2017

Predictive modelling in credit risk: a survival analysis case

Allan Anyona Omoga
Strathmore Institute of mathematical Sciences (SIMs)
Strathmore University

Predictive Modelling in Credit Risk
A Survival Analysis Case.

Omoga, Allan Anyona
093452

Submitted in partial fulfilment of the requirements for the degree of
Master of Science in Statistical Science at Strathmore University.

Institute of Mathematical Sciences
Strathmore University
Nairobi, Kenya

June, 2017

2017

Predictive modelling in credit risk: a survival analysis case

Allan Anyona Omoga
Strathmore Institute of mathematical Sciences (SIMs)
Strathmore University

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 thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Omoga, Allan Anyona

…….....
10th June 2017.

Approval

The thesis of Omoga, Allan Anyona was reviewed and approved by the following:

Prof. Samuel Mwalili.
Lecturer - Institute of Mathematical Sciences.
Strathmore University.

Ferdinard Otieno
Dean – Institute of Mathematical Sciences
Strathmore University

Prof. Ruth Kiraka
Dean, School of Graduate Studies.
Strathmore Univerity.

Abstract

Six survival analysis techniques are accessed by applying the techniques to a dataset consisting of 33,238 active credit facilities from a financial institution operating in Kenya. Namely, the Accelerated Failure Time (AFT) Models, Cox proportional hazard (PH) Model and the Mixture Cure Model (MCM) are considered in the comparisons. Evaluation of the techniques is conducted from a Statistical approach evaluation using the Area under the Curve (AUC) and financial evaluation using the annuity theory. The Cox Proportional Hazard (PH) and the Mixture cure model performs significantly well.

Keywords: Credit Event, Mixture Cure model, Survival Analysis, Credit risk modelling.

Table of Contents

List of Figures

List of Tables

Acknowledgements

Thanks to the Almighty for the gift of life and energy. Special thanks to my Supervisor Prof. Samuel Mwalili, for his tremendous guidance and support. To my Classmates for the overwhelming support through this thesis. May God, bless you all.

Dedication

I dedicate this thesis to God for seeing me through my academic studies and for the gift of life, my parents who always believed in me, my colleagues for the support during my studies and to my girlfriend, thanks for being patient with me throughout my studies.

Chapter 1 Introduction

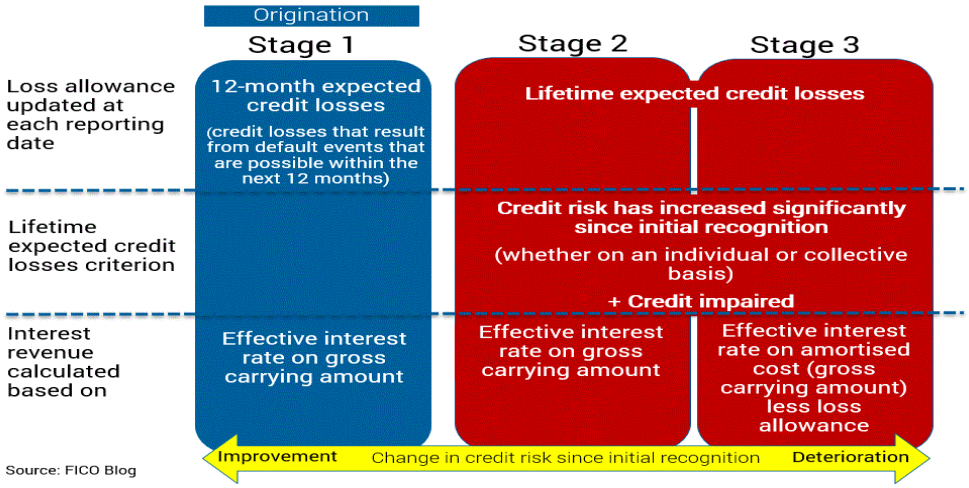
1.0 Introduction

Assessment of credit event is a critical element to any financial institutions, the ability of a lending entity to clearly identify and segment customers based on the credit risk profile still remains a challenge for institutions operating in Kenya. This was evident by 47.5% rise in non-Performing facilities reported in the 2016 financial results by financial institutions operating in Kenya (Central Bank of Kenya, 2016). This coupled up with the introduction of International Financial Reporting Standards 9 (IFRS9) which is set to replace the current International Accounting Standards (IAS 39) from January 2018. The standard defines the methodology upon which Impairment calculations should be calculated. IFRS9 requires for impairment losses and impairment gains, the calculation to be based on a lifetime expected credit loss, rather than the current standard, IAS 39, which is anchored on an incurred loss principle.

IFRS9 defines three key stages;

- Stage 1 where a facility is performing (0 days past due).
- Stage 2 where a facility is delinquent (1 to 90 days past due).
- Stage 3 where a facility is in default (>91 days past due).

Figure 1 Illustration of IFRS9 Key Credit Stages (Source: FICO Blog)



This is expected to cause seismic changes in the way facilities are originated since higher provision will be taken once an account moves to Stage 2 and Stage 3,

where impairment would be taken at the lifetime level. The estimation of Time to default and probability of default breathes in a new wave for credit risk modeling via survival analysis, where Probability of Default (PD) can be estimated from a lifetime expected loss approach, by allowing extensions of effects of covariates on the predicted time to a credit event to fluctuate as the facility evolves.

Survival techniques are utilised extensively in the medical field as a technique for modelling time to an event of interest (Hosmer et al., 2008). Narain (1992) first introduced the technique in Credit risk context, this was later developed and improved by different authors over the years; Banasik et al. (1999), Hand and Kelly (2001) and Stepanova and Thomas (2002). Survival analysis provides a framework where one can not only predict whether a facility will default but also when the facility is likely to experience a credit event.

1.1 Background of Study

We access various survival modelling techniques in the consumer credit risk environment to deduce on the model that best provides the Probability of Default (*PD*) estimation, specifically the Accelerated Failure Time (AFT) models, Cox proportional hazards (PH) models and the Mixture Cure models for a single event and Mixture cure model for multiple events. Evaluation of the models is done using financial (Annuity Theory) and statistical evaluation (AUC).

1.2 Problem Definition

Financial Institutions utilise the Internal Risk Based models in accessing the likelihood of Insolvency risk as a result of a credit event. Currently, Regression Models are largely applied across Financial Institutions LR) (Stepanova and Thomas, 2002). However, with the Introduction of the International Financial Reporting Standard 9 (IFRS9), set to start from January 2018, which requires, for impairment losses and impairment gains, the calculation will be based on a lifetime expected credit loss, rather than the current standard based on, the International Accounting Standard (IAS39), which is anchored on an incurred loss principle. Previous researchers have suggested that survival model might improve the Probability of Default *(PD)* estimation (Thomas, Crook, & J, 1999). This can provide a framework where Probability of Default *(PD)* can be estimated from a lifetime expected loss

approach, by allowing extension of effects of covariates on the predicted time to a credit event to fluctuate as the facility evolves. Thus, providing an opportunity for Risk Professionals in managing their Impairment numbers.

Main Objective

To access the performance of various survival analysis models, and their application in the credit risk environment, and to conclude which model provides the best PD estimation.

Specific Objective

i. To formulate various survival predictive models for credit risk scoring.
ii. To fit the proposed predictive models to credit risk scoring.
iii. To determine the best predictive survival analysis model.

Significance of Study

The Survival model developed is expected to improve;

1. ***Profit Scoring:*** The Time to a Credit Event will provide an informative view of the profitability of a facility during onboarding, providing the first step to facility Profit Scoring.
2. ***Impairment Provisioning:*** The Probability of Default (PD) Estimates will provide a better forecast of the 12 months expected credit losses for the Performing book and lifetime expected credit loss for the Delinquent book hence opportunity on improving on PD and LGD rates.
3. ***Credit Lending Policy:*** The Time to Default estimates may guide credit lending policy over the Consumer Credit Facilities Tenure.

Chapter 2 Literature Review

Previous researchers have employed survival analysis techniques in determining the probability of default, *PD*. This is seen Cao, Juan, & Andr´es (2009); Belloti & Crook (2008); Madorno, Mecatti, & Figini (2013); Andreeva (2006); and Stepnova & Thomas (2002).

Cao, Vilar, and Andr´es (2009) found sufficient evidence supporting the effectiveness of survival analysis techniques by utilising Cox Proportional hazards (PH) model and the nonparametric conditional distribution estimation. Their results from these methods find *"powerful discrimination between default and non-default credits"* Cao, Vilar, and Devia (2009), which shows the strength of these models in determining the difference between credit event that will or will not occur.

Banasik, Crook, and Thomas (1999) performed survival analysis on Consumer loan data from a leading United Kingdom (UK) financial institution. Their sample composed 50,000 loans spanning the period from June 1994 to March 1997. Andreeva's (2006) data consists of a retail card issue in Belgium, the Netherlands, and Germany over a 25-month period from October 1998 to December 2000. Andreeva's implementation of survival analysis thus moves beyond the fixed-term credit studied in previous papers and into the arena of revolving credit. Andreeva uses various models, including accelerated failure time (AFT), proportional hazards, and the more traditional logistic regression. Bellotti and Crook (2008) conducted survival analysis on data from over 200,000 credit card accounts opened in the United Kingdom (UK) from 1997 to 2005. Tong, Mues, and Thomas (2012) utilise a dataset of 27,527 observations of Retail facilities from a single consumer bank in the United Kingdom (UK) covering loan terms of 12, 24, and 36 months.

Several studies have shown findings that support the strength of survival analysis methods. Tong, Mues, and Thomas (2012) associated application of survival models in biostatistics to utilising the same technique in econometrics. By predicting the probability of default (PD) on a United Kingdom (UK) Based Financial Institution loan portfolio, using 3 methods: A mixture cure model, a Cox proportional hazards method and a standard logistic regression. The survival approaches were shown to provide more robust probability estimates in comparison to the standard logistic regression. Banasik, Crook, and Thomas (1999) examine the probability of when a borrower will default as opposed to just the probability of if the

borrower will default. They find that the ability of proportional hazard models to predict a credit event in the first 12 months rivals that of logistic regression models. Furthermore, their findings suggest that proportional hazard models are better than logistic regression models in predicting if a borrower will pay off their facility before maturity date within the first year. Bellotti and Crook (2007) found that survival analysis provides more predictive power than logistic regression due to its ability to incorporate time-variant macro-financial variables.

Madorno, Mecatti, and Figini (2013) use the Cox regression model *"in order to write the default probability in terms of the conditional distribution function of the time to default."* The findings from their analysis conclude that using survival analysis provides more robust results. Andreeva (2006) examines different *"timescales of default"* and finds strong support for the notion that the predictive power of survival analysis is comparable to that of logistic regression.

From current literature, questions remain to be explored. Firstly, apart from Zhang and Thomas (2012), little attempt to compare a several survival analysis techniques in one paper has been made, and, the majority of the research, the evaluation techniques remains anchored mainly on the area under the curve (AUC) of the receiver operating characteristics(ROC) curve and classification.

Survival models are set to play a pivotal role in credit risk following the introduction of the International Financial Reporting Standards 9 (IFRS9) which establishes a new approach for impairing loans and advances. In the new approach, Impairment is to be calculated from an expected loss of twelve-month for the performing book, and Lifetime expected loss for the Delinquent and Default book. Estimating the Time to default breathes in a new wave for Impairment modelling via Survival Models.

This research contributes to existing literature by analysing a dataset from a Leading Financial Institution in Kenya, by utilising the survival techniques itemized in Table 1, and employing existing statistical evaluation measures (Area under the Curve and the error measurement) and financial evaluation measures (eventual true future value of the facility), applicable to the various survival model employed.

Table 1: Rundown of current literature on survival analysis in credit risk setting

Paper	AFT	Cox PH	Mixture cure	Multi-event Mixture Cure	Sample Size	Number of Inputs	Evaluation Measure
Narain(1992)	Y				1,242	7	***None.***
Banasik t al. (1999)	Y	Y			50,000	>7	***Classification.***
Stepnova and Thomas (2001)		Y			11,500	16	***Classification, AUC, Profit Measure.***
Stepnova and Thomas (2002)		Y			50,000	16	***Classification,AUC.***
Cao et al. (2009)	Y	Y			25,000	1	***AUC***
Tong et al. (2012)		Y	Y		27,527	14	***AUC, H-Measure, Kolmogrov-Smirnov.***
Zhang and Thomas (2012)	Y	Y			27,000	21	***Error in default time prediction.***
Dirick et al (2015)			Y	Y	7,521	8	***AUC***

**The number of inputs is before final variable selection.*

Chapter 3 Research Design

3.1 Survival Analysis

The conditional survival function for modelling credit risk was measured via the conditional distribution of the random variable; time to default or early payoff, and maturity period for the mixture cure model (MCM). *T*, given a vector of measurements *x*.

Two competing risk exists, time to default and early payoff, time to default, T_1, was estimated with the assumption of censoring for the other observed performance; and, independently, time to early payoff , T_2, assuming all other observation to be censored. Survival analysis was then trained separately on, T_1, and T_2. The forecasted lifetime of the facility was then estimated as, $T = \min\{T_1, T_2, faclity\ term\}$ (Thomas, Crook, & J, 1999)

The probability density function $f(u)$ was then expressed as

$$f(u) = \frac{d}{du} S(u).$$

The distribution of T was then expressed by the hazard function;

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T < t + \delta t \,|T \geq t)}{\delta t} \right\}.$$

This model the instantaneous risk. Expressing the function in regards to the cumulative distribution function and probability density function;

$$h(t) = \frac{f(t)}{1 - F(t)} \ \ where\ F(t) = P(T < t).$$

A proportion of the observations in survival data is normally censored, meaning that for these observations, an event of interest has not yet been recorded as at the point of data aggregation. The following are the definitions for censoring;

- ***Censored cases:*** Those credit facilities which had not attain the predefined maturity date at the point of data aggregation, and did not experience default or early payoff.
- ***Uncensored cases:*** The credit facility where default has been observed by the end of the observation period. Hence, mature cases and early payoffs cases are censored, along with the censored cases according to the first definition.

When modelling the time to default, T, the second definition (Uncensored Cases) is applied. However, for the case of multiple event mixture cure models, where

competing risks were factored into account, the first definition applied. The censoring indicator for the $i-th$ case which is denoted by δ_i, is equal to 1 for an uncensored observation and equals to zero for censored observations.

3.1.1 Accelerated Failure Time (AFT)

These are parametric survival models where the explanatory variables are acceleration factors that accelerate or decelerate the survival process when contrasted with to the baseline survival function. This can be expressed by;

$$S(t|x) = S_0\{t.\exp(-\beta' x)\}.$$

The event rate decelerates when $0 < \exp(-\beta' x) < 1$, and accelerates when $\exp(-\beta' x) > 1$.

The hazard function expressed as; $h(t|x) = h_0\{\text{t}.\exp(-\beta' x)\}\exp(-\beta' x)$.

The general form, the AFT model represented as a log-linear model for the time to occurrence of an event of interest. $\log(T) = \beta^i x + \sigma\varepsilon$, where ε is a random error, and σ a parameter that rescales ε. Since most classical survival distributions, tend to have event times that are log-linear, AFT models tend to be the model of choice as a starting point in order to parameterize these distributions, Collett (2003) and Kleinbaum and Klein (2011).

3.1.2 Weibull Accelerated Failure Time Model

These are expressed by the survival and hazard function. The Weibull Model has a scale λ and shape p which is expressed in the classical form.

$$S(t) = \exp(-\lambda t^p), h(t) = \lambda p t^{p-1}.$$

Using the relationship $\sigma = \frac{1}{p}$, the random event time for the Weibull-distribution $T_i = \exp(\beta' x_i + \sigma\varepsilon_i)$ resembles to $S(t|x) = \exp\left(-\lambda t^{\frac{1}{\sigma}}\right)$, where $\lambda_{i,} = \exp(\frac{\beta' x_i}{\sigma})$ is the reparameterization that incorporates the explanatory variables.

3.1.3 Exponential Accelerated Failure Time Model

The Model is a case of the Weibull distribution, where $p = 1$. Leading to a survival function;

$$S(t) = \exp(-\lambda t),$$

And the corresponding Hazard Function.

$$h(t) = \lambda.$$

This distribution employs the strong assumption of constant hazard rate λ, and for each case $\lambda_i = \exp(-\beta' x)$.

3.1.4 Log-Logistic Accelerated Failure Time Model

The log-logistic distribution is expressed by parameters $\theta \; and \; k$. The corresponding survival function is given as;

$$S(t) = \frac{1}{1 + \exp(\theta)\, t^k}$$

Hazard function is then expressed as;

$$h(t) = \frac{\exp(\theta)\, k t^{k-1}}{1 + \exp(\theta)\, k t^k}.$$

Using the AFT re-parameterization, the strong assumption of a constant hazard rate $\sigma = \frac{1}{k}$ and the log-logistically distributed event time T_i has a survival function defined as;

$$S_i(t|x_i) = \frac{1}{1 + \exp(\theta_i)\, t^{\frac{1}{\sigma}}}, where \; \theta_i = \frac{\beta' x_i}{\sigma}.$$

3.1.5 Cox Proportion Hazard Model

Cox (1972) proposed the proportional hazard model, with the assumption that the hazard of a credit facility characteristic x is proportional to an unknown baseline hazard. This model is flexible than the AFT models since it has a non-parametric baseline hazard function,$h_0(t)$, along with a parametric part. The Cox proportional hazard model follows;

$$h(t|x) = exp^{(\beta' x)}\, h_0(t), \qquad (1)$$

and the survival function;

$$S(t|x) = \exp(-\exp(\beta' x)) \int_{o}^{t} h_0(u)du) = \exp(-\exp(\beta' x))\, H_0(t)),$$

with $H_0(t)$, the cumulative baseline hazard function. This equation becomes challenging to calculate, hence Breslow (1974) and Efron (1977) proposed and easier approximation method. The study used Breslow's method to estimate the cumulative baseline hazard rate, which is expressed as;

$\widehat{\mathrm{H}}_0(t) = \sum_{t_i \leq t} \frac{1}{\sum_{r \varepsilon R(t_i)} exp(x'_r \beta')}$,

Where $R(t_i)$ denotes the group of individuals at risk at time t_i (the ones that have not had a credit event by time t_i).

3.1.6 Mixture cure model

These models are inspired by the presence of a segment of long-haul survivors, or a "*cured*" segment (Sy & Taylor, 2000). The subgroup incorporates into the model through a mixture distribution where a logistic regression model provides a mixing proportion of the "*non-susceptible*" cases (Dirick, Claeskens, & Baesens, 2015). A survival model portrays the cases inclined to the event of interest. This model is of particular interest in credit risk modelling since the key event of interest, default, will not occur for a very large fraction of the subjects. Tong et al. (2012) introduced the idea in the credit risk context first. Dirick et al. (2015) introduced and applied a model selection criterion adapted to these models to credit reapplication data. The unconditional survival function for the mixture cure model is given by:

$$S(t|x) = \pi(x) S\,(t|Y = 1, x) + 1 - \pi(x). \qquad (2)$$

With Y denoting the susceptibility indicator, ($Y = 1$ indicating susceptibility of a given facility, and $Y = 0$ indicating non-susceptibility of a given facility,). A new covariate vector x for the logistic regression model is introduced, in this case the binomial logit.

$$\pi(x) = P(Y = 1|x) = \frac{\exp(b'x)}{1 + \exp(b'x)}.$$

Farewell (1982) & Tong et al. (2012) illustrated the difference between an unconditional survival curve and plain survival curves in a mixture cure model via Figure 2. The full lines denote the plain survival curves (modelled using an Accelerated Failure Time - Weibull model for the grey curve, and a log-logistic AFT

model for the black curve), the specked lines speak to the unconditional survival curves in a mixture cure model with a cure rate of 30% with corresponding parameter vector *b.*

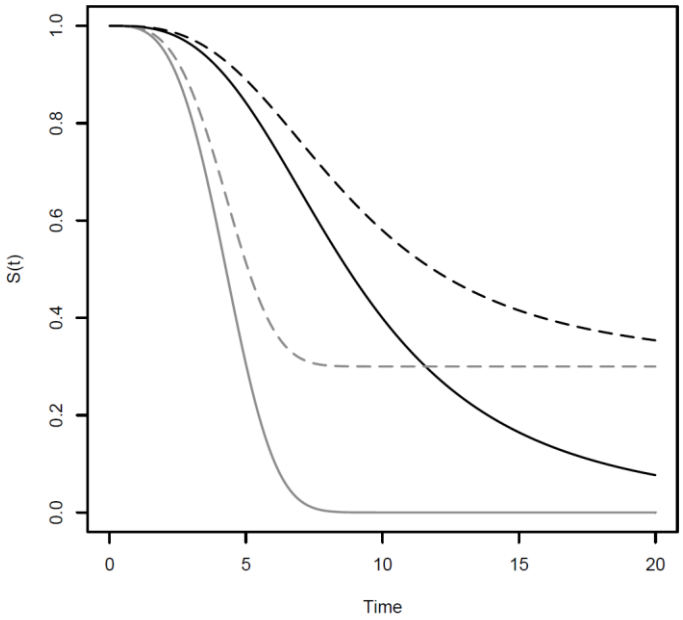


Figure 2 Illustration of the comparison of two plain survival curves and the unconditional survival curve in a mixture cure model.

The conditional survival function modelling the facilities that are inclined to default is expressed by a Cox proportional hazards model,

$$S(t|Y=1,x) = \exp(-\exp(\beta' x) \int_{o}^{t} h_0(u|Y=1)du).$$

3.1.7 Mixture cure model - Multiple events

It's bizarre to ever really record cure in the medical context. However, an observed cure exists in the credit risk setting, where a facility moves from charge off/default status to a performing status since as a facility attains maturity and there are no residual balances, default cannot occur. The censoring indicator in the mixture cure model gives information on if or not default took place, information on maturity is not applied in the model.

An approach that offers simultaneous modelling of multiple events, alongside a mature group was proposed by Watkins et al. (2014). This model was later

extended by Dirick et al. (2015), where a semi-parametric Cox proportional hazards models the survival times, as an alternative to the parametric survival models which was applied by previous researchers.

- Y_m, denoting the facility to be matured, hence repaid at the maturity date.
- Y_d , denoting occurrence of the event of interest, default.
- Y_e , denoting that early payoff/repayment takes place.

The set (Y_m, Y_d, Y_e) is exhaustive and mutually exclusive. However, for censored, it is unknown which event will transpire. In analogy, to

$$S(t|x) = \pi(x)S\,(t|Y = 1, x) + 1 - \pi(x),$$

the unconditional survival function denoted as;

$S(t|x) = \pi_e(x)S_e\,(t|Y_e = 1, x) + \pi_d(x)S_d\,(t|Y_d = 1, x) + \big(1 - \pi_e(x) - \pi_d(x)\big)$.

With $S_e(t|Y_e = 1, x)$ and $S_d(t|Y_d = 1, x)$ signifying the conditional survival functions for, early payoff and default,respectively , this modeled using a Cox proportional hazards model, as in equation (2).

3.2 Model Parameters Estimation

3.2.1 The Proportional Hazard Model

The information about β can be obtained from the relative orderings of the survival times. Let A_i be the incident that a facility *i* will experience default in $[u, u + \Delta u)$ condition to the facility being open with a debit balance at u. Let $t_1, .., t_u$ define the individual default times, then

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

Where $Y_{i(u)}(u) = 1$ when the facility is at risk at u. The partial Likelihood can then be expressed as;

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

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;

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

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

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

This maximum $\hat{\beta}$ defined the Maximum Partial Likelihood estimate (MPLE) of β.

3.2.2 The Mixture Cure Model - Single Event

The incidence model component utilises a logistic regression. The latency model uses, 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;

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

Given the full information of Y. The complete likelihood is then expressed as;

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

3.2.3 The Mixture Cure Model - Multiple Event

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

$$L_{obs}(\boldsymbol{\Theta}) = \prod_{i=1}^{n} \{\prod_{j=1}^{2} \pi_j(x_i; \boldsymbol{b}_j) f_j(t_i | Y_{j,i} = 1, x_{j,i}; \beta_j)^{Y_{j,i}} \left(1 - \sum_{j=1}^{2} \pi_j(x_i; \boldsymbol{b}_j)\right)^{Y_{m,i}}\}^{\delta_i} * \{\left(1 - \sum_{j=1}^{2} \pi_j(x_i; \boldsymbol{b}_j)\right) + \sum_{j=1}^{2} \pi_j(x_i; b_j) S_j\left(t_{j,i} \middle| Y_{j,i} = 1, x_{j,i}; x_{\beta_j}\right)\}^{1-\delta_i}$$

Where $\boldsymbol{\Theta} = (\boldsymbol{b_e}, \boldsymbol{b_d}, \boldsymbol{\beta_e}, \boldsymbol{\beta_d})$**.** Zeng and Lin 2007 discussed that the Maximum of the observed likelihood does not exist. Therefore, the maximization of the Kernel-smoothed profile likelihood as proposed by Zeng and Lin 2007 using an EM Algorithm is employed.

The flexibility of the model; different segments employed its different set of covariates, therefore the vectors $\boldsymbol{x}_d, \boldsymbol{x}_e$ and $\boldsymbol{x}$ may vary. The model can then be rewritten starting from the complete likelihood, hence the likelihood expression under the assumption that the full information on $Y = (Y_m, Y_d, Y_e)$ is present;

$$L_{complete}(\boldsymbol{\Theta}) = \prod_{i=1}^{n} \{\prod_{j=1}^{2} (\pi_j(x_i; \boldsymbol{b}_j))^{Y_{j,i}} (1 - \prod_{j=1}^{2} (\pi_j(x_i; \boldsymbol{b}_j))^{Y_{m,i}}\} * \{\prod_{j=1}^{2} h_j(t | Y_{j,i} = 1, x_{j,i}; \beta_j)^{\delta_i} \; S_d(t_{j,i} | Y_{j,i} = 1, x_{j,i}; \beta_j)^{Y_{j,i}}\}$$

Using the model density with parameters $\boldsymbol{\Theta}_1$ the expected value can be computed by converting the Likelihood to a log likelihood translating to the $Q - function$

$$\mathrm{Q}(\boldsymbol{\Theta}_1 | \boldsymbol{\Theta}_2) = E_f[\log L_{complete}(\boldsymbol{\Theta}_2; \mathrm{T}_i, \delta_i, \boldsymbol{\Theta}_1]$$

$$= \sum_{i=1}^{n} \left\{ w_{ji} \log\left(\pi_j(x_i; b_j)\right) + w_{mi} \log(1 - \sum_{j=1}^{2} \pi_j(x_i; b_j)) + \sum_{j=1}^{2} w_{ji}\, \delta_i \log\left(h_j(t_i | Y_j = 1, x_{j,i}; \beta_j)\right) + w_{ji} \log\left(h_j(t_i | Y_j = 1, x_{j,i}; \beta_j)\right)\right\}$$

The conditional expectations of $Y_{i,j}(j = 1,2)$, $E_f\left[Y_{i,j}|T_i|\delta_i, \boldsymbol{\Theta}_1\right]$, are calculated with respect to the model density using parameter $\boldsymbol{\Theta}_1$ denoted by w_{ji} with $w_{mi} = 1 - w_{1i} - w_{2i}$ and for $j = 1,2$,

$$\boldsymbol{w_{mi} = w_{mi}(\Theta) = P\left(Y_{i,j} = 1 \middle| T_i = t; \delta_i; \Theta\right)}$$

$$= \begin{cases} \dfrac{\pi_j(x_i; b_j) S_j(t_i; \beta_j)}{\sum_{k=1}^{2} \pi_k(x_i; b_k)\, S_k(t_i; \beta_k) + (1 - \sum_{k=1}^{2} \pi_k(x_i; b_k))} & for\ \delta_i = 0 \\ 1 & for\ Y_{i,j} = 1\ and\ \delta_i = 1 \\ 0 & for\ Y_{i,j} = 0\ and\ \delta_i = 1 \end{cases}$$

3.3 Population and sampling

Data was obtained from a foremost financial institution in Kenya, consisting of mainly Personal Instalment Loans. For the period 01 January 2010 to 31 December 2016. The raw dataset composed of 33,238 active accounts. Accounts entering the study after the observation period, are left truncated. Granular details provided per facility for each observation month.

3.4 Data Analysis

3.4.1 Missing Data

As some survival analysis methods can't cope with missing data, measures were put to ensure the final data-set was free of missing data, by putting a where condition limiting only to non-null observations in the Structured Query Language (SQL) statements during data extraction.

3.4.2 Experimental Setup

The SAS software 9.3 was used. The dataset was randomly split using the *PROC SURVEYSELECT* function into a training set of (67%) and (33%) of the observation respectively. Survival models were induced to the training sets, and the corresponding test datasets used for evaluation.

The *PROC PHREG* and *BASELINE* function are used in observing the effect of the continuous variables separately in the model. A smoothing method is then used to provide an estimate of the effect of the set of covariates on the survival rate.

The *PROC PHREG* is used to fit the Cox-Proportional Hazard model and the *PROC LIFEREG* to fit the AFT Models. The *PSPMCM SAS Macro* developed by Fabien & Pierre, 2007 was used to fit the Mixture Cure Models. The maximisation of the likelihood function in the macro is performed using SAS PROC NLMIXED for parametric models and through an EM algorithm for the Cox PH mixture cure model (Fabien & Pierre, 2007). The variance of parameters estimates are obtained by inverting the Hessian matrix or by non-parametric bootstrap methods.

3.5 Performance evaluation metrics

The following evaluation measures were employed.

3.5.1 Area under the Curve (AUC)

The receiver operating curve is the universal method of evaluating binary classifiers. The curve illustrates each double classifier performance, for every conceivable limit value, by plotting the genuine positive rate against the false positive rate, (Dirick, Claeskens, & Baesens, 2015). The particular performance metrics of intrigue is the region under the curve, which can likewise be generated from a survival analysis setting. In this setting, assessment is conceivable at any time-point of the survival curve (Heagerty & Saha, 2000).

3.5.2 Default times prediction

Prediction of default through time concentrates on the capacity to foresee the default times of the credit events in the observation data. A survival curve gives the time estimates distribution. However, with a high censoring level, the mean estimate of the survival analysis does not provide good estimates. Zhang and Thomas (2012) derived a forecaster for the recovery rate by taking a gander at every percentile of the training set and computing the squared and absolute deviations from the predictors to the actual figures of the default cases. Then, the percentiles with the lowest deviations were withheld and used to calculate the deviations in the observation set. A similar approach to Zhang and Thomas (2012) was applied in this Thesis, but default times were considered instead of recovery rates.

3.5.3 Annuity theory (The Profitability Test)

The *"bottom line"* for Lending for Financial institutions when granting Credit facility is the expected future value of the credit facility at the maturity term. Kellison and Irwin (1991) proposed the use of Principles of annuity theory to compute this value. However, these functions make the assumption that the no credit event will occur during the lifetime of the facility, this doesn't incorporate the aspect of credit risk. Incorporating the risk aspect is done using survival analysis since it gives a precise gauge of the probability of serviceability of a credit facility at each time point of the survival curve.

This thesis calculates the true future value (TFV) for the uncensored facilities, while accounting for their true final-state (maturity, default or early payoff), and compares the values to their estimated values utilising the survival models. The following assumptions were made when evaluating the model to make the results comparable;

- Facilities are repaid fully each month, with a fixed instalment amount.
- Facilities are dealt with as though they all originated at the same vintage, keeping in mind the end goal to make them practically identical.

Introducing the following;

1. $\boldsymbol{L_s}$, the Original facility amount for the facility
2. $\boldsymbol{R_s}$, the fixed sum of the monthly installment for the facility.
3. $\boldsymbol{n}$, the number of periods
4. $\boldsymbol{i}$, the defined interest rate, monthly $(i = \left(1 + i_y\right)^{1/12} - 1)$
5. $(\boldsymbol{E})\boldsymbol{FV}$ the expected future value of a facility

The fixed sum $\boldsymbol{R_s}$ consists of a repayment of the facility, $\boldsymbol{a_{s,j}}$ and some interest paid, $p_{s,j}$ each in month$\boldsymbol{j}$. Hence, $\boldsymbol{R_s} = \boldsymbol{a_{s,j}} + \boldsymbol{p_{s,j}}$. Note that where $\boldsymbol{R_s}$ remain constant, $a_{s,j}$ and $\boldsymbol{a_{s,j}}$ change over time, where $a_{s,j}$ increases and p_s, j decreases. Then,

$$L_s = \sum_{j=1}^{n} a_{s,j} = a_{s,1} * \sum_{j=1}^{n} (1+i)^{j-1} .$$

Hence it can be shown that

$$R_s = \frac{i}{1-(1+i)^{-n}}\, L_s.$$

The future value is expressed as;

$$FV_s = R_s((1+i)^{n-1} + (1+i)^{n-2} + \cdots + (1+i)^0)$$
$$= R_s \frac{(1+i)^n - 1}{i}. \qquad (3)$$

3.5.4 The true future facility values (TFV)

Eventual true future value (*TFV*) is dependent on the position of the facility at the end. For mature facility equation 3 is expressed with n as the facility term;

$$FV_{s\varepsilon mature} = R_s\, \frac{(1+i)^n - 1}{i}.$$

The future value (FV) for a credit facility which has experienced early payoff/repayment, the actual instalments received in the time period k is expressed as;

$$L_{s,k} = \left(1 - \frac{(1+i)^k - 1}{(1+i)^n - 1}\right) L_s.$$

If an early payoff occurs in the period k, assumption is, the credit facility is amortizes normally as per schedule until the period k, and the total $\boldsymbol{L_k}$ is fully settled in the period. The amount is then plugged back for reinvestment for $n-k-1$ periods.

$$FV_{s\varepsilon early} = R_s \left(\sum_{j=1}^{k} (1+i)^{n-j}\right) + L_{s,k}(1+i)^{n-k-1}.$$

The future value of a facility given defaults take occurs k months after is denoted as;

$$FV_{s\varepsilon default} = R_s \left(\sum_{j=1}^{k} (1+i)^{n-j}\right).$$

Subsequently, assumption is that when default occurs, zero recoveries are made on the remaining sum $\boldsymbol{L_k}$.

3.5.5 The Expected Future Value

Given the survival probability estimate, Dirick (2015) denoted $\hat{\mathrm{S}}(t)_{s,m}^{d}$ as the estimated probability that a given credit facility has not experienced a credit event by time *t*, using the model *m*. The eventual value of the facility using the non-mixture survival models is then computed following a specified model *m*;

$$\boldsymbol{EFV_{s,m}} = \boldsymbol{R_s}\left(\sum_{j=1}^{n} \hat{\mathrm{S}}(t=j)_{s,m}^{d}\left(1+i\right)^{n-j}\right) \quad where\ j = month.$$

The mixture cure model, the probabilities of susceptibility to a credit event is *PD*, and *1-PD* for a non-event. The P*D* is expressed as;

$$PD_s = \hat{\pi}(x_s) = \frac{\exp(\hat{b}_i x_s)}{1+\exp(\hat{b}_i x_s)}.$$

The eventual value of the facility using the mixture survival models with single event is then computed following a specified model *m*;

$$\boldsymbol{EFV_{s,m}} = PD_s * \boldsymbol{R_s}\left(\sum_{j=1}^{n} \hat{\mathrm{S}}(t=j)_{s,m}^{d}\left(1+i\right)^{n-j}\right) + (1-PD_s) * \boldsymbol{R_s}\frac{(1+i)^n-1}{\boldsymbol{i}}.$$

The methodology applied by Dirick 2015 where a Multinomial logit was applied was replicated in this thesis. Addition probabilities of Maturity and early payoffs/repayment are obtained. The estimated probability$\hat{\mathrm{S}}(t)_{s,m}^{e}$, that a credit facility has not experienced early payoff/repayment at a given time *t*. The eventual value of the facility using the mixture survival models with multiple event was then computed by;

$$\begin{aligned}\boldsymbol{EFV_{s,m}} = {} & PD_s * \boldsymbol{R_s}\left(\sum_{j=1}^{n} \hat{\mathrm{S}}(t=j)_{s,m}^{d}\left(1+i\right)^{n-j}\right) + (1-PD_s) * \boldsymbol{R_s}\frac{(1+i)^n-1}{\boldsymbol{i}} \\ & + PE_s \\ & * \left(\boldsymbol{R_s}\left(\sum_{j=1}^{n} \hat{\mathrm{S}}(t=j)_{s,m}^{e}\left(1+i\right)^{n-j}\right) + \sum_{j=1}^{n-1} \hat{\mathrm{S}}(t=j-1)_{s,m}^{e}\right. \\ & \left. - \hat{\mathrm{S}}(t=j)_{s,m}^{e}\right)\boldsymbol{L_{s,j}}\left(1+i\right)^{n-j-1}.\end{aligned}$$

Chapter 4 Results

The main results are summarised and grouped per evaluation measure, an empirical comparison between the models was conducted and convention notational assigned where the best result is marked with an asterisk. Significantly unique performances at a 5% level contrasting with the top performance with respect to a one-sided Mann -Whitney test are denoted in boldface.

Table 2 contains AUC values which represent the point estimate, with the highest values marked with an asterisk. The AUCs in Table 2 are stacked close together making it difficult to conclude on the preferred survival model when comparing the AUC alone (ties are due to rounding). This can be observed in Table 5, as the range for the average ranking is 4.2 to 6.19. An AFT Log Logistic models show to be the best preferred techniques. An AFT Model has all the earmarks of being a better option, when comparing using the average rankings, in spite of the fact that it shows up once in Table 2.

Table 2: Evaluating at different maturity points (1/3, 2/3, 3/3) - Area under the Curve.

Method/AUC	1/3	2/3	3/3
AFT Weibull	0.857	0.854	0.874
AFT Exponential	0.857	0.851	0.874
AFT Log-Logistic	0.857	0.854*	0.874
Cox PH	0.855*	0.851	0.876*
Mixture Cure Model	0.857*	0.856*	0.875*
Multi-Event Mixture Cure Model	0.857	0.853	0.873

* Denotes the best Values

The performance measure, for the AFT-Model, is significantly different when comparing to the best performing model at the 5% level for Table 3. The prevailing trend is that the Mixture Cure models and the Cox Proportional Hazard (PH) model outperform the AFT-models. Specifically, the default times prediction for the exponential AFT model seems to be significantly distant to the actual default times. The mean rankings have a more extensive range contrast with ROC (from 1.99 to 8.85), it's obvious that the default time prediction measures rightly favours the mixture cure model and the Cox PH model.

Table 3: Measures for Forecasting the default times for the observed defaults.

Method/Deviation Measure	MSE	MAE
AFT Weibull	**336.29**	13.63
AFT Exponential	**438.04**	**16.08**
AFT Log Logistic	**347.48**	13.67
Cox PH	231.04*	12.02*
Mixture Cure Model	235.33	12.09
Multi-Event Mixture Cure Model	265.06	12.76

* Denotes the best Values

Table 4 tabulates the mean of the absolute differences between the survival models and the expected future facility value estimates and the actual values. Based on domain knowledge the differences are expected to wider for facilities with lengthier facility tenure since the original sanctioned amount are larger too.

Table 4: Model performance using financial metrics. Mean absolute deviations (MAD) from the real future values.

Method	MAD for FV
AFT Weibull	362.67
AFT Exponential	380.37
AFT Log Logistic	363.11
Cox PH	360.39*
Mixture Cure Model	359.09*
Multi-Event Mixture Cure Model	444.96

From Table 5 the observed mean expected figures per facility tend to be closer to the mean real value of the credit facility with respect to the different survival methods. This information supports the applicability of survival models in the credit risk environment.

Table 5: Evaluating using financial metrics. Mean expected future values.

Method	Mean Future Value Per facility
AFT Weibull	15360.98*
AFT Exponential	15459.12*
AFT Log Logistic	15496.91
Cox PH	15504.59
Mixture Cure Model	15343.93*
Multi-Event Mixture Cure Model	15506.67
Mean FV Per facility	15445.37

Observing Table 5 the mean absolute difference regarding the future value of the facility, it is evident that the AFT-Weibull outperforms the other models, followed by the Mixture Cure Model**.**

Table 6: Mean ranking of the methods used.

	Area Under the Curve						
	1/3	2/3	3/3	MSE	MAE	MAD (FV)	EFV Vs. FV
AFT Weibull	4.75*	4.37*	4.86*	5.69	5.53	3.43*	6.41
AFT Exponential	5.03	5.47	6.03	8.85	8.85	7.30	2.43
AFT Log-Logistic	4.37*	4.20*	4.37	6.25	6.41	4.64	5.20
Cox PH	5.42	5.97	5.92	1.99*	2.21	3.65	6.52
Mixture Cure Model	5.20	5.20	4.64*	3.54*	3.54	3.57*	2.99
Multi-Event Mixture Cure Model	5.97	6.19	4.92	5.09	4.86	8.73	6.08

* Denotes the best Values

Evaluating using the 3 Measures; AUC, default time prediction, and future facility value estimation, it is evident that the Cox Proportional Hazard (PH) Model performs significantly well, additionally the mixture cure model does not perform differently. Table 7 demonstrates the determinants factors of interest.

To illustrate the interpretation of the parameter estimates Table 8 tabulates the output for the Cox-PH and the Mixture Cure Model.

Table 7 Description of the Variables

	Description	Type
v1	Gender (*1=M, 2=F*)	Categorical
v2	Net Monthly Income (*After Statutory Deductions Only*)	Continuous
v3	Age	Continuous
v4	Sanctioned Facility Amount	Continuous
v5	Years in Current Employment	Continuous
v6	Relationship with Bank (Years)	Continuous
v7	Debt Service Ratio *(DSR)*	Continuous

The survival component of the mixture cure model narrowed down to two covariates that actually influenced time to default – Debt Service Ratio (DSR) and Years in current Employment. The Logistic component likewise discovered comparative covariates to the Cox PH Model with the exclusion of Income. From the mixture cure results, we can conclude that the five Covariates in the logistic component of the model are significant predictors of the probability of being predisposed to default. The given survival model recommends just two covariates to be prescient of when a credit event will happen given that the borrower is inclined to default.

Table 8: Parameter estimate from the Cox PH Model and the mixture cure.

Mixture Cure				Cox Proportional Hazards			
Logistic model component	*OR*	*95% CI*	*P-Value*		*HR*	*95% CI*	*P-Value*
Gender	0.99	0.98-1.00	0.0048	Gender	0.97	0.96–0.99	0.001
Sanctioned facility Amount	2.71	2.18-3.34	<.0001	Sanctioned facility Amount	2.48	1.93–3.19	<.001
Relationship with Bank (Years)*	0.98	0.96-0.99	0.143	Years in Current Employment*	0.92	0.90–0.94	<.001
Age*	0.99	0.98-1.00	<.0001	Relationship with Bank (Years)*	0.97	0.96–0.99	0.001
Income*	1.03	0.80-1.32	0.843	Age*	0.98	0.96-0.99	0.146
				Debt Service Ratio	0.72	0.56–0.91	0.007
Survival Model Component	*HR*	*95% CI*	*P-Value*				
Years in Current Employment*	0.57	0.46-0.72	<.001				
Debt Service Ratio*	0.53	0.35-0.79	0.002				

* Continuous Covariate centred on mean

Chapter 5 Discussion

The current thesis evaluated the strength of six survival analysis models applicable in the consumer credit risk scoring. AUC, Default time prediction differences and future facility estimation were the main evaluation measures used in accessing model performance. The Cox-PH Model performed particularly well, the Mixture cure model performed significantly better and is among the top models.

The results of the covariates support previous research which concluded that survival techniques can perform identically to the commonly utilised logistic regression in regards to discrimination capacity by having the ability to utilise necessary information from the data. This is very useful in Credit risk where Survival analysis can provide a framework where PD can be estimated from a lifetime expected loss approach, by allowing the extension of effects of covariates on the predicted time to a credit event to fluctuate as the facility evolves. This provided substantial benefit relating to the estimation of accounts flowing into defaults and the actual profit realised from the portfolio at various time points.

This study shows that establishing a fitting assessment measure for comparing survival techniques remains a challenge since the AUC on its' own does not have the appropriate properties to clearly tell apart the different survival model.

Limitations

There were a number of limitations in the study. The dataset employed was composed of credit facility of personal Unsecured facility from of financial institution operating in Kenya, given that only a single sample was analysed, the results cannot be readily generalised to another portfolio such as mortgages were observed facility term are 15 years or more are common.

Future Research

Future research could involve the benchmarking of survival analysis techniques on revolving products (Overdrafts and Credit Cards) and the applicability of Mixture cure model to revolving products where the facility term is lengthy (Open ended), because of the revolving nature of the facility. This data would be characterised by lengthy observation periods in order to have non-vulnerable

subpopulations that do not achieve default status amid the credit line, since, some facilities, classified as transactors, always pay off their balance at the end of the interest-free period. Research can also be extended to observe the Mobile Lending propositions where credit Institutions issue loan to customers through their mobile phone, an example of the "*Mshwari"* loan offered by Commercial Bank of Kenya (CBA).

References

Andreeva, G. (2006). European generic scoring models using survival analysis. *Journal of the Operational research Society*, 1180-1187.

Belloti, T., & Crook, J. (2008). Credit scoring with macroeconomic variables using survival analysis. *The Journal of the Operational Research Society*, 1699-1707. doi:10.1057/jors.2008.130

Cao, R., Juan, M. V., & Andr´es, D. (2009). *Modelling consumer credit risk via survival.* Universidade da Coru˜na., Departamento de Matem´aticas., Spain. Retrieved Novemeber 13, 2016

Central Bank of Kenya. (2016). *Developments in the Kenyan Banking Sector for the quarter ended 31st March 2016.* Nairobi.

Collet, D. (2003). *Modelling Survival Data in Medical Research, Second Edition.* Chapman & Hall/CRC Texts in Statistical Science. CRC Press.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological),*, 34(2):pp. 187–220.

Curry, B. (2017, 04 05). *IFRS 9 and Collections – The 31-Day Time Bomb.* Retrieved 04 23, 2017, from FICO: http://www.fico.com/en/blogs/collections-recovery/ifrs-9-and-collections-the-31-day-time-bomb/

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.

Doorsselaere, J. V. (2015, April 07). *Risk Solutions.* Retrieved from Wolters Kluwerfs: http://www.wolterskluwerfs.com/onesumx/commentary/expected-losses-accounting-under-IFRS9.aspx

Fabien, C., & Pierre, J. (2007). A SAS macro for parametric and semiparametric mixture cure. *Computer Methods and Programs in Biomedicine*, 73-80.

Heagerty, P., & Saha, P. (2000). SurvivalROC: time-dependent ROC curve estimation from censored survival data. In *Biometrics,* (pp. 56(2):337–344.).

Kellison, S. G., & Irwin, R. D. (1991). *The theory of interest, volume 2.* Irwin Homewood,IL.

Kleinbaum,, D., & Klein, M. (2011). *Survival Analysis: A Self-Learning Text, Third Edition.* Statistics for Biology and Health. Springer.

Madorno, F., Mecatti, F., & Figini, S. (2013, June). Survival models for credit risk estimation. *Advances in Latent Variables-Methods, Models and Applications.*

Narain, B. (1992). Survival analysis and the credit granting decision. (L. C. Thomas, J. N. Crook, & D. B. Edelman, Eds.) *Credit Scoring and Credit Control*, 109-121.

Peng, Y., & Dear, K. (2000). A nonparametric micture model for cure rate estimation. In *Biometrics* (pp. 56(1):227–236).

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

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

Thomas, L. C., Crook, J. N., & J, B. (1999). Not if but When will Borrowers Default. *The Journal of the Operational Research Society, 50*(12), 1185-1190. doi:10.2307/3010627

Tong, E. N., Mues, C., & Thomas, L. C. (2012). Mixture cure models in credit scoring: If and when borrowers default. *European Journal of Operational Research, 218(1),* 132-139.

Watkins, J. T., Vasnev, A. L., & Geralach, R. (2014). Multiple event incidence and duration analysis for credit data incorporating non-stochastic loan maturity. *Journal of Applied Econometrics*, 29:627–648.

Zhang, J., & Peng, Y. (2007). A new estimation method for the semiparametric accelerated failure time mixture cure model. *Statist. Medicine 26,* pp. 3157-3171.

Zhang, J., & Thomas, L. (2012). Comparisons of linear regression and survival analysis using and mixture distributions approaches in modelling LGD. *International Journal of Forecasting, 18(2):204-215*.

Appendix

Selected SAS Code

The *pspmcm* Macro (The pieced-up segment)

```
/*The PSPMCM Model*/
%macro pspmcm(DATA=,SURVPART=, AFT= ,ID=,CENSCOD=,TIME=,
                VAR=,
              INCPART=,
                TAIL=, SU0MET=,
                FAST= ,BOOTSTRAP=,
                NSAMPLE=, STRATA=,
                MAXITER=,CONVCRIT=,ALPHA= ,
                BOOTMET=, JACKDATA=,
              GESTIMATE=,
                BASELINE=,
                SPLOT= ,
              PLOTFIT= );
option  nonotes nomlogic nomprint nosymbolgen nosource;
*options notes  mprint source symbolgen mlogic  ;
options formdlim=' ' nodate nonumber  ls=100;

ods listing;

/* Change to uppercase */
%let SURVPART       = %qupcase(&SURVPART);
%let AFT=           = %qupcase(&AFT);
%let ID             = %qupcase(&ID);
%let CENSCOD        = %qupcase(&CENSCOD);
%let TIME           = %qupcase(&TIME);
%let VAR            = %qupcase(&VAR);
%let LINK           = %upcase(&INCPART);
%let BOOTSTRAP      = %qupcase(&BOOTSTRAP);
%let strata         = %qupcase(&strata);
%let BASELINE       = %qupcase(&BASELINE);
%let GESTIMATE      = %qupcase(&GESTIMATE);
%let TAIL           = %qupcase(&TAIL);
%let FAST           = %qupcase(&FAST);
%let SU0MET         = %qupcase(&SU0MET);
%let BOOTMET        = %qupcase(&BOOTMET);
%let SPLOT          = %qupcase(&SPLOT);
%let PLOTFIT        = %qupcase(&PLOTFIT);


/* Default values  */
%let HYB=N ; %let PCTL=N; %let BOOTN=N; %let JACK=N ;
%let BCA=N ; %let BC=N;
%let jack_n=0;
%let conv_0=0;
```

```
%let crit_0=0;
%let done_0=0;
%let nokm=0;


%if &INCPART=  %then  %let LINK=LOGIT;
%if &MAXITER= %then %let maxiter=200;
%if &TAIL= %then %let tail=ZERO;
%if &SU0MET= %then %let SU0MET=PL;
%if &CONVCRIT= %then %let CONVCRIT=1e-5;
%if &ALPHA= %then %let ALPHA=0.05;
%if &GESTIMATE ne Y %then %let GESTIMATE = N;
%if &BASELINE ne Y %then %let BASELINE= N;
%if &BOOTMET= %then %let BOOTMET=N;
%if (&SURVPART ne COX) %then %do;
     %let BOOTSTRAP=N;
     %let tail=;
%end;
%if &BOOTSTRAP ne Y %then %do;
     %let nsample=0;
     %let BOOTMET=N;
     %let GESTIMATE= N;
%end;
%if ((&BCA=Y or &JACK=Y) AND &JACKDATA= ) %then %let
JACKDATA=Jackdist_t;
%if &SPLOT = %then %let SPLOT=N;
%if &SPLOT=Y %then %let BASELINE=Y;
%if &PLOTFIT=Y %then %do;
     %let BASELINE=Y;
     %let SPLOT=Y;
%end;
%if &PLOTFIT= %then %let PLOTFIT=N;

/*Time at the begenning */
data _null_;
time1=time();
call symput('time1',trim(left(time1)));
run;

*** compute confidence level;
data _null_;
conf=100*(1-&ALPHA);
call symput('conf',trim(left(put(conf,best8.))));
run;

/* Ckeck information*/
%prepare
%if (&exiterr ne 0) %then %goto exit;

/* Parametric mixture cure models */
%if (&SURVPART2=PARA) %then %do;
```

```
      %parametric
      %if (&BASELINE=Y) %then %do;
        %baseline
      %end;
     %if (&SPLOT=Y) %then %do;
       %survplot
     %end;
     %if (&PLOTFIT=Y) %then %do;
       %plotfit
     %end;
     %goto exit;
%end;

/* Error messsage if the specified number of replicates
is not high enough to compute bootstrap CI */
%if (&BOOTSTRAP=Y and &BOOTMET ne N) %then %do;
     data _null_;
     n=1/(&ALPHA/2);
     call symput('nes',left(n));
     run;

     %if (&nes>&NSAMPLE) AND (&NSAMPLE>=0) %then %do;
          %put ** The number of bootstrap sreplicates
specified is too low to allow computation of bootstrap
Confidence Intervals **;
          %let exiterr=1;
          %goto exit;
     %end;
%end;

/* If Bootstrap=N then number of sample=0 ans sample0=
intitial sample */
%if &BOOTSTRAP NE Y %then %do;
     %let nr=1;
     %let nb=0;
     %let msg1=original sample;
     data sample; set sample0;
     run;
%end;


/*If bootstrap=Y */
%if &BOOTSTRAP=Y %then %do;

 /* Number of loop needed for the specified number of
replicates (max size of data set set to 5.10e5 records)
*/
     data _null_;
     n=&NSAMPLE*&nobs;
     nr=ceil(n/5e5);
     nb=ceil(&NSAMPLE/nr);
```

```
     call symput('nr', trim(left(nr)));
     call symput('nb', trim(left(nb)));
     run;

     %if &nr>1 %then %do;
          %put **The dataset containing &NSAMPLE
bootstrap replicates would by very large   **;
          %put **SAS will resampling in &nr times of &nb
replicates each **;
     %end;
%end;

/* EM algorithm */
%let run=1;

%do %until(&run>&nr);

     %if &BOOTSTRAP=Y %then %do;
          %let nsample_b=&nb;
          %let seed=%eval(1234475+&run);
          %let msg1= &nb bootstrap replicates;
          %let boot_res=bootdist_t;
          %bootstrap
     %end;

     %put ;
     %put ** SAS will now compute estimates for &msg1;
     %max

     %let it=0;
     %let crit_g=0;
     %let nb_sep=0;
     %let conv_g=0;

     %do %until (&conv_g=1 or &it>=&MAXITER);
          %let it=%eval(&it+1);
          %let crit_u=0;
          %let conv_u=0;
          %let conv_1=0;
          %let nb_nc=0;
          %let nb_c=0;
          %maximisation
          %expectation
          %if &it>1 %then %do;
               %compare
          %end;
          %end_it
          %if ((&conv_0=1  or &it>=&MAXITER) and &done_0
ne 1)  %then %do;
               %result_0
               %if (&BASELINE=Y) %then %do;
```

```
                    %baseline
                %end;
                %if (&SPLOT=Y) %then %do;
                    %survplot
                %end;
                %if (&PLOTFIT=Y) %then %do;
                    %plotfit
                %end;
                %let done_0=1;
            %end;
            %if ((&conv_1=1 or &it>=MAXITER) and
&BOOTSTRAP=Y ) %then %do;
                %result
                %if (&BASELINE=Y) %then %do;
                    %baseline
                %end;
            %end;
     %end;
    proc datasets lib=work nolist; delete comp&it;
    run;

     %let run=%eval(&run+1);

%end;


/* Itération for BCA ou JACKKNIFE*/
%if (&JACKDATA= and (&BCA=Y or &JACK=Y)) %then %do;
          %put ** Creating the jackknifed
replicates.....;
          %put ;

          proc sort data=sample0 out=sample_j(drop=_obs_
_sample_); by &TIME descending &CENSCOD;
          run;

     %jacknife;

     /* Calculating the number of run(s) needed to
compute jackknife estimates */

     data _null_;
     nb_j=ceil(1e6/(&nobs-1));
     nr_j=ceil(&nobs/nb_j);
     call symput('nb_j',trim(left(nb_j)));
     call symput('nr_j',trim(left(nr_j)));
     run;

     %if &nr_j>1 %then %do;
          %put ** The number of jackknife replicates is
very large **;
```

```
          %put ** Then macro will perform &nr_j loops of
&nb_j replicates each**;
     %end;
     %else %do ; %let nb_j=&nobs; %let nr_j=1; %end;

     %let msg1=Jackknife re_samples;
     %let boot_res=jackdist_t;
    %let run_j=1;

     %do %until (&run_j>&nr_j);
     %put ** Calculating Jacknife estimates : loop
&run_j/&nr_j (&nb_j/&nobs) replicates) **;

          data sample ; set jackdata;
          %if &nr_j>1 %then %do;
             %if (&run_j<&nr_j) %then %do;
                 if &nb_j*(&run_j-
1)+1=<_sample_=<(&nb_j*&run_j);
             %end;
             %if (&run_j=&nr_j) %then %do;
                 if _sample_>=&nb_j*(&run_j-1)+1;
             %end;
          %end;
          run;

          %max

          %let it=0;
          %let crit_g=0;
          %let nb_sep=0;
          %let conv_g=0;

          %do %until (&conv_g=1 or &it>=&MAXITER);
               %let it=%eval(&it+1);
               %let conv_1=0;
               %let nb_nc=0;
               %let nb_c=0;
               %maximisation
               %expectation
               %if &it>1 %then %do;
                    %compare
               %end;
               %end_it
               %if (&conv_1=1) %then %do;
                    %result
               %end;
          %end;

          %let run_j=%eval(&run_j+1);
     %end;
%end;
```

```
/* Selection of converged replicates for CI intervals
compuation and graph */

%if &BOOTSTRAP=Y %then %do;
     %convrep
%end;

/* Bootstrap confidence interval computation*/
%if &BOOTMET ne N %then %do;
 %bootci
 %output
%end;

/*remane variable in output datasets */
%if (&SURVPART=COX) %then %do;
 %outset
%end;

/* QQ plot and distribution of parameters estimates*/
%if &GESTIMATE=Y %then %do;
 %graph
%end;

data _null_ ;
format time_tot time.;
time2=time();
time_tot= time2-&time1;
put 'Total time:' time_tot;
run;

%exit :;
%if (&exiterr ne 0) %then %do;
     %put ** The macro exited due to errors.**;
%end;

%else %do;
proc datasets lib=work nolist;
delete sample0
%if &SURVPART=COX %then bas compare conv b lp l lp2
min_max n_conv sample sep sortie1 sp sp2   _cov
%if (&BASELINE=Y and &BOOTSTRAP=Y) %then bas_boot
_baseline_t;
%if &BOOTSTRAP=Y %then fvar _acttr_ _boot_  ci_bootdist
bootdist_t boot boottran;
%if &TAIL=WTAIL %then weib weib1;
%if &BOOTN=Y or &jack=Y %then _tmp2_ ;
%if &BCA=Y %then jackskew _jack_ sample_j bootpctl_bc;
%if &BC=Y %then boottran _bc_ bootpctl_bc;
%if &HYB=Y %then bootpctl_hyb;
```

```
%if &PCTL=Y %then bootpctl_pctl;
%if &JACK=Y %then jackdist_t ;
%if (&SPLOT=Y) or (&BASELINE=Y) %then  _baseline ;
%if (&SPLOT=Y) %then _graphkmu _graphkm graph;
%if &SURVPART2=PARA %then param ;
%if &PLOTFIT=Y and &nokm=0 %then cm _leg;;
run;
quit;
%end;
options notes source;
%mend pspmcm;
```

Calling the *pspmcm* Macro

```
%macro mcm(Data,Censored,stime,out);
*call the PSPMCM macro;
%pspmcm(DATA = &Data,ID=CustAcc,CENSCOD=&
Censored,TIME=&stime,
VAR= Gender(IS,0) Income (IS,0) Age (IS,0)
Original_Loan_Amount(IS,0) DSR(IS,0) Time_In_Employment
(IS,0) Time_With_Bank DSR (IS,0),
INCPART=logit,SURVPART=cox,TAIL=zero , SU0MET=pl,
FAST=Y,BOOTSTRAP=Y,NSAMPLE=2000, STRATA=,
MAXITER=200,CONVCRIT=1e-5, ALPHA=0.05,
BASELINE=Y, BOOTMET=ALL, JACKDATA=, GESTIMATE=Y, SPLOT=Y,
PLOTFIT=Y);
option  nonotes nomlogic nomprint nosymbolgen nosource;
options formdlim=' ' nodate nonumber  ls=100;
ods listing;
run;
proc sql ;
create table cox as
select
Fast_or.*,
Fast_inci.ProbChiSq as or1,
Fast_surv.HazardRatio,
Fast_surv.HRLowerCL,
Fast_surv.HRUpperCL,
Fast_surv.ProbChiSq as hrp
from Fast_or, Fast_inci, Fast_surv
where Fast_surv.Parameter = Fast_inci.Variable and
Fast_inci.Variable = Fast_or.Effect;
data cox; set cox;
coxOR =
put(OddsRatioEst,4.2)||"("||put(LowerCL,4.2)||","||put(Up
perCL,4.2)||")";
coxORp = or1;
coxHR =
put(HazardRatio,4.2)||"("||put(HRLowerCL,4.2)||","||put(H
RUpperCL,4.2)||")";
coxHRp = hrp;
```

```
drop OddsRatioEst LowerCL UpperCL HazardRatio HRLowerCL
HRUpperC or1 hrp HRUpperCL;
run;
proc datasets library=work;
delete Fast_or Fast_inci Fast_surv;
run;

%pspmcm(DATA = &Data ,ID= CustAcc,CENSCOD=&
Censored,TIME=&stime,
VAR= Gender(IS,0) Income (IS,0) Age (IS,0)
Original_Loan_Amount(IS,0) DSR(IS,0) Time_In_Employment
(IS,0) Time_With_Bank DSR (IS,0), INCPART=logit,
SURVPART=WEIB, TAIL=zero , SU0MET=pl, FAST=Y,BOOTSTRAP=Y,
NSAMPLE=2000, STRATA=, MAXITER=200,CONVCRIT=1e-5,
ALPHA=0.05,
BASELINE=Y, BOOTMET=ALL, JACKDATA=, GESTIMATE=Y, SPLOT=Y,
PLOTFIT=Y);
run;
proc sql ;
create table weibull as
select
Fast_or.*,
Fast_inci.ProbChiSq as or1,
Fast_surv.HazardRatio,
Fast_surv.HRLowerCL,
Fast_surv.HRUpperCL,
Fast_surv.ProbChiSq as hrp
from Fast_or, Fast_inci, Fast_surv
where Fast_surv.Parameter = Fast_inci.Variable and
Fast_inci.Variable = Fast_or.Effect;

data weibull; set weibull;
weibullOR =
put(OddsRatioEst,4.2)||"("||put(LowerCL,4.2)||","||put(Up
perCL,4.2)||")";
weibullORp = or1;
weibullHR =
put(HazardRatio,4.2)||"("||put(HRLowerCL,4.2)||","||put(H
RUpperCL,4.2)||")";
weibullHRp = hrp;
drop OddsRatioEst LowerCL UpperCL HazardRatio HRLowerCL
HRUpperC or1 hrp HRUpperCL;
run;

proc sql;
create table &out as
select cox.* , weibull.*
from cox, Weibull
where cox.Effect = weibull.Effect;
%mend;
```