

Model	AIC
Cox PH Model with time independent	39,747
Frailty Model	42,100
Mixture Cure model	44,503
Non- Mixture Cure model	44,478

Table 4: AIC Results for the Survival Analysis Models

5.2 Conclusions and Recommendations

Survival analysis is advantageous in that the time to default can be modeled, and not just whether an applicant will default or not. Furthermore, the models revisited collectively have the advantage of not requiring the survival function to go to zero when time goes to infinity; a situation that is seldom and appropriate for credit risk data

In the study, there was a challenge of finding an appropriate evaluation measure that is evidenced across all the methods for survival analysis comparison. In the future, it could be appropriate to extend the mixture cure and non-cure model and study the performance of these models in contrast with a Cox PH model and some of its extensions. It would be also interesting to run all the models again over data that have been coarse-classified and compare its results with other researchers studies.

References

Allen, L., Delong, G., and Saunders, A. (2004). Issues in the credit risk modeling of retail markets. *Journal of Banking and Finance, 28(4)*, 727-752.

Baesens, J., Crook, J. N., and Thomas, L. C.(1999). Not if but when will Borrowers Default. *The Journal of Operational Research Society*, 50(12), 1185.

Bellotti, T. and Crook J. (2009). Credit scoring with macroeconomic variables using survival analysis. *The Journal of the Operational Research Society* 60(12): 1699–1707.

Bellotti T. and Crook J. (2013). Forecasting and stress testing credit card default using dynamic models. *International Journal of Forecasting* 29(4): 563 – 574.

Bellotti T. and Crook J. (2014). Retail credit stress testing using a discrete hazard model with macroeconomic factors. *Journal of the Operational Research Society* 65(3): 340–350.

Bugnard F., Ducrot C. and Calavas D (1994). Advantages and inconveniences of the Cox model compared with the logistic model: application to a study of risk factors of nursing cow infertility. *Veterinary Research* BioMed Central, 1994, 25 (2-3), pp.134-139.

CBK. (2013a). Credit survey Report. Nairobi: Central Bank of Kenya.

CBK. (2013b). Risk-Based Supervisory Framework. Nairobi: Central Bank of Kenya.

CBK. (2015). The Kenya financial sector stability report. Nairobi: Central Bank of Kenya.

CBK. (2016). Bank Supervision Annual Report. Nairobi: Central Bank of Kenya.

CBK. (2017). Bank Supervision Annual Report 2017. Nairobi: Central Bank of Kenya.

CBK. (2018). Credit Survey Report for the Quarter ended March 2018. Nairobi: Central Bank of Kenya.

CBK. (2018). CBK Annual Banking Report. Nairobi: Central Bank of Kenya.

Dirick, L., Claeskens, G., & Baesens, B. (2015). An Akaike information criterion for multiple

event mixture cure models. *European Journal of Operational Research*, 241:449–457.

Gitahi J. N., & Othieno F. (2017). Survival Analysis Approach To Credit Risk Modeling. *Unpublished Bachelors' Research Proposal* . Strathmore University.

Gupta, V. (2017). A survival approach to the prediction of default drivers for India listed companies. *Theoretical Economics Letters*, 07(02), 116-138.

Gaynor M. and Town R. (2011). Competition in Health Care Markets. *Chapter for the Handbook of Health Economics*, Volume 2. T. McGuire, M.V. Pauly, and P. Pita Barros, Editors 2011.

Jacobson, T., & Roszback, K. (2003). Bank lending policy, credit scoring, and value at risk. *Journal of Banking and Finance*,27(4), 615-633.

J-K Im, DW Apley, C Qi and X Shan (2012). A time-dependent proportional hazards survival model for credit risk analysis. *Journal of the Operational Research Society*, 63(3) : 306-321.

Jose A. D., Edna S. L., & Jaime O. (2013). Interest rates and default in unsecured loan markets. *Quantitative Finance*, 13:12, 1925-1934,

Kagri, H. S. (2011). Credit risk and performance of Nigerian banks. *Ahmadu Bello University, Zaria.*

Lore, D., Gerda, C., and Bart, B.(2016). Time to default in credit scoring using survival analysis. *Journal of the Operational Research Society*, 68, 652–665

Marimo, M. (2015). Survival analysis of bank loans and credit risk prognosis. *Unpublished master's thesis.* The University of Witwatersrand.

Merton, R. (1974). On the pricing of corporate debt: The risk structure of interest rates. *Journal of Economics and International Finance.*

Narain B. (1992). Survival analysis and the credit granting decision. In: Thomas LC, Crook JN and Edelman DB, editors, *Credit Scoring and Credit Control*, pp. 109–121. Clarendon Press: Oxford.

Nevo A. (2001). Measuring Market Power in the Ready-to-Eat Cereal Industry. *Econometrica*

2001;69(2); 307-342.

Obuda, F. (2016). Analysis of credit risk on bank loans using the Cox proportional hazard model. (Unpublished master's thesis). The University of Nairobi.

Omoga, A. A. (2017). Predictive modeling in credit risk: a survival analysis case. (Thesis). Strathmore University.

Pauline N., Lucy M. and Collins O. (2018). Modeling Factors Affecting Probability of Loan Default: A Quantitative Analysis of the Kenyan Students' Loan. *International Journal of Statistical Distributions and Applications.* Vol. 4, No. 1, 2018, pp. 29-37. doi: 10.11648/j.ijsd.20180401.14

Tong, E., Mues, C., & Thomas, L. (2012). Mixture cure models in credit scoring. *European journal of operational research*, 218(1), 132-139.

Stepanova M. and Thomas L. (2002). Survival analysis methods for personal loan data. *Operations Research Quarterly* 50(2): 277–289.

Sy, J., & Taylor, J. (2000). Estimation in a Cox proportional hazards cure model. *In Biometrics* (pp. 56(1):227–236.)

Valle, C. A. (2013). Credit Risk Modeling in a Semi-Markov Process Environment. (Unpublished doctoral dissertation). University of Manchester.

Wekesa, O. A., Samuel, M., & Peter, M. (2012). Modeling Credit Risk for Personal Loans Using Product-Limit Estimator.*International Journal of Financial Research*, 3(1).

Goel, M. K., Khanna, P., Kishore, J. (2010). Understanding survival analysis: Kaplan-Meier estimate. *International journal of Ayurveda research*, 1(4), 274.

Zhang, A. (2009). Statistical methods in credit risk modeling. (Unpublished doctoral dissertation). The University of Michigan.

Appendix

The R Code

Packages required

```
library(survival)
library(smcure)
library(arm)
library(dplyr)
library(ranger)
library(ggplot2)
library(survminer)
library(rcure)
library(flexsurvcure)
```

$\#\#\#\#\#\#\#\#\#Fitting Kaplan Meier curves\#\#\#\#\#\#\#$

```
head(LoanDefalt_Sammy)
```

fit1 <- survfit(Surv(time, Status) ~ *marital_status*, *data=LoanDefalt_Sammy*)

ggsurvplot(*fit*1, *data=LoanDefalt_Sammy*)

summary(*fit*1)

*fit*2¡−*survfit*(*Surv*(*time*, *Status*) ~ *gender*, *data=LoanDefalt_Sammy*)

ggsurvplot(*fit*2, *data=LoanDefalt_Sammy*)

summary(*fit*2)

*fit*3¡−*survfit*(*Surv*(*time*, *Status*) ~ *age*, *data=LoanDefalt_Sammy*)

ggsurvplot(*fit*3, *data=LoanDefalt_Sammy*)

summary(*fit*3)

*fit*4¡−*survfit*(*Surv*(*time*, *Status*) ~ *product_name*, *data=LoanDefalt_Sammy*)

ggsurvplot(*fit*4, *data=LoanDefalt_Sammy*)

summary(*fit*4)

$\#\#\#\#\#\#\#\#\#\#Cox PH Model with time independent\#\#\#\#\#\#\#\#$

fit_CPH <- coxph(Surv(time, Status) ~ *age_bracket+marital_status+product_name+*

$original\_amount, data{=}LoanDefalt\_Sammy)$

$summary(fit\_CPH)$

$\#\#\#\#\#\#\#\#\#\#Frailty\#\#\#\#\#\#$

rfit <- survreg(Surv(time, Status) $\sim age\_bracket + marital\_status + product\_name +$
$original\_amount$
$+ frailty.gaussian(age, df{=}13, sparse{=}TRUE), data{=}LoanDefalt\_Sammy)$
$summary(rfit)$

$\#\#\#\#\#\#\#\#\#\#spline\#\#\#\#\#\#$

par(mfrow = c(1,2))
sfit <- survreg(Surv(time, Status)$\sim pspline(age, df{=}2), data{=}LoanDefalt\_Sammy)$
$plot(LoanDefalt\_Sammy\&age, predict(sfit), xlab{=}'age', ylab{=}"Splineprediction",$
$main{=}"SplineGraphforPredictionwithage")$
$sfit1¡-survreg(Surv(time, Status) \sim pspline(original\_amount, df{=}2), data{=}LoanDefalt\_Sammy)$
$plot(LoanDefalt\_Sammy\&original\_amount, predict(sfit1), xlab{=}'original\_amount',$
$ylab{=}"Splineprediction", main{=}"SplineGraphforPredictionwithoriginalamount")$

$\#\#\#\#\#\#\#\#\#Curemixturemodel\#\#\#\#\#\#\#$

cure_model <- flexsurvcure(Surv(time, Status)$\sim age, data{=}LoanDefalt\_Sammy,$
$link{=}"logistic", dist{=}"weibullPH", mixture{=}T)$
$print(cure\_model)$
$cure\_model2¡-flexsurvcure(Surv(time, Status) \sim maritalstatus, data{=}LoanDefalt\_Sammy,$
$link{=}"logistic", dist{=}"weibullPH", mixture{=}T)$
$print(cure\_model2)$
$cure\_model3¡-flexsurvcure(Surv(time, Status) \sim age\_bracket, data{=}LoanDefalt\_Sammy,$
$link{=}"logistic", dist{=}"weibullPH", mixture{=}T)$

$print(cure\_model3)$

$cure\_model4 <- flexsurvcure(Surv(time, Status) \sim gender, data=LoanDefalt\_Sammy,$
$link="logistic", dist="weibullPH", mixture=T)$

$print(cure\_model4)$

$\#\#\#\#\#\#\#\#\#\#\#\#cure mixture model(uncured individuals)\#\#\#\#\#\#\#\#\#$

cure_model<- fltext exsurvcure(Surv(time,Status) age, data=LoanDefalt_Sammy,
link="logistic", dist="weibullPH",mixture=T, anc=list(scale=$\sim age$))

$print(cure\_model)$

$\#\#\#\#\#\#\#\#non - cure mixture model\#\#\#\#\#\#\#\#\#$

cure_model_nmix <- flexsurvcure(Surv(time, Status)$\sim age, data=LoanDefalt\_Sammy,$
$link="logistic", dist="weibullPH", mixture=F)$

$print(cure\_model3)$

$cure\_model\_nmix1 <- flexsurvcure(Surv(time, Status) \sim age\_bracket, data=LoanDefalt\_Sammy,$
$link="logistic", dist="weibullPH", mixture=T)$

$print(cure\_model3)$